A Search For Contingency Genes In *Candida albicans*

A thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in the Institute of Fundamental Sciences
Massey University, Palmerston North
New Zealand

Synodalia Chrisma Wattimena

2014

Abstract

Many microbial pathogens have been known to use repeats in their cell wall proteins to generate diversity, and this has been found to contribute to their virulence. In bacteria, these genes are called contingency genes, and function to facilitate adaptation of bacteria to the host environments as they invade different host parts and to evade the host's constantly evolving immune system. In the diploid *Candida albicans*, few genes have been classified as contingency genes due to the variation in the length of their repeat regions in different clinical isolates. This study attempts to answer a question of whether *YWP1*, *HWP1*, and *EAP1* of *C. albicans* are contingency genes. These three genes encode cell wall proteins and contain repeats. For this purposes, allelic distributions of the genes in the general purpose genotype (GPG) and non-GPG strains (two groups with different genetic backgrounds), in commensal and infection strains, and in strains isolated from different sites of the humans body were examined. Based on the allelic distributions of the genes in GPG and non-GPG strains, it can be inferred that *YWP1* and *HWP1* can be categorized as contingency genes, while *EAP1* cannot be categorized as a contingency gene. The allelic distributions of the genes in commensal and infection strains indicate that *YWP1*, *HWP1*, and *EAP1* do not act as contingency genes when *C. albicans* state changes from commensal to pathogenic. Although the allelic distributions of the genes cannot distinguish commensal from infection strains, the non-random association between alleles of *YWP1*, *HWP1*, and *EAP1* does distinguish these two groups, i.e. the *YWP1-HWP1-EAP1* association is stronger in commensal strains that it is in infection strains. Based on the allelic distribution of the genes in strains isolated from different sites of the human body, it can be inferred that *YWP1* and *EAP1* do not act as contingency genes, but *HWP1* may act as a contingency gene, when *C. albicans* moves to particular sites of the human body.

Acknowledgments

I would like to thank Dr. Jan Schmid and Dr. Mark Patchett as my supervisors, who introduce me to *Candida albicans* field. I found this field interesting, and am ready to continue this study in years to come. I am thankful for Dr. Mark Patchett for his helpful comments in shaping my thesis. I am grateful for the encouragement I received from Dr. Kathryn Stowell as IMBS graduate advisor in many occasions. I really appreciate the helps I received from Richard Fong, Lorraine Berry, and Trish McLenachan from the Allan Wilson Centre Genome Service at Massey University for helpful discussions on the sample preparations. I would also like to thank the members of Dr. Scmhid's Lab; special thanks for Zhuo Zhou for her collaboration work with me in preparing some samples, in particular the 300 generation sample, Pany Vijayan, Dennys and Yeying Wang for introducing me to many new stuffs in the beginning of my program, and Shuguang Zhang for many helpful discussions. I am really grateful for the generous funding from NZAID (the Ministry of Foreign Affairs & Trade) for my PhD program, and for Sylvia Hooker, Sue Flynn, Olive Pimental, and Jamie Hooper from Massey International Office for their supports. My special thanks go to my family in Ambon, Indonesia for their prayers and supports. Finally, I would like to give my very special Thanks to my daughter Wilda, and my husband Philip, for their supports, loves, understandings, helps, and prayers.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Literature Review

1.1.1 *Candida albicans* and Humans Host

Candida albicans is a normal member of the microflora on the mucosal surfaces of most healthy persons. It is uniquely associated with humans or animals, and rarely found in other environments such as soil. Although it is not a dangerous organism and most of the time is not able to infect healthy people, overgrowth of this fungus can cause superficial as well as life-threatening systemic infections in immunosuppressed patients due to cancer treatments or diseases such as AIDS [1]. For these patients, candidiasis is frequently caused by a resident strain and is triggered by changes in the host immune system [2, 3]. Therefore, *C. albicans* is well adapted to humans environments. However, there is evidence that *C.albicans* strains can be transmitted between persons and then replace existing commensal strains in immunosuppressed patients [4].

In order to colonize and infect humans, *C. albicans* cell has to come into contact with the humans body. There are several known ways for *C. albicans* to enter the humans body. For examples, contact with carriers [5] and transfer via contaminated food [6]. After it is in contact with the humans body, the cells must adhere on host surfaces in order to replicate and then colonize the body. Without this adherence,

the cells will be washed away from the body surface and colonization of the body will not occur. Colonization occurs when a certain number of *C. albicans* populations are maintained in the host without producing disease symptoms [5]. It usually takes place on the gastrointestinal tract (from the oral cavity to the rectum and anal tissues), genitourinary tract, and in some cases on the skin. Colonization depends on the ability of the cells to evade innate primary host defenses such as the epithelia physical barrier and antimicrobial molecules (peptide, lysosime and lactofferin) [7, 8, 9]. On the skin, colonization is enhanced by heat and moister. In warm moist body areas such as skin folds, the growth of the cells may increase rapidly. The breakdown of the outer layer of the skin promotes the cell's overgrowth. If the number of cells increase in high numbers (cell overgrowth) due to reduced activity of the host immune system in immunocompromised individuals, the penetration of tissue will occur and the symptom of candidiasis, such as thrush on the mouth and the gastrointestinum system and rush on the skin, will appear. The fungus then may enter the blood stream and then disseminates via the blood stream to the deep organs such as liver, spleen and kidney and then later invades those organs [1].

The ability of *C. albicans* to colonize and infect many parts of the humans body indicates that this fungus must have numerous mechanisms that allow the fungus to adapt and survive in different environments of the body. Different environments may have different temperature, pH, oxygen, carbon dioxide, osmolarity etc. Adhesion factors, phenotypic switching and extracellular proteolitic or lipolytic activity have been recognized as virulence factors required by *C. albicans* to cause infection [1]. Virulence factors expressed by *C. albicans* depends on the type and site of infection, and on the host response.

1.1.2 Cell Wall of *C. albicans*

The cell wall of *C. albicans* contains 80-90% carbohydrate, 6-25% protein, and 1-7% lipid [10]. It is composed of three major carbohydrates: β-1.3 glucan, β-1.6-glucan and chitin. Mannose polymers (Mannan) are normally found in association with cell wall protein (CWPs). Phosphopeptidomannans (PPM) form a fibria layer at

the outermost layers of the *C. albicans* cell wall [11], and are covalently linked to cell wall structural polysaccharide via two ways. First, they link to β-1,6 glucan via a glycophosphatidylinositol-anchor remnant (GPI-anchor) and second, they link directly to β-1,3 glucan [12, 13, 14].

Several GPI-anchor cell wall proteins of *C. albicans* have been identified, i.e. ALS family, EAP1p, HWP1p, YWP1p [15]. The ability of *C. albicans* to alter its cell wall protein composition has been suggested to play important roles to the virulence of this fungus and enable this fungus to adapt well to its host condition. The new composition of the cell wall proteins may make the fungus to cope better with the new environmental conditions. Since cell wall proteins play roles in many events including adhesion, biofilm formation and tissue invasion [16, 17, 18], which are important factors for the success of host infection, the protein composition of fungal cell walls of virulent strains may relate to the ability of the strains to cause disease on their hosts. By varying their cell protein surface the fungal pathogens may escape the host immune system and infect abroad range of tissues/substrates.

Cell wall protein (CWP) composition of *C. albicans* can be altered by either regulating their expressions or by changing the characterization of individual protein through the variation on its repeats.

A. Alteration of CWP composition by regulating gene expression

CWP composition of *C. albicans* can be altered by regulating the expression of CWP genes [11]. Based on computer analysis of C-terminal hydrophobic amino acid residues that characterize GPI-anchor attachment site, around 115 *C. albicans* proteins have been predicted to be modified with GPI-anchors [19]. This modification occurs in the endoplasmic reticulum and is a posttranslational modification. Proteins with GPI-anchors are targeted to fungal cell wall and attached to cell wall β-1.6-glucan [20]. From these 115 proteins, only around 20 GPI-anchored proteins are found in the cell wall of *C. albicans* cells growing *in vitro* [21].

Studies using mass spectrometry-based proteomics of fungal wall glycoproteins

show that at any one time, the cell walls of *Saccharomyces cerevisiae* and *C. albicans* consisted of more than 20 dissimilar cell wall glycoproteins [16]. Environmental conditions such as nutrient accessibility, temperature and pH, changed the protein composition of cell walls of these two fungi [22, 23]. The cell wall proteome variation has also been found on the cells growing under different conditions e.g. between yeast and hyphae, and between hypoxic and iron-limiting conditions [24, 25].

B. Alteration of CWP composition by variation of the repeats

Cell surface composition of *C. albicans* can also be altered by changing the characterization of individual CWPs through the variation of their repeats [11]. Addition or reduction of the number of repeat units within genes that encode cell wall proteins can change the functional characteristics of the proteins and also can determine the position of functional region of the proteins on the cell wall surface.

Repeats have been located more frequently in Fungal CWPs than in other classes of proteins, and the number of repeat containing cell wall protein is found to be larger in *C. albicans* compared to *S. cerevisiae* or *A. fumigates* [26]. Approximately 115 genes in *C. albicans* have been predicted to encode cell surface proteins [19], and 50 % of the genes have repeat sequences [11]. Many repeat-containing CWPs in *C. albicans* are adhesins including *EAP1*, *YWP1*, *HWP1*, and ALS family genes [15]. They mediate adhesion to the host. The function of ALS proteins in adhesion seemed to be associated with variation in the number of tandem repeat copies. This conclusion is based on the facts that ALS3 proteins with 9 tandem repeat copies played less roles in *C. albicans*'s adhesion to endothelial and epithelial surfaces than do ALS3 proteins with 12 tandem repeat copies [27]. Loza *et al.* [28] found that when each of the ALS1 with 5 repeat units or the ALS1 with 20 repeat units was expressed in non-adherent *S. cerevisiae* cells, adhesion of the cells with ALS1-5 repeat units decreased by 50% compared with the cells with ALS1-20 repeat units. They also found that the cells with ALS1-0 repeat unit had no adhesion ability at all.

Rauceo *et al.* [29] found that when *ALS5* with different number of repeat units (0-6) were expressed in non-adherent *S. cerevisiae* cells, there was a correlation between

adhesion of the cells and the number of ALS5 repeat units contained by the cells.

Variation in repeat numbers of ALS genes has been found among strains of *C. albicans*. The variability is thought to generate diversity in the population, which enable the organism to survive in different host environment. The ALS genes (*ALS1* to *ALS7*, and *ALS9*) of *C. albicans* have repeat regions which consist of tandemly repeated copies of a 108-bp sequence. Rearrangement of the units in the repeat regions of these genes results in highly allelic diversity expressed by different numbers of 108-bp tandemly repeated copies [30]. Examination of *ALS1* allelic diversity in over 100 isolates of *C. albicans* revealed that the isolates had alleles with a copy number of the tandem repeat sequence between 4 to 37 but the most common allele had 16 copies [31]. Zhang *et al.* [32] studied *ALS7* allelic diversity in a group of strains called the general-purpose-genotype (GPG) cluster which cause humans diseases 10-100 times more often than other strains and found that this group had between 14 and 17 copies of the tandem repeat in the central domain and these alleles were uncommon in other group. Repeat variability and clade-specific allele combinations which relate to tandem repeat copy numbers in *ALS3*, *ALS5* and *ALS6* were also found [27, 33].

1.1.3 Repeat Sequences

Repeat sequences (three or more nucleotides) are found frequently in coding genomic DNA sequences of both prokaryotes and eukaryotes [32, 34, 35]. Repeats are also termed as simple sequence repeats (SSR) or short tandem repeats (STR). Repeats of DNA sequences of 1-9 bp in length are called microsatellites, while repeats of DNA sequences of 10-100 are called minisatellite [36]. Repeats are caused by replication slippage and recombination between homolog sequences [37, 38]. The regions are unstable and have a higher mutation rate compare to the rest of the genome.

Many algorithms have been used to identify repeat regions in DNA sequences, for example Tandem Repeat Finder, REPuter, Stepstone, the EMBOSS ETANDEM, SERV [39, 40, 41, 42]. One of the programs, SERV, calculates a VAR score that can determine the probability of the repeat variation in a population. The score is calculated based on the length of an individual repeat, the purity of the repeat and

the number of the repeat. The variability of both microsatellites and minisatellites in many organisms can be predicted using this program [42].

1.1.4 Contingency Genes

In bacteria, many genes have been found to contain repeats. These genes are called contingency genes when a change in the gene repeat numbers is used for adaptation purposes [43, 44, 45]. These genes are important in bacterial pathogenicity since they function to facilitate adaptation of bacteria to the host environments as they invade different host parts and to evade the host's constantly evolving immune system. These genes create variability in the bacterial cell population that allows the pathogen to survive in changing host environments. Bacteria use contingency genes as one of the evolvable mechanisms to change their rate of mutation [46]. This diversity-generating mechanism may contribute to bacterial adaptation both short term adaptation and long term evolution.

Many *C. albicans* repeat-containing genes have been identified such as: *EF3* [47], *CEK1*, *HYR1*, *HYR2* [31], *Rlm1* [48], *ALS* family [49], *ALS7* [32], *ALS3* [27], *ALS5*, *ALS6* [33, 50], *PNG2* [51], and *SSR1* [52]. From the list of these studies, only two genes have been claimed as contingency genes: *ALS7* [32] and *SSR1* [52].

1.2 Research Background

Contingency genes have been the focus of many studies in bacteria. The genes contain repeats within their coding sequences, and play a role in facilitating adaptation to different host microenvironments [44, 45]. In bacteria, a contingency gene has relatively high mutation rates, which result in a high variation of gene repeats in the population. This variation helps bacteria to evade their host immune system, which constantly evolves, or to colonize different host microenvironments [44, 53, 54]. The high mutability of the genes is expected to produce a pool of variant alleles from which the most suitable ones are selected, thus the variability in alleles are used for adaptation purposes. Since contingency genes in bacteria have relatively high mutation rates,

the term contingency genes are frequently used interchangeably with hypermutable genes. The mutation rate for contingency genes in bacteria has been determined to be in the range of 1×10^{-5} to 1×10^{-2} per cell division [55].

Repeats in *C. albicans* genes have been the focus of many studies, which include repeats in *EF3* [47], *CEK1*, *HYR1*, *HYR2* [31], *Rlm1* [48], *ALS* family [49], *ALS7* [32], *ALS3* [27], *ALS5*, *ALS6* [33, 50], *PNG2* [51], and *SSR1* [52]. From the list, only few studies discussed the relationship between the repeat variation and its effect on the fungus adaptation to its host, thus contingency genes; these studies include *PNG2* [51], *ALS7* [32], and *SSR1* [52].

Using Ca3 fingerprinting, three clusters of *C. albicans* strains have been identified from 266 infection strains collected from six different countries: 37% of strains were genetically homogeneous and classified as cluster A, while 63% of strains were genetically diverse and classified as either cluster B or C (see Figure 1.1 on page 8) [56]. Cluster A is called the general purpose genotype (GPG) cluster, while clusters B and C are non-GPG clusters. Figure 1.1 indicates that GPG and non-GPG strains have different genetic backgrounds. Since GPG and non-GPG strains have different genetic backgrounds, these groups can be used to determine whether a *C. albicans* gene is a contingency gene, i.e. whether the repeat variation is used for adaptation purposes. If there is a selective pressure on the gene, the influence of the genetic background is diminished [57], which will result in the same alleles selected by the two groups. Viceversa, if there is no selective pressure on the gene, the two groups may select different alleles due to the influence of the genetic background.

Studies to compare allele combinations of the *C. albicans* genes including *ALS7*, *PNG2*, and *SSR1*, between GPG and non-GPG strains have been conducted [32, 51, 52], and shared at least three common key findings. First, there was a variability in the number of repeat units observed for each gene in the strains tested. Second, there were predominant allele combinations associated with the group of GPG strains, while no obvious predominant allele combinations were observed in the group of non-GPG strains. If the repeat variation was used for adaptation purposes, the predominant allele combinations should be observed in non-GPG strains, and should be the same

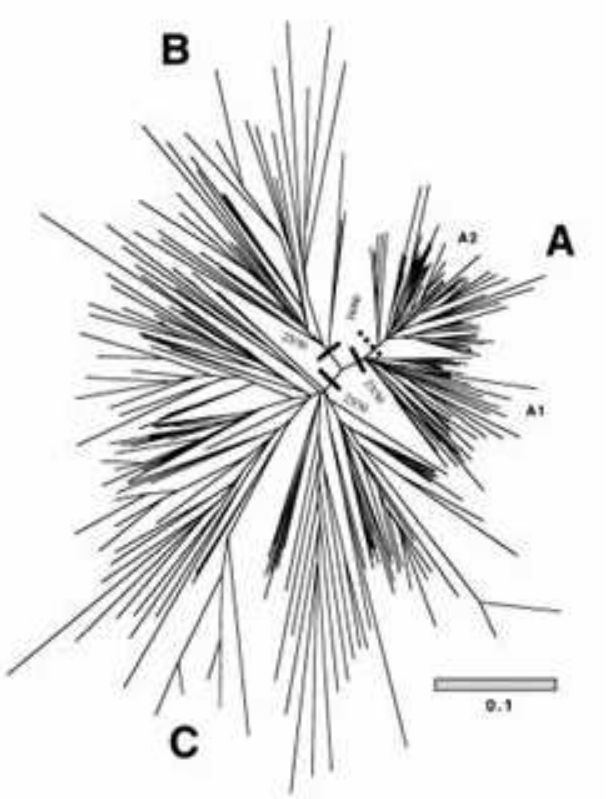

Figure 1.1: A neighbour-joining tree of 266 *C. albicans* strains. 37% of strains were genetically homogeneous and classified as cluster A, while 63% of strains were genetically diverse and classified as either cluster B or C [56]. Cluster A is called the general purpose genotype (GPG) cluster, while clusters B and C are non-GPG clusters.

as that in GPG strains. Thus, based on the genetic background, these three genes cannot be categorized as contingency genes. Third, the results of the mutation rate measurements of *PNG2*, and *SSR1* indicate that these two genes are hypermutable genes. Although the same measurements were not conducted for *ALS7*, by considering the number of different allele combinations observed, it was claimed that *ALS7* is a hypermutable gene [32]. The findings that *ALS7*, *PNG2*, and *SSR1* are hypermutable genes seem to contradict with the facts that there are predominant allele combinations of *ALS7*, *PNG2*, and *SSR1* observed in GPG strains; high mutation rates should produce more allele diversity with less or no predominant allele combinations.

If the criteria for a *C. albicans* contingency gene is based on the gene allelic distributions in GPG and non-GPG strains, it can be inferred that *PNG2*, *ALS7* and *SSR1* which are hypermutable genes, are not contingency genes. This seems to reveal that hypermutable genes of *C. albicans* may not act as contingency genes. Beside the genetic background, the criteria for a *C. albicans* contingency can also can be based on the gene allelic distributions in commensal and infection strains, i.e. whether there is a change in the repeats when the *C. albicans* state changes from commensal to pathogenic. Based on this criteria, *SSR1* can be categorized as a contingency gene, but no information is available for *ALS7* and *PNG2*. Another criteria is based on the allelic distributions in strains isolated from different sites of the humans body, i.e. whether there is a change in the repeats when *C. albicans* moves to particular sites of the humans body.

In this thesis, the results of a search for contingency genes in *C. albicans* are described. Since contingency genes and hypermutable genes are frequently used interchangeably, it is important to emphasize that in this study, the two terms are separated. A contingency gene is defined as a gene which has a role in adaptation by changing the repeats within the coding sequences, and the criteria used for a contingency gene is based on the allelic distributions of the gene in GPG and non-GPG strains, i.e. the allelic distribution of a gene which plays a role in adaptation should be the same in GPG and non-GPG strains. In addition, allelic distributions in commensal and infection strains, and strains isolated from different sites of the humans

body are also used to determine whether a *C. albicans* gene is a contingency gene or not. A hypermutable gene is defined as a gene with mutation rates ranging from 1×10^{-5} to 1×10^{-2} per cell division, the same criteria for bacteria [55].

Three potential contingency genes are chosen based on two criteria: 1) the genes encode cell wall proteins (CWPs), and 2) the genes contain repeats. These three potential genes are *YWP1* of chromosome 2, *HWP1* of chromosome 4, and *EAP1* of chromosome 2. These 3 genes have been predicted to encode GPI-anchored cell wall proteins [13, 58], and contain repeats [59]. The three genes also have been found to involve in adherent ability of the *C. albicans* cells. The YWP1p has been suggested to promote the dispersal ability of the *C. albicans* [60, 61]. The HWP1p has been found to be a surface adhesin required for both covalent attachment to host epithelial cells and hypha adhesion one with another [62, 63, 64]. The EAP1p has been found to have three domains, which relate to the adherent ability of the *C. albicans* cells to polystyrene and mammalian epithelial cells [65, 66].

1.3 Research Questions

In this study, alleles of *YWP1*, *HWP1*, and *EAP1* were characterized and examined to find out whether *YWP1*, *HWP1*, and *EAP1* are contingency genes. In addition, the mutation rates of *YWP1*, *HWP1*, and *EAP1* were determined to find out whether *YWP1*, *HWP1*, and *EAP1* are hypermutable genes. Therefore, this study was conducted to answer the following general questions:

1) Are *YWP1*, *HWP1*, and *EAP1* contingency genes ?

2) Are *YWP1*, *HWP1*, and *EAP1* hypermutable genes ?

To answer the first general question, the allelic distributions of *YWP1*, *HWP1* and *EAP1* in two different genetic background strains (GPG and non-GPG strains), among commensal and infection strains, and among strains isolated from different sites of the humans body were examined. For a contingency gene, it is expected that two different genetic background strains select the same alleles of the gene, which are the advantageous alleles. The examination of the allelic distribution of the three

genes in commensal and infection strains is to answer a question of whether *YWP1*, *HWP1* and *EAP1* act as a contingency gene when the *C. albicans* state changes from commensal to pathogenic. The examination of the allelic distribution of the three genes in strains isolated from different sites of the humans body is to answer the question of whether *YWP1*, *HWP1* and *EAP1* act as contingency genes when the *C. albicans* moves to particular sites of the humans body. Finally, these studies also try to answer the question of whether there is an interconnectedness, i.e. a non random association of alleles of *YWP1*, *HWP1* and *EAP1*. The analysis of the gene interconnectedness can reveal whether there is a selection of certain genes to interconnect to each other when the *C. albicans* faces a particular circumstance, e.g. when the *C. albicans* state changes from commensal to pathogenic.

To answer the second general question, the mutation rate of *YWP1*, *HWP1*, and *EAP1* was determined. The mutation rate for a hypermutable gene in bacteria has been determined to be in the range from 1×10^{-5} to 1×10^{-2} per cell division [55], and will be used as a standard to determine whether or not *YWP1*, *HWP1*, and *EAP1* are hypermutable genes.

Chapters 3, 4, and 5 contain the descriptions to answer the question of whether *YWP1*, *HWP1*, and *EAP1*, respectively, are contingency genes. Each chapter includes the description of the allelic distribution of the genes in GPG and non-GPG strains, commensal and infection strains, and strains isolated from different sites of the humans body. In addition to these, the analysis of the combination of the two alleles in an individual strain is also described. This analysis describe the composition of the homozygous and heterozygous strains, and the pattern of the difference between the two alleles in an individual strain for the three genes in the strains of interest.

Chapter 6 contains the description to answer the question of whether there is an interconnectedness among the three genes.

Chapter 7 contains the description to answer the question of whether *YWP1*, *HWP1*, and *EAP1* are hypermutable genes.

Chapter 8 contains the summary of the results, the conclusion and proposed future experiments.

Chapter 2

Materials and Methods

2.1 Fungal and Bacterial Strains

2.1.1 Fungal Strains

C. abicans strains used in this study were infection and commensal strains. Infection strains are strains isolated from sick (unhealthy) persons and commensal strains are strains isolated from healthy persons. For infection strains, two groups were investigated: GPG and non-GPG strains. For commensal strains, only GPG strains were investigated. The strains were isolated from different sites of the human body, and collected from different parts of the world. From 266 strains originally studied, 123 strains were selected, of which 49 were GPG strains isolated from infected sites, 47 were non-GPG isolated from infected sites, and 37 were GPG strains isolated from uninfected sites (healthy person). Tables 2.1, 2.2, and 2.3 contain the list of GPG infection strains, non-GPG infection strains, and GPG commensal strains, respectively.

Table 2.1: List of GPG infection strains used in this study.

No	Strain	Site	Country of origin	Reference
1	AU1	respiratory/oral	New Zealand	[67]
2	AU19	urine	New Zealand	[67]
3	AU27	sputum	New Zealand	[67]
4	AU39	skin/wounds	New Zealand	[67]
5	AU90	skin /wounds	New Zealand	[67]
6	CH14	skin/wounds	New Zealand	[67]
7	CH35	urine	New Zealand	[67]
8	CH42	skin/wounds	New Zealand	[67]
9	CHOB5	skin/wounds	New Zealand	[67]
10	CLB42	vagina/vulva	Columbia	[56]
11	CLB53	skin/wounds	Columbia	[56]
12	COUR-C	respiratory/oral	USA	[56]
13	FJ9	respiratory/oral	Fiji	[56]
14	FJ23	urine	Fiji	[56]
15	FJ26	skin/wounds	Fiji	[56]
16	HUN93	sterile	Great Britain	[56]
17	HUN95	sterile	Great Britain	[56]
18	HUN96	Sterile	Great Britain	[56]
19	HUN122	Sterile	Great Britain	[56]
20	HUN127	Sterile	Great Britain	[56]
21	Jam-2c	anal	USA	[67]
22	KO-2c	stool	USA	[67]
23	OD8807	respiratory/oral	Great Britain	[56]
24	OD8826	respiratory/oral	Great Britain	[56]

continued on the next page...

Table 2.1 – continued

No	Strain	Site	Country of origin	Reference
25	OD8911	respiratory/oral	Great Britain	[56]
26	OD8916	respiratory/oral	Great Britain	[56]
27	OD9014	respiratory/oral	Great Britain	[56]
28	RIHO9	sterile	USA	[56]
29	RIHO10	sterile	USA	[56]
30	RIHO13	sterile	USA	[56]
31	RIHO16	sterile	USA	[56]
32	ROLO-C	respiratory/oral	USA	[56]
33	SIM-C	respiratory/oral	USA	[56]
34	VAR1.1VAG	vagina/vulva	USA	[56]
35	VAR1.3VAG	vagina/vulva	USA	[56]
36	VAR1.4VAG	vagina/vulva	USA	[56]
37	VAR1.10VAG	vagina/vulva	USA	[56]
38	VARE1.8VAG	vagina/vulva	USA	[56]
39	W3	respiratory/oral	New Zealand	[67]
40	W26	anal	New Zealand	[67]
41	W43	respiratory/oral	New Zealand	[67]
42	W59	sputum	New Zealand	[67]
43	W68	respiratory/oral	New Zealand	[67]
44	W132	respiratory/oral	New Zealand	[67]
45	W134	sputum	New Zealand	[67]
46	YSM073	respiratory/oral	Malaysia	[56]
47	YSU568	urine	Malaysia	[56]
48	YSU649	urine	Malaysia	[56]
49	YSU751	urine	Malaysia	[56]

Table 2.2: List of Non-GPG infection strains used in this study.

No	Strain	Site	Country of origin	Reference
1	AU11	skin	New Zealand	[67]
2	AU134	respiratory/oral	New Zealand	[67]
3	AU2	respiratory/oral	New Zealand	[67]
4	AU33	skin/wounds	New Zealand	[67]
5	AU36	catheter	New Zealand	[67]
6	CH3	urine	New Zealand	[67]
7	CH9	vagina/vulva	New Zealand	[67]
8	CH20	vagina/vulva	New Zealand	[67]
9	CH41.1	urine	New Zealand	[67]
10	CLB44	skin/wounds	Columbia	[56]
11	CLB45	skin/wounds	Columbia	[56]
12	CLB49	respiratory/oral	Columbia	[56]
13	FJ3	respiratory/oral	Fiji	[56]
14	FJ12	respiratory/oral	Fiji	[56]
15	FJ27	catheter	Fiji	[56]
16	GAYMC-C	respiratory/oral	USA	[56]
17	HUN61	respiratory/oral	Great Britain	[56]
18	HUN64	skin/wounds	Great Britain	[56]
19	HUN66	skin/wounds	Great Britain	[56]
20	HUN68	respiratory/oral	Great Britain	[56]
21	HUN91	sterile	Great Britain	[56]
22	HUN92	sterile	Great Britain	[56]
23	HUN123	sterile	Great Britain	[56]
24	OD8824	respiratory/oral	Great Britain	[56]

continued on the next page...

Table 2.2 – continued

No	Strain	Site	Country of origin	Reference
25	OTG1	respiratory/oral	New Zealand	[67]
26	OTG2	respiratory/oral	New Zealand	[67]
27	OTG4	vagina/vulva	New Zealand	[67]
28	OTG6	skin/wounds	New Zealand	[67]
29	OTG10	anal	New Zealand	[67]
30	OTG18	respiratory/oral	New Zealand	[67]
31	RIHO2	no info	USA	[56]
32	RIHO5	no info	USA	[56]
33	RIHO30	sterile	USA	[67]
34	SW17C	sterile	USA	[56]
35	VAR1.5Vag	vagina/vulva	USA	[56]
36	VAR1.7VUL	vagina/vulva	USA	[56]
37	W17	respiratory/oral	New Zealand	[67]
38	W53	respiratory/oral	New Zealand	[67]
39	W55	anal	New Zealand	[67]
40	W137B	respiratory/oral	New Zealand	[67]
41	W142	respiratory/oral	New Zealand	[67]
42	YASU709	urine	Malaysia	[56]
43	YSM1	sterile	Malaysia	[56]
44	YSM42	respiratory/oral	Malaysia	[56]
45	YASU63	urine	Malaysia	[56]
46	YASU363	urine	Malaysia	[56]
47	YASU123	urine	Malaysia	[56]

Table 2.3: List of GPG commensal strains used in this study.

No	Strain	Site	Country of origin	Reference
1	CFR2.1vul	vulva	USA	[68]
2	CPR2.2fec	stool	USA	[68]
3	CFR2.3vul	vulva	USA	[68]
4	CFR2.4vul	vulva	USA	[68]
5	CFR2.8vag	vagina	USA	[68]
6	CFR2.9vag	vagina	USA	[68]
7	CFR2.10vul	vulva	USA	[68]
8	COD21	respiratory/oral	UK	[68]
9	HMHC1	respiratory/oral	USA	[69]
10	HMHC2	respiratory/oral	USA	[69]
11	HMHC4	respiratory/oral	USA	[69]
12	HMHC5	respiratory/oral	USA	[69]
13	HMHC6	respiratory/oral	USA	[69]
14	HMHC9	respiratory/oral	USA	[69]
15	HP2bt	respiratory/oral	USA	[69]
16	HP3ch	respiratory/oral	USA	[69]
17	HP10bt	respiratory/oral	USA	[69]
18	HP11an	anal	USA	[68]
19	HP11vw	vagina	USA	[69]
20	HP12bt	respiratory/oral	USA	[69]
21	HP13vw	vagina	USA	[69]
22	HP31an	anal	USA	[69]
23	HP31ch	respiratory/oral	USA	[69]
24	HP31vu	vulva	USA	[69]

continued on the next page...

Table 2.3 – continued

No	Strain	Site	Country of origin	Reference
25	HP31vw	vagina	USA	[69]
26	HP33vu	vulva	USA	[69]
27	HP33vw	vagina	USA	[69]
28	HP42bt	respiratory/oral	USA	[69]
29	HP42vp	vagina	USA	[69]
30	HP55bt	respiratory/oral	USA	[69]
31	W104	respiratory/oral	New Zealand	[67]
32	W105	respiratory/oral	New Zealand	[67]
33	w106	respiratory/oral	New Zealand	[67]
34	w107	respiratory/oral	New Zealand	[67]
35	w108	respiratory/oral	New Zealand	[67]
36	w109	respiratory/oral	New Zealand	[67]
37	w111	respiratory/oral	New Zealand	[67]

In addition, 6 strains, strains 3207-3212, which are a series of sequential isolates from an AIDS patient with recurring candidiasis [70] were also used in this study.

2.1.2 Bacterial Strains

The bacterial strain used in this study is *E. coli* DH5. This is used as competent cells for cloning experiments.

2.2 Media and Solutions

2.2.1 Media

YPD Agar (1L):

Yeast extract 5 g, Peptone 10 g, Glucose 10 g, Agar 10 g, ddH_2O 1 L. The medium is sterilized by autoclaving at 121°C and is cooled down to 60°C. The medium is then

poured into sterile petridishes (approximately 25 ml/petridish) to make agar plates. To make agar slants, the medium is poured into bottles with caps (approximately 15 ml/bottle) before sterilization. After sterilization, the bottles are put in room temperature to harden the agar in slanted position. The plates and the slants are put in room temperature for 24 hours before used for yeast inoculation. Uninoculated plates and slants are stored at 4°C until used for microbial inoculation.

YPD broth (1 L):

Yeast extract 5 g, Peptone 10 g, Glucose 10 g, ddH_2O 1 L. The medium is sterilized by autoclaving at 121°C and is cooled to 60°C. The medium is then poured into sterile erlenmeyers or other sterile containers. Unioculated broths are stored at 4°C until used for yeast inoculation.

LB Agar (1 L):

Tryptone 10 g, Yeast Extract 5 g, NaCl 5 g, 1N NaOH 1 ml, Agar 15 g, ddH_2O 1 L. The medium is sterilized by autoclaving at 121°C and is cooled to 60°C. The medium is poured into sterile petridishes to make agar plates. The plates are put in room temperature for 24 hours before used for yeast inoculation. Uninoculated plates are stored at 4°C until used for bacterial inoculation.

LB Agar + Ampicillin (1 L) :

Tryptone 10 g, Yeast Extract 5 g, NaCl 5 g, 1N NaOH 1 ml, Agar 15 g, ddH_2O 1 L. The media is sterilized by autoclaving at 121°C and is cooled to 60°C. Ampicillin 12.5 ml (from 4 mg/ml stock solution) is added to get final concentration of 50 μg/ml. The medium is then poured into sterile petridishes to make agar plates.

LB broth (1L):

Tryptone 10 g, Yeast Extract 5 g, NaCl 5 g, 1N NaOH 1 ml, ddH_2O 1 L. The medium is sterilised by autoclaving at 121°C and is cooled to 60°C. The medium is then poured into sterile erlenmeyers, or other sterile containers. Uninoculated broths are stored at 4°C until used for bacterial inoculation.

LB broth + Ampicillin:

Tryptone 10 g, Yeast Extract 5 g, NaCl 5 g, 1N NaOH 1 ml, ddH_2O 1 L. The medium is sterilized by autoclaving at 121°C and is cooled to 60°C. Ampicillin 12.5 ml (from 4

mg/ml stock solution) is added to get final concentration of 50 μg/ml). The solution is poured into sterile erlenmeyers or other sterile containers.

2.2.2 Solutions

Ampicillin solution 4mg/ml (20 ml):

Ampicillin 100 mg, dd H_2O 20 ml. The solution is sterilized by filter sterilization.

$CaCl_2$ Solution (1L):

PIPE 3 g,$CaCl_2$ 6.66 g, Glycerol 80% 187.5 ml, ddH_2O 812.5 ml. The pH solution is adjusted into 7 by adding NaOH 8 M and then is sterilized by autoclaving.

DNA ladder 1 kb+ 40ng/μl (500 μl):

Distilled water 400 μl, loading dye 80 μl, 1kb+DNA ladder 20 μl.

EDTA 0.5M (1L):

NaEDTA2 H_2O 186 g, ddH_2O 700ml. The pH solution is adjusted into 8.0 with NaOH 10 M(50ml), then ddH_2O is added to the solution to get 1 liter.

Gel loading buffer10X (50 ml):

Ficoll 400 10 g; 0.1M disodium EDTA 0.5 M 10 ml (pH 8), SDS 0.5 g, Bromophenol blue 0.125 g, Xylene cyanol 0.125 g. The mixture is brought to 50 ml with H_2O and then heated at 65°C to dissolve.

Glycerol solution (100 ml):

Glycerol 65 ml, $MgSO_4$ 1M 10 ml, Tris 1M 2.5 ml (pH 8). The solution is autoclaved and stored at room temperature.

IPTG solution 0.1M (50 ml):

IPTG 1.2 g, bring to 50 ml final volume with ddH_2O. The solution is filter-sterilized and stored at 4°C.

X-Gal (2ml):

5-bromo-4-chloro-3-indolyl-β-D-galactoside 100 mg, N,N'- dimethyl-formamide 2 ml. The solution is covered with aluminum foil and stored at -20°C.

TBE 10 (1L)) : Tris 108 g; Boric acid 55 g, EDTA 0.5M (pH 8) 40ml, H_2O is added to bring to 1 L. The mixture is autoclaved and stored in room temperature.

TBE 1x (1L): TBE 10x 100 ml, add H_2O to bring to 1 L and store in room temperature.

Agarose Gel 1% (100 ml):

Agarose 1g, ddH_2O is added to bring to 100 ml, and then put in microwave to dissolve. The solution is then cooled to 50°C and poured to gel electrophoresis apparatus.

2.3 Culture Conditions

2.3.1 Short-term Cultures

A single colony was taken from the stock by a loop and the loop was streaked on an agar plate (YPD agar for yeast and LB agar for bacteria) to get single colonies. The plates were then incubated on 37°C overnight. The plates were then stored at 4°C for about 1-2 months. Single colonies were also used to inoculated liquid media (YPD broth for yeast and LB broth for bacteria).

2.3.2 Long-term Cultures

For long term cultures, two methods are used, slant stock and glycerol stock.

Slant stock:

A loop of cells taken from a single colony of yeast or bacterial cells was streaked on a slant agar in a bottle with a cap, the bottle was then incubated at 37°C overnight. The slants then were stored at 4°C up to 6 months.

Glycerol Stock:

A single colony of yeast or bacterial cells was inoculated into 50 ml liquid media (YPD broth for yeast and LB broth for bacteria) in a 500 ml flask, and then the culture was grown overnight at 37°C with moderate shaking (250 rpm). The overnight culture (0.5 ml) was added into a 2 ml plastic cryovial with a screw cap together with glycerol solution (0.5 ml). The tube was vortexed to mix the two components and then stored at -80°C.

2.4 *E. coli* Compotent Cell Preparation

A single colony of *E. coli* cells (strain DH 5 α) was inoculated into 50 ml of LB medium (500 ml flask), and was grown overnight at 37°C (18 h 10 mins) with moderate shaking (250 rpm). The culture (4 ml) was inoculated into 400 ml LB medium, and was grown at 37°C with shaking at 250 rpm to an OD 590 of 0.375. The culture was then aliquoted into 8 tubes (50-ml prechilled, sterile polypropylene tubes), and the tubes were left on ice for 10 mins. After that the tubes were centrifuged for 7 mins at 1600g (3000 rpm). The supernatants were poured off and each pellet was resuspended in a 10 ml-ice cold $CaCl_2$ solution. The resuspended cells were kept on ice for 30 mins. After centrifugation at 1100g (2500 rpm), 4°C for 5 mins, the supernatants were discarded and each pellet was resuspended in 2 ml ice-cold $CaCl_2$ solution. The tubes were then put on ice in a cold room for 24 hours. After that, the cells were dispensed into prechilled sterile ependrof tubes (250 μl aliquots per tube) and the tubes were freezed immediately at -80°C.

2.5 Primers

Primers used in these studies were designed using MacVector software (MacVector Inc, www.macvector.com). Primers used for sequencing of the whole *YWP1* gene are shown in Table 2.4. All of these primers were purchased from Invitrogen life technologies (Invitrogen New Zealand Limited).

Primers used for PCR amplification and the sequencing of repeat regions of *YWP1*, *HWP1*, and *EAP1* genes are shown in Table 2.5. Primer HWP1F was purchased from Applied Biosystems (Foster City CA, USA), and all other primers were purchased from Invitrogen life technologies (Invitrogen New Zealand Limited).

Table 2.4: Primers for sequencing of the *YWP1* gene.

No	Primer Name	Primer Sequence
1	YWP1F1	CTTTCCTTTTCCTCTTTCCCTCC
2	YWP1R1	GGTGGGGGTATATTGTCTTATGCG
3	YWP1F2	GTTGTTGATTTGGACACTGG
4	YWP1R2	TTTCACCTTGAGTTGGGC
5	YWP1F3	GGTGCTAATGGTGAATCCAC
6	YWP1R3	TAGCACCAGTAGCAGTAGCG
7	YWP1F4	GGTGAACAACATCAACCAGG
8	YWP1R4	CAAGAGTAGAACCTTCAAGAGC
9	M13F	GTAAAACGACGGCCAG
10	M13R	CAGGAAACAGCTATGAC

2.6 PCR

2.6.1 PCR (Gel Electrophoresis)

Amplification was directly performed on *C. albicans* colonies from YPD plates. A single colony was picked using a sterile white pipette tip and cells were then suspended in 20 μl of the reaction mixture including 2 μl 10x PCR buffer (Roche), 2.5 μl 2mM concentration of each deoxynucleoside triphosphate (Roche), 1 μl 10 μmol/l of the nonlabeled reverse and forward primers of each gene (Invitrogen or Applies Biosystems) (see Table 2.5 on page 24), 0.2 μl of 5 units/μl Taq Polymerase (Roche), and 13.3 sterile distilled water. The samples were then initially incubated for 5 mins at 96°C to break the cells and denature the DNA. The temperature cycling (30 cycles at 94°C for 15 s, 51°C for 30 s and 72°C for 1 min) was performed in a thermal cycler. The final cycle was followed by an additional 10 mins at 72°C to complete the polymerization.

2.6.2 Gel Electrophoresis

To make agarose gel, 1 g agarose was mixed with 100 ml 1x TBE buffer. The mixture was heated in a microwave oven to dissolve the agar. The solution was cooled to 50°C before poured into a gel casting unit. After the gel was formed, the comb was

Table 2.5: Primers for PCR amplification and the sequencing of repeat regions of the *YWP1*, *HWP1*, and *EAP1* genes. For sequencing, the forward primers used were without label.

Primer Name	Primer Sequence	Note
YWP1AF	AGTAGTGGTGATTCTGCC	without label
YWP1AR	CCAGCAACGGTATTTTCC	without label
YWP1BF	TCAAGTTCTGCTTCCCCATCG	label with HEX On 5'
YWP1BR	CGTGGACCGTAGTGACACCAATAC	without label
HWP1F	ACAGGTAGACGGTCAAGGTGAAAC	label with NED On 5'
HWP1R	GGAATAGGAGGATTGTCATCAGGC	without label
EAP1AF	TAAGTGTTCGGTTTCATCAG	label with HEX On 5'
EAP1AR	ACATCCACCTTCGGGACAGC	without label
EAP1BF	CCAAATGTGATGGCGGTTC	label with 6-FAM on 5'
EAP1BR	GGTGTACTTGAAGCTGTCTCAGTG	without label

removed and the gel casting unit was put into an electrophoresis tank. Into the tank, 1x TBE buffer was added until the entire gel surface was covered by the buffer. Two μl of 10xgel loading buffer was mixed with the PCR products and then the mixture was loaded into a gel well. Twenty μl of 1KB+ DNA ladder was also loaded into a gel well as a size marker. The gel was run on 90 V until the loading buffer dye reached approximately 75% of the gel length. The gel was removed from the tank, and then was immersed in the ethidium bromide solution (0.5 μg/ml water) for 20 mins. The gel was immersed again in the water for another 20 mins before it was photographed using a Gel Doc UV transilluminator.

2.6.3 PCR (Genotyping)

Amplification was directly performed on *C. albicans* colonies from YPD plates. The colonies were harvested with a single use pipette tip and cells were then suspended in 20 μl of the reaction mixture. For *EAP1*(1) or *EAP1*(2), the reaction mixture included 2 μl 10x PCR buffer (Roche), 2.5 μl 2mM concentration of each deoxynucleoside triphosphate (Roche), 1 μl 10 μmol/l of the nonlabeled reverse and labeled forward primers of the gene (Invitrogen or Applies Biosystems) (see Table 2.5), 0.2 μl of 5 units/μl Taq Polymerase (Roche), and 13.3 sterile distilled water. The samples were

then initially incubated for 5 mins at 96°C to break the cells and denature the DNA. The temperature cycling (30 cycles at 94°C for 15 s, 51°C for 30 s and 72°C for 1 min) was performed in a thermal cycler. The final cycle was followed by an additional 20 mins at 72°C to complete polymerization. For multiplex reaction of *YWP1* and *HWP1*, the reaction mixture included 2 μl 10x PCR buffer (Roche), 2.5 μl 2mM concentration of each deoxynucleoside triphosphate (Roche), 0.5 μl 10 μmol/l of the nonlabel reverse and forward primers of each gene(Invitrogen or Applies Biosystems) (see Table 2.5), 0.2 μl 5 units/μl Taq Polymerase (Roche) and 13.3 sterile distilled water. The samples were then initially incubated for 5 mins at 96°C to break the cells and denature the DNA. The temperature cycling (30 cycles at 94°C for 15 secs, 51°C for 30 s and 72°C for 1 min) was performed in a thermal cycler. The final cycle was followed by an additional 20 mins at 72°C to complete the polymerization.

2.6.4 Genotyping

For genotyping, fluorescent dye 5'-end labeled primers: EAP1(2)-FAM (blue), YWP1-HEX (Green), Eap1(1)-HEX (Green), and HWP1-NED (Yellow) were used. The PCR products were diluted by 10 times and sent to The Allan Wilson Centre (Massey University) for genotyping using a 3730 genetic analyzer. The 1200 LIZ- size standard (Applied Bio systems) was used as the internal standard for fragment sizing. The software Peak Scanner (Applied Bio system was used to analyse the results).

2.7 DNA Sequencing

2.7.1 PCR

The whole gene or repeat regions of the genes were amplified by PCR in 50 μl reaction mixture per strain (see section 2.6.1 on page 23). Ten μl of the PCR product was used to examine the product in a gel electrophoresis (see section 2.6.2 on page 23). The rest of PCR product was purified with high pure PCR product purification kit (Roche) according to the product's instructions. DNA concentration of the PCR products

was measured using Quant-iT TM DS DNA BR Assay Kit and QuBit fluorometer according to the manufacturer's instructions.

2.7.2 Ligation

The purified PCR products were ligated to TA-cloning vector using pLUG-Multi TA-cloning vector kit (Intron) according to the manufacturer's instructions. Ligation mixture (10 μl) included 1 μl TA vector, 1-4 μl PCR product (depend on the concentration of PCR product), 3-6 μl distilled buffer (depend on volume of PCR product), 1 μl 10xligation buffer and 1 μl T4 DNA ligase. The tubes were incubated overnight at 4°C.

2.7.3 Transformation

For each transformation, 100 μl of *E. coli* competent cells were dispensed into the tube containing 10 μl of ligation mixture. Another 100 μl of *E. coli* competent cells were dispensed into another tube for negative control. The tubes were swirled gently and placed on ice for 10 mins. The cells were then heat -shocked by placing the tubes into a 42°C water bath for 2 mins. After that, the tubes were put immediately on ice for 2 mins. One ml of LB medium was added to each tube and the tubes then were placed on a shaker at 250 rpm for 1 h at 37°C.

2.7.4 Selection of Transformed Cells

On the surface of LB-Amp plates, 100 μl of 100 mM IPTG and 20 μl of 50 mg/μl X-Gal were spread and dried at 30°C for 15 mins. The transformed cells (100 μl) were then spread on the plate surface and the plates were then incubated at 37°C for 16 hours. PCRs were conducted to amplify the insert region using white colonies and M13F or M13R. The PCR reaction and PCR programme were the same as explained above except the reaction mixture was initially incubated for 3 mins at 96°C. Ten μl of the PCR product were separated on 1% agarose gel to identify the colonies that had the right insert size. The rest of the PCR products (40 μl) of the colonies

with the right insert size were purified using high pure PCR product purification kit (Roche) according to the product's instructions and then the DNA concentrations were measured using QuBit fluorometer according to the manufacturer's instructions.

2.7.5 Sample Preparation for Sequencing

The DNA sequencing reaction (15 μl) included the DNA prepared above (2-10 μl), 1 pmol/ μl forward primer or reverse primer of each gene (3.2 μl), and distilled water (adjusted). The reaction mixtures were then sent to The Allan Wilson Centre (Massey University) for analysis on the ABI3730 DNA Analyser.

2.7.6 Sequencing Data Analysis

The sequencing results received from The Allan Wilson Centre were analysed using MacVector software.

2.8 Preparations and Analysis of Isolates Derived from Repeated Sub-Culture (300 Generations) of A singe Parental Strains of *C. Albicans*

A single colony of strain RIHO 30 taken from a YPD overnight-grown plate culture was inoculated into a tube containing 2 ml of sterile YPD medium. The culture was grown at 37°C for 12 hours. The cultures (10 μl) were then inoculated into another tube containing 2 ml of sterile YPD medium and the tube was incubated for 12 hours or 24 hours. Ten μl of the cultures were then transferred again into a new tube containing 2 ml of sterile YPD medium. The process was repeated until the cells reached 300 generations. After the incubation time, the optical density of the culture was measured using spcectrophotometer (Nova Tech.) at 600 nm to determine the number of cell generations. For the first 8 transfers, after the incubation time the optical density of the culture were measured using spcectrophotometer (Nova

Tech.) at 600 nm to determine the number of cell generations. The data were then used to calculate the number of transfers required for reaching 300 generations. Two hundreds μl of the last culture were spread on a the surface of YPD agar plate, and the plates were incubated at 37°C overnight. The cells grown on the agar surface were then transferred into a tube containing 5 ml of distilled water. The culture was diluted and spread on three YPD agar plates with approximately 100 cells per agar. The plates were incubated at 37 °C overnight and 60 single colonies were randomly chosen to determine allele size of the repeat regions of *YWP1*, *HWP1* and *EAP1* by genotyping. The 300 generation samples were prepared by Zhuo Zhou [52].

2.9 Statistical Analysis

Statistical analysis used in this study include the Chi-square tests and the t test. There are 2 Chi-square tests used: the Chi-square goodness of fit test, and the Chi-square test for contingency tables. The Chi-square goodness of fit test was used to check whether particular allele combinations or an alleles predominated in a group of strains compared to other allele combinations or alleles in the same group. Thus the test checks whether the frequency of a certain allele combinations or alleles is significantly higher than the observed frequencies of the other allele combinations or alleles. The Chi-square test for contingency tables was used to check whether the predominant allele combinations or alleles in a group of strains overrepresents compared to those particular allele combinations or alleles in the other group of strains. Thus it checks whether the frequency of predominant allele combinations or alleles in a group of strains is significantly higher than those particular allele combinations or alleles in the other group of strains. The t-test was used to check the significant difference between two means of the differences in the number of repeat units in the two alleles in an individual strain. The procedure of statistical analysis in this study follows the standard statistics textbook written by Sheskin [71].

Chapter 3

Alleles of the *YWP1* Gene

This chapter describes the results of the allelic characterization of the *YWP1* gene in strains of interest, which is aimed to determine whether *YWP1* is a contingency gene, i.e. whether *YWP1* has a role in adaptation by changing the number of repeat units within the coding sequences. For this purpose, the allelic distribution of *YWP1* in GPG and non-GPG strains, in commensal and infection strains, and in strains isolated from different sites of the humans body were compared one to another. For a contingency gene, GPG and non-GPG strains, two groups of strains with different genetic backgrounds, should select the same alleles, which are advantageous alleles. A comparison of the allelic distributions of the *YWP1* gene in commensal and infection strains, and in strains isolated from different sites of the humans body, respectively, is to observe whether *YWP1* acts as a contingency gene, when the *C. albicans* state changes from commensal to pathogenic, and when it moves to particular sites of the humans body.

The chapter begins with a description of the identification of the repeat units in *YWP1*, where two regions in the gene were identified to contain repeat units. This is followed by two sections containing the results of the characterization of the *YWP1* alleles for each repeat region. In the first section, since there was no variability in the number of repeat units observed, the analysis was not continued. The second section contains the results of the allelic characterization of *YWP1* for GPG and non-GPG strains of infection strains, two groups of strains with different genetic

backgrounds, and the results of the allelic characterization for the commensal strains, for comparison with the infection strains. Analysis of the diversity of the alleles and allele combinations, and analysis of combination of the two alleles in an individual strain for all strains of interest are explained before the description of the results of the allelic characterization of *YWP1* for strains isolated from different sites of the humans body. The chapter ends with a discussion of the biological implications of the results.

3.1 Identification of the Repeat Units in the *YWP1* Gene

The repeat units in *YWP1* were identified using MacVector software (MacVector Inc, www.macvector.com) to analyse the DNA sequence of *YWP1* from strain SC5314, downloaded from www.candidagenome.org. In MacVector, the nucleotide chain of a particular gene was arranged from the top to the bottom of the vertical axis, and from the left to the right of the horizontal axis starting from the 5' end. This forms a square matrix, and each edge repeats the length of the *YWP1* sequence. With this method, there are dots forming a diagonal line from top left to the bottom right. The line is formed by dots which are a consequence of the fact that at those positions, the vertical and horizontal axis refers to the same nucleotide. The presence of a symmetric pattern of dots in the matrix distinct from the diagonal line indicates regions containing repeat units. The result of the DNA sequence analysis of *YWP1* from strain SC5314 is shown in Figure 3.1.

Primers YWP1AF and YWP1AR (see Table 2.5 on page 24) were designed to amplify the region from nucleotide 607 - 1424 (818 bp), which is a region containing repeat units, and for convenience is called repeat region 1. Primers YWP1BF and YWP1BR (see Table 2.5 on page 24) were designed to amplify the region from nucleotide 493 - 766 (274 bp), which is a region containing repeat units, and for convenience is called repeat region 2.

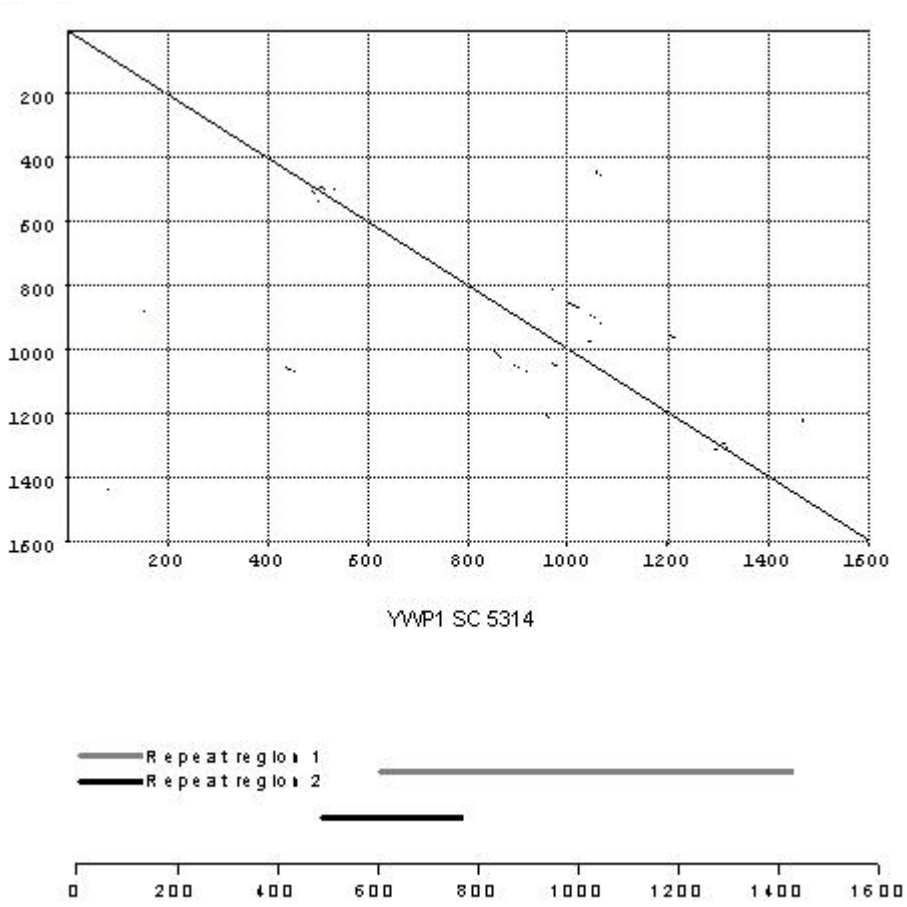

Figure 3.1: Identification of the repeat unit in *YWP1* using MacVector software (top figure). The identification was performed on the DNA sequence of *YWP1* from strain SC5314 downloaded from www.candidagenome.org. The grey and black thick lines indicate regions amplified, i.e. nucleotide 607 - 1424 (818 bp) for region 1 and nucleotide 493 - 766 (274 bp) for region 2 (bottom figure).

3.2 Results of Allelic Characterization of the *YWP1* Gene For Repeat Region 1

Repeat region 1 of the *YWP1* gene from each strain of interest was amplified by PCR, and the amplicons were characterized by length using gel electrophoresis. Later, genotyping replaced gel electrophoresis as the characterization method of choice.

3.2.1 Allelic Characterization of the *YWP1* Gene For Repeat Region 1 in GPG and non-GPG strains (Infection Strains)

Alleles of the *YWP1* gene for repeat region 1 in two groups of infection strains, GPG and non-GPG strains, were characterized. For convenience, in the rest of the text, GPG and non-GPG infection strains will be referred to as GPG and non-GPG strains.

Results of fragment separation using gel electrophoresis for repeat region 1 of *YWP1* in GPG strains showed that all 20 strains had allele combinations containing 850 and 850 bp fragment length. Figure 3.2.a shows the gel electrophoresis fragment separation of repeat region 1 of *YWP1* for some GPG strains. Identical results were observed for 33 non-GPG strains tested. Figure 3.2.b shows the gel electrophoresis fragment separation of repeat region 1 of *YWP1* for some non-GPG strains. The results indicate that there is no repeat variability in repeat region 1 of *YWP1* in infection strains.

3.2.2 Allelic Characterization of the *YWP1* Gene For Repeat Region 1 in Commensal Strains

As GPG strains have similar genetic background, only GPG commensal strains were characterized to be compared to GPG infection strains. For convenience, in the rest of the text, when comparing commensal and infection strains, GPG commensal strains will be referred to as commensal strains, and GPG infection strains will be referred to as infection strains.

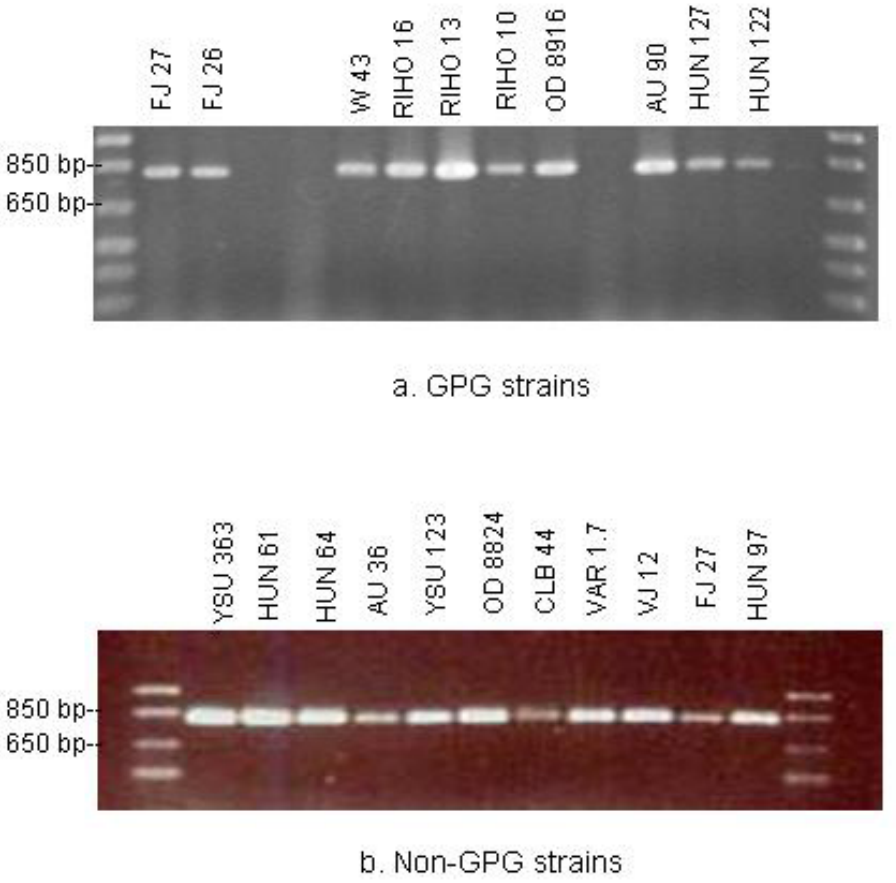

Figure 3.2: The gel electrophoresis fragment separation of repeat region 1 of *YWP1* for: a. GPG strains, and b. non-GPG strains. All strains show the same fragment size (850 bp).

Results of fragment separation using gel electrophoresis for repeat region 1 of *YWP1* in commensal strains showed that 19 of 20 strains had allele combinations containing 850 and 850 bp fragment length, i.e. the same allele combination observed in infection strains. One strain (HMHC8) had an allele combination containing 600 and 850 bp fragment length. Figure 3.3 shows the gel electrophoresis fragment separation of repeat region 1 of *YWP1* for some commensal strains.

3.2.3 No Variability in Alleles of the *YWP1* Gene For Repeat Region 1

Since almost all strains have allele combinations containing 850-850 bp fragment length, the *YWP1* whole gene of four strains was sequenced to check the accuracy of the gel electrophoresis results. The strains were HUN127, AU90, OD8824, and Var 1.7. Eight primers (see Table 2.4 on page 23) were designed for the sequencing. The results showed that all strains have the same nucleic acid sequence on repeat region 1. This indicates that repeat region 1 of the *YWP1* gene is constant, i.e. there is no repeat variability, and no allele diversity.

As indicated previously, there is a second repeat region in the gene between nucleotides 493 and 766 in SC5314 (see Figure 3.1 on page 31). Figure 3.4.a shows the results of the sequencing of *YWP1* in the region containing repeat region 2 and Figure 3.4.b shows the associated amino acid sequences. Based on the former figure, it was determined that the repeat unit in repeat region 2 is GTTCTG. Thus, SC5314, HUN127, and AU90 have 5 repeat units, WO1 and Var 1.7 have 8 repeat units, and OD8824 has 10 repeat units. Hence, there is a variability in the number of repeat units for repeat region 2 of *YWP1*.

Since there is no repeat variability in repeat region 1 of *YWP1*, this chapter will focus more on repeat region 2. The results of the allelic characterization of *YWP1* for repeat region 2 will be described in the following sections, and for convenience, alleles of *YWP1* for repeat region 2 will be referred to as alleles of *YWP1*.

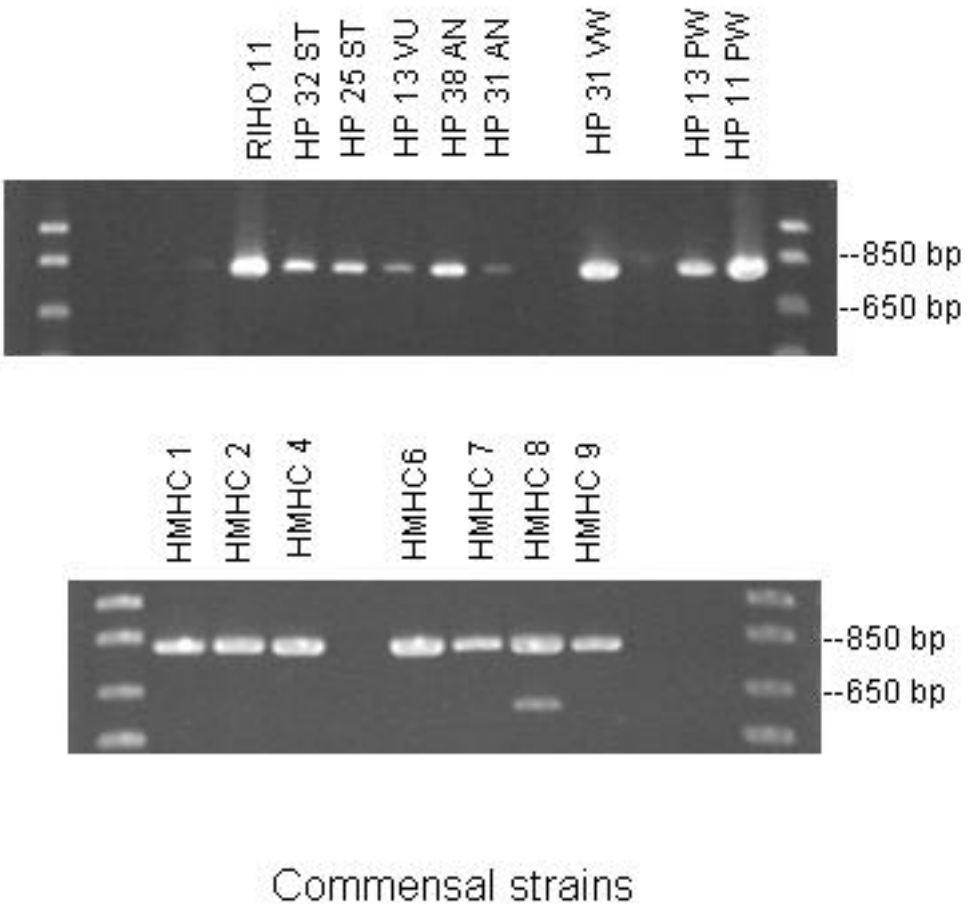

Figure 3.3: The gel electrophoresis fragment separation of repeat region 1 of *YWP1* for some commensal strains. All strains show the same fragment size (850 bp), the only exception being strain HMHC8 with an allele combination containing 600 and 850 bp.

```
YWP1 SC 5314   451 CTTGAAGGTTCTACTCTTGAAGTTGTTGATTACGTTCCAGGTTCAAGTTC  500
YWP1 HUN127    451 CTTGAAGGTTCTACTCTTGAAGTTGTTGATTACGTTCCAGGTTCAAGTTC  500
YWP1 AU 90     451 CTTGAAGGTTCTACTCTTGAAGTTGTTGATTACGTTCCAGGTTCAARTTC  500
YWP1 W01       451 ATTGAAGGTTCTACTCTTGAAGTTGTTGATTACGTTCCAGGTTCAAGTTC  500
YWP1 OD8824    451 ATTGAAGGTTCTACTCTTGAAGTTGTTGATTACGTTCCAGGTTCAAGTTC  500
YWP1 Var 1.7   451 ATTGAAGGTTCTACTCTTGAAGTTGTTGATTACGTTCCAGGTTCAAGTTC  500
                    ****************************************  **** ***

YWP1 SC 5314   501 TGCTTCCCCATCGGGTTCTGCTTCTCCATCTGGTTCCGAATCCGGTTCTG  550
YWP1 HUN127    501 TGCTTCCCCATCGGGTTCTGCTTCTCCATCTGGTTCCGAATTCGGTTTTG  550
YWP1 AU 90     501 TGCTTCCCCATCGGGTTCTGCTTCTCCATCTGGTTCCGAATTCGGTTTTG  550
YWP1 W01       501 TGCTTCCCCATCGGGTTCTGCTTCTCCATCTGGTTCCGAATCCGGTTCTG  550
YWP1 OD8824    501 TGCTTCCCCATCGGGTTCTGCTTCTCCATCTGGTTCCGAATCCGGTTCTG  550
YWP1 Var 1.7   501 TGCTTCCCCATCGGGTTCTGCTTCTCCATCTGGTTCCGAATCCGGTTCTG  550
                   ********************************* ******** ***** **

YWP1 SC 5314   551 GTTCTG ------------------------------ ATTCTGCTACTATC  570
YWP1 HUN127    551 GTTCTG ------------------------------ ATTCTGCTACTATC  570
YWP1 AU 90     551 GTTCTG ------------------------------ ATTCTGCTACTATC  570
YWP1 W01       551 GTTCTGGTTCTGGTTCTAGTTCTG ------------ATTCTGCTACTATC  588
YWP1 OD8824    551 GTTCTGGTTCTGGTTCTGGTTCTGGTTCTGGTTCTGATTCTGCTACTATC  600
YWP1 Var 1.7   551 GTTCTGGTTCTGGTTCTGGTTCTG ------------ATTCTGCTACTATC  588
```

(a)

```
prot YWP1 SC531 151 LEGSTLEVVDYVPGSSSASPSGSASPSGSESGSGS————————DSATI 190
prot YWP1 HUN12 151 LEGSTLEVVDYVPGSSSASPSGSASPSGSESGSGS————————DSATI 190
Prot YWP1 AU 90 151 LEGSTLEVVDYVPGSSSASPSGSASPSGSESGSGS————————DSATI 190
prot YWP1 WO1   151 IEGSTLEVVDYVPGSSSASPSGSASPSGSESGSGSGSGSSS——DSATI 196
protYWP1 OD 882 151 LEGSTLEVVDYVPGSSSASPSGSASPSGSESGSGSGSGSGSGSGSDSATI 200
prot ywp1 Var 1 151 IEGSTLEVVDYVPGSSSASPSGSASPSGSESGSGSGSGS——DSATI 196
                     ************************************        *****
```

(b)

Figure 3.4: Results of the sequencing of the *YWP1* gene in the area containing repeat region 2 (a) and the associated amino acid sequences (b).

3.3 Results of Allelic Characterization of the *YWP1* Gene

3.3.1 Allelic Characterization of the *YWP1* Gene in GPG and non-GPG Strains (Infection Strains)

Alleles of the *YWP1* gene in 49 GPG strains and 47 non-GPG strain were characterized. Since *C. albicans* is diploid, the results are described for both alleles and allele combinations.

Based on allele length, three different *YWP1* alleles were found in GPG strains, and eight in non-GPG strains. In total there were eight different alleles found in infection strains. For the purpose of this study, each allele was characterized by the number of repeat units it contained as well as by the fragment length. In order to determine the number of repeat units, repeat region 2 of *YWP1* of seven strains having seven different alleles observed was sequenced. Primers YWP1BF and YWP1BR (see Table 2.5 on page 24) were used for sequencing. The repeat unit identified by DNA sequencing was GTTCTG which encodes the amino acid sequence SG; this is consistent with the results of sequencing of the *YWP1* gene for four strains shown previously. The frequency of alleles of *YWP1* found in GPG and non-GPG strains is summarized in Table 3.1, and presented graphically in Figure 3.5.

Figure 3.5 shows that *YWP1* alleles containing 5 repeat units predominated in GPG strains, i.e. the frequency of alleles containing 5 repeat units is significantly higher than the combined observed frequency of other alleles (the Chi-square goodness of fit test, $p < 0.001$). Likewise, *YWP1* alleles containing 5 and 8 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). Alleles containing 5 repeat units were overrepresented in GPG strains (95%), compared to non-GPG strains (38%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of *YWP1* alleles containing 5 repeat units between GPG and non-GPG strains ($p < 0.001$). In addition, *YWP1* alleles containing 8 repeat units were overrepresented in non-GPG strains (38%), compared to GPG

Table 3.1: Frequency of alleles of *YWP1* in GPG and non-GPG strains. Alleles are represented by the fragment length and the number of repeat units.

No	Allele		Frequency		
	Length (bp)	No. of repeats	GPG strains	Non-GPG strains	Total
1	275	5	93	36	129
2	287	7	0	7	7
3	293	8	4	36	40
4	299	9	1	1	2
5	305	10	0	10	10
6	317	12	0	2	2
7	329	14	0	1	1
8	335	15	0	1	1
Total			98	94	192

strains (4%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of *YWP1* alleles containing 8 repeat units between GPG and non-GPG strains ($p < 0.001$). Similar results were also observed in strains isolated from oral sites as shown in Figure 3.6. These results show that GPG and non-GPG strains have the same predominant alleles, i.e. alleles containing 5 repeat units, but non-GPG strains have additional predominant alleles, i.e. alleles containing 8 repeat units.

The number of different allele combinations of *YWP1* found in GPG strains was four, compared to 12 combinations in non-GPG strains. In total there were 13 different allele combinations of *YWP1* found in infection strains. The frequency of allele combinations of *YWP1* found in GPG and non-GPG strains is summarised in Table 3.2, and presented graphically in Figure 3.7.

Figures 3.7 shows that allele combinations of *YWP1* containing 5 and 5 repeat units predominated in GPG strains, i.e. the frequency of allele combinations containing 5 and 5 repeat units in GPG strains is significantly higher than the combined observed frequency of other allele combinations (the Chi-square goodness of fit test, $p < 0.001$). Likewise, allele combinations of *YWP1* containing 5 and 8 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). Allele combinations containing 5 and 5 repeat units were overrepresented in GPG strains

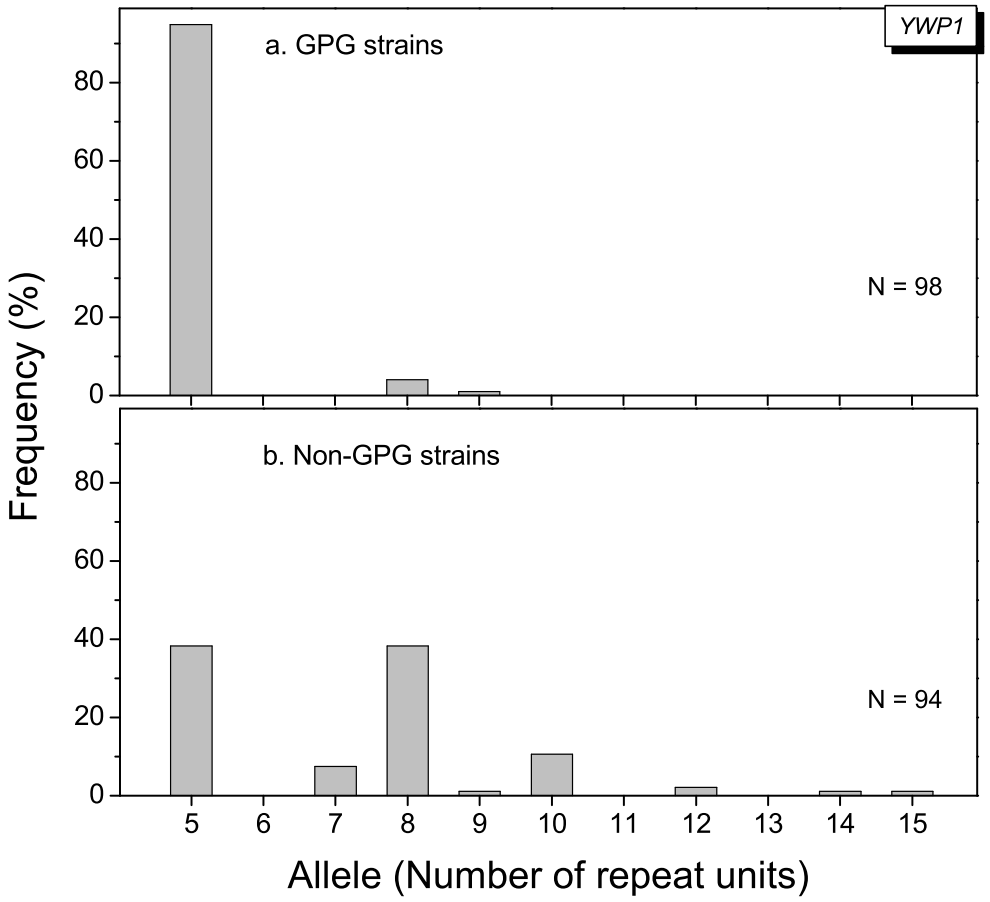

Figure 3.5: Distribution of alleles of *YWP1* for: a. GPG strains, and b. non-GPG strains. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 5 repeat units predominated in GPG strains, and alleles containing 5 and 8 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). Alleles containing 5 repeat units were overrepresented in GPG strains compared to non-GPG strains, and alleles containing 8 repeat units were overrepresented in non-GPG strains compared to GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

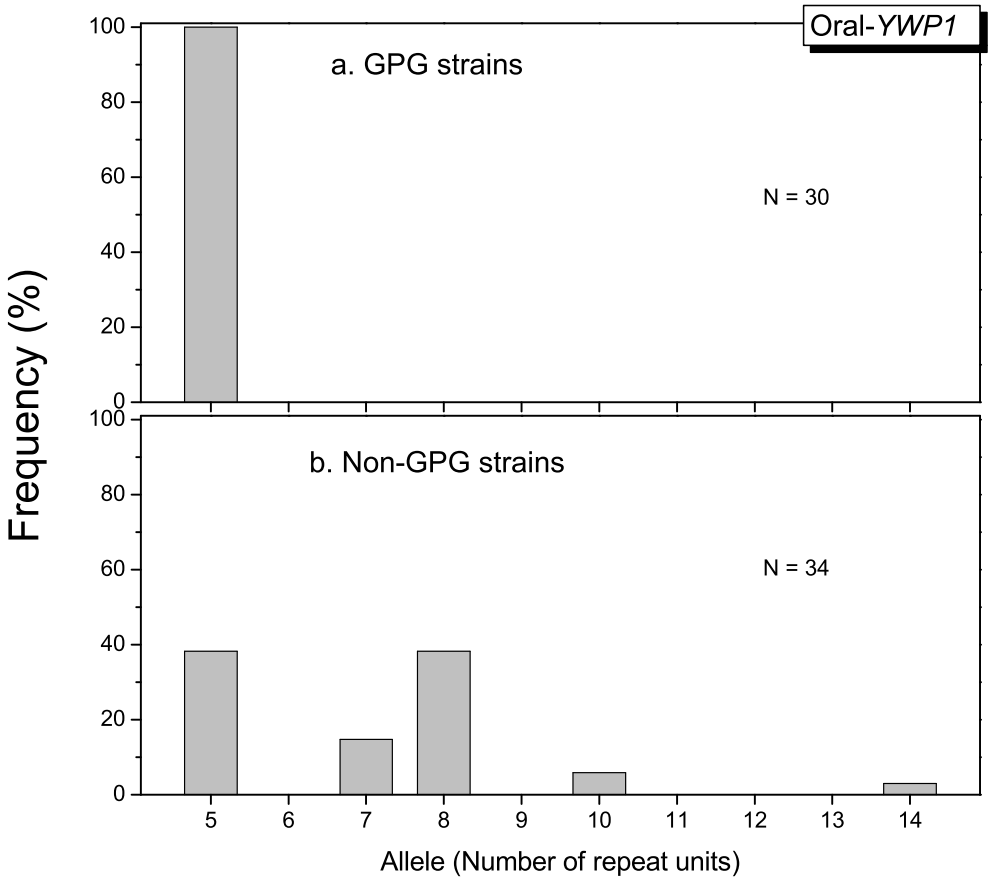

Figure 3.6: Distribution of alleles of *YWP1* for: a. GPG strains, and b. non-GPG strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 5 repeat units predominated in GPG strains, and alleles containing 5 and 8 repeat units predominated in non-GPG strains (The Chi-square goodness of fit test, $p < 0.001$). Alleles containing 5 repeat units were overrepresented in GPG strains compared to non-GPG strains, and alleles containing 8 repeat units were overrepresented in non-GPG strains compared to GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

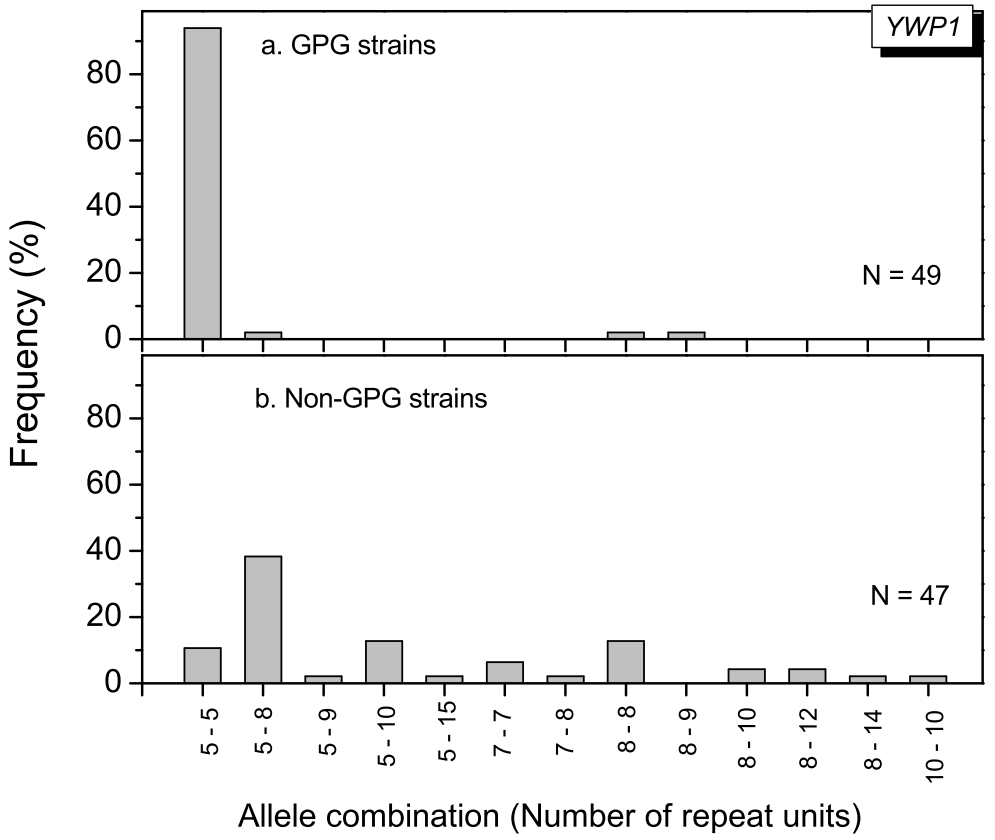

Figure 3.7: Distribution of allele combinations of *YWP1* for: a. GPG strains, and b. non-GPG strains. N represents the total number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing containing 5 and 5 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). Likewise, allele combinations containing 5 and 8 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). Allele combinations containing 5 and 5 repeat units were overrepresented in GPG strains, compared to non-GPG strains, and allele combinations containing 5 and 8 repeat units were over-represented in non-GPG strains, compared to that in GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

Table 3.2: Frequency of allele combinations of *YWP1* in GPG and non-GPG strains. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

No	Allele Combination		Frequency		
	Length (bp)	No. of repeats	GPG strains	Non-GPG strains	Total
1	275-275	5-5	46	5	51
2	275-293	5-8	1	18	19
3	275-299	5-9	0	1	1
4	275-305	5-10	0	6	6
5	275-335	5-15	0	1	1
6	287-287	7-7	0	3	3
7	287-293	7-8	0	1	1
8	293-293	8-8	1	6	7
9	293-299	8-9	1	0	1
10	293-305	8-10	0	2	2
11	293-317	8-12	0	2	2
12	293-329	8-14	0	1	1
13	305-305	10-10	0	1	1
Total			49	47	96

(94%), compared to non-GPG strains (11%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of these allele combinations between GPG and non-GPG strains ($p < 0.001$). In addition, allele combinations of *YWP1* containing 5 and 8 repeat units were overrepresented in non-GPG strains (42%), compared to GPG strains (2%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of these allele combinations between GPG and non-GPG strains ($p < 0.001$). Similar results were also observed in strains isolated from oral sites as shown in Figure 3.8. These results show that GPG and non-GPG strains have distinct predominant allele combinations of *YWP1*.

3.3.2 Alleles of the *YWP1* Gene in Commensal Strains

Alleles of the *YWP1* gene from 35 commensal strains were characterized by length using genotyping. Based on allele length, two different *YWP1* alleles were found in commensal strains, compared to three found in infection strains. The number of

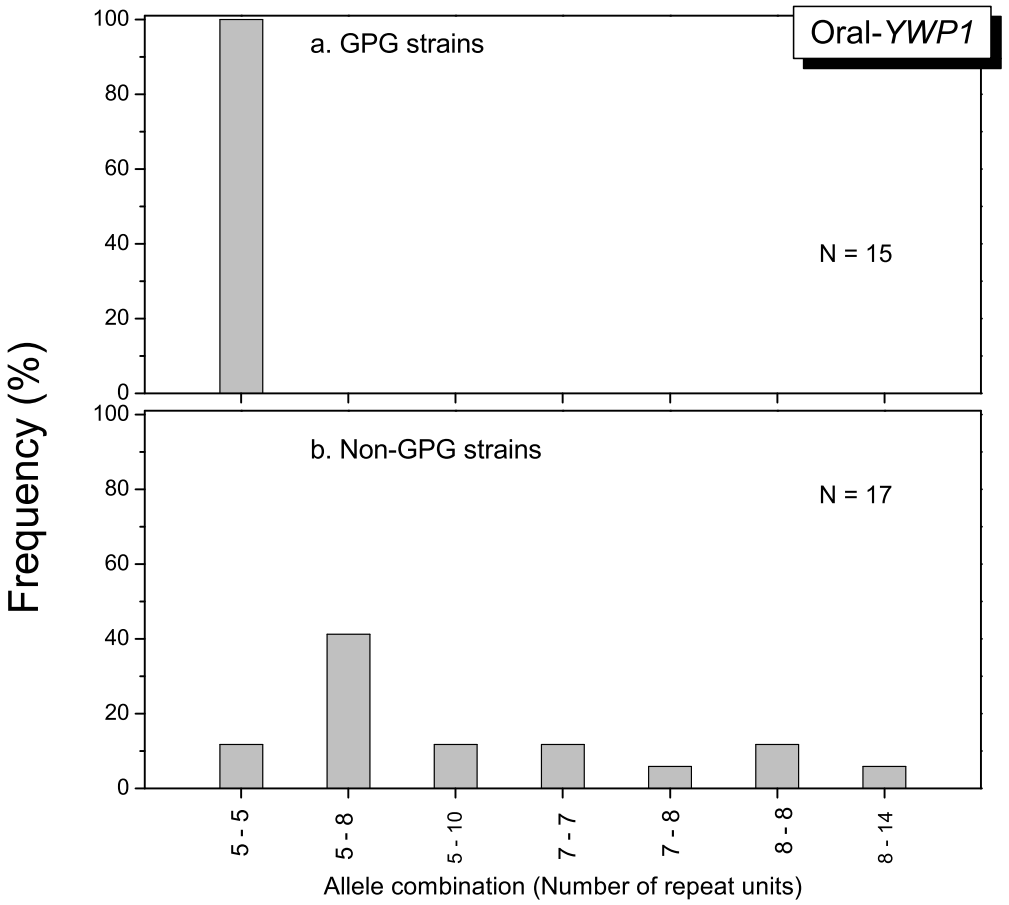

Figure 3.8: Distribution of allele combinations of the *YWP1* gene for: a. GPG strains, and b. non-GPG strains, where all strains were isolated from oral sites. N represents the number of strains in a particular group. All allele combinations of GPG strains contained 5 and 5 repeat units, and alleles containing 5 and 8 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). Allele combinations containing 5 and 5 repeat units were overrepresented in GPG strains compared to non-GPG strains, and allele combinations containing 5 and 8 repeat units were overrepresented in non-GPG strains compared to GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

repeat units was determined using the sequencing results shown previously for the infection strains. The frequency of alleles of *YWP1* found in commensal strains is summarized in Table 3.3, and presented graphically in Figure 3.9. For comparison, the data from infection strains have been added in the table and the figure.

Table 3.3: Frequency of alleles of *YWP1* in commensal strains. Alleles are represented by the fragment length and the number of repeat units. For comparison, the data from infection strains have been added.

No	Allele		Frequency		
	Length (bp)	No. of repeats	Commensal strains	Infection strains	Total
1	275	5	66	93	159
2	293	8	4	4	8
3	299	9	0	1	1
Total			70	98	168

Figure 3.9 shows the same pattern of allele distribution of *YWP1* between commensal and infection strains. Alleles containing 5 repeat units predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$). The percentage of alleles containing 5 repeat units in commensal strains is 94% comparable to the one in infection strains 95%. The result of the Chi-square test for contingency tables showed no significant difference in the distribution of these alleles between commensal and infection strains ($p > 0.05$). Similar results were also observed for strains isolated from oral sites as shown in Figure 3.10. These results show that commensal and infection strains have the same predominant alleles of *YWP1*.

The number of different allele combinations of *YWP1* found in commensal strains was two, compared to four combinations in infection strains. The frequency of allele combinations of *YWP1* found in commensal strains is summarized in Table 3.4, and presented graphically in Figure 3.11. For comparison, the data from infection strains have been added in the table and the figure.

Figures 3.11 shows the same pattern of the distribution of allele combinations between commensal and infection strains. Allele combinations containing 5 and 5 repeat units predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$). The percentage of allele combinations containing 5 and 5 repeat units in

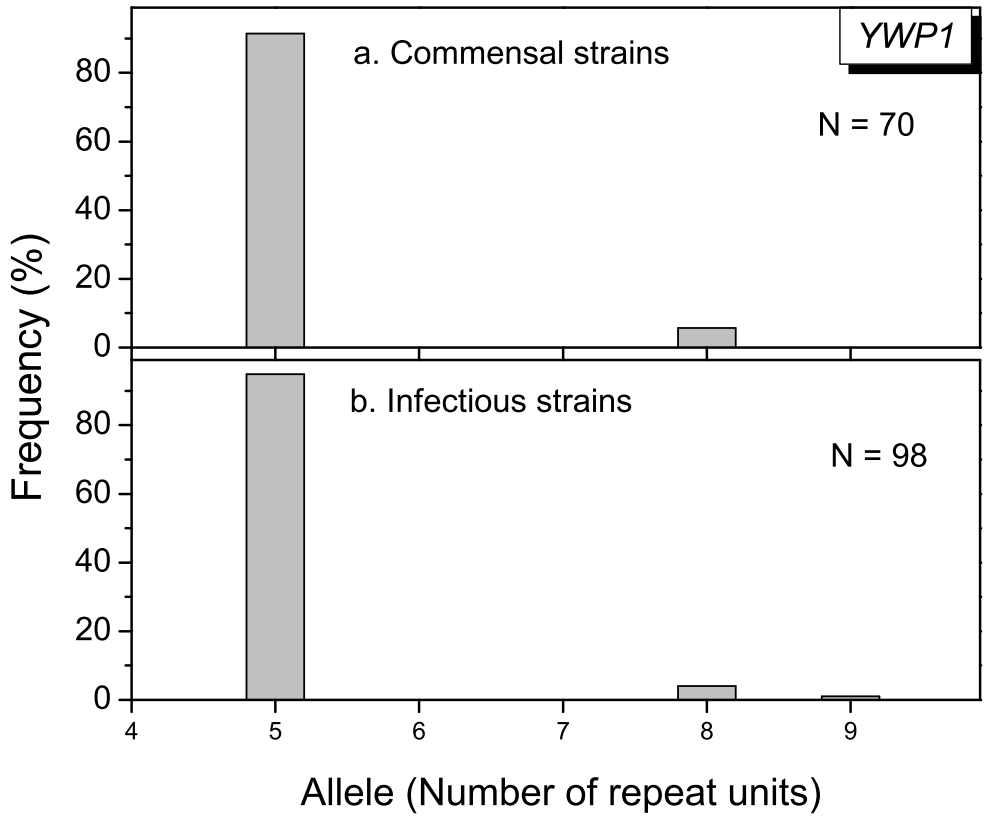

Figure 3.9: Distribution of alleles of *YWP1* for: a. commensal strains, and b. infection strains. N represents 2 times the number of strains in the groups, as there are 2 alleles per strain. Alleles containing 5 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

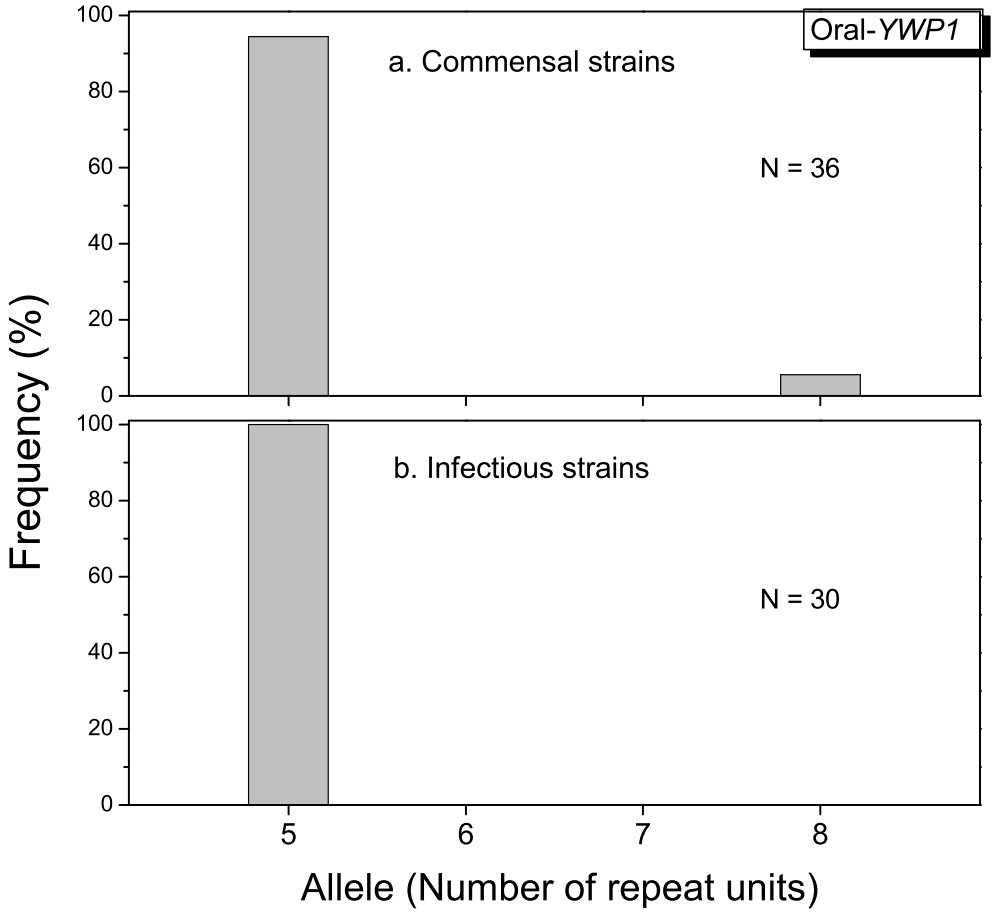

Figure 3.10: Distribution of alleles of *YWP1* for: a. commensal strains, and b. infection strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 5 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

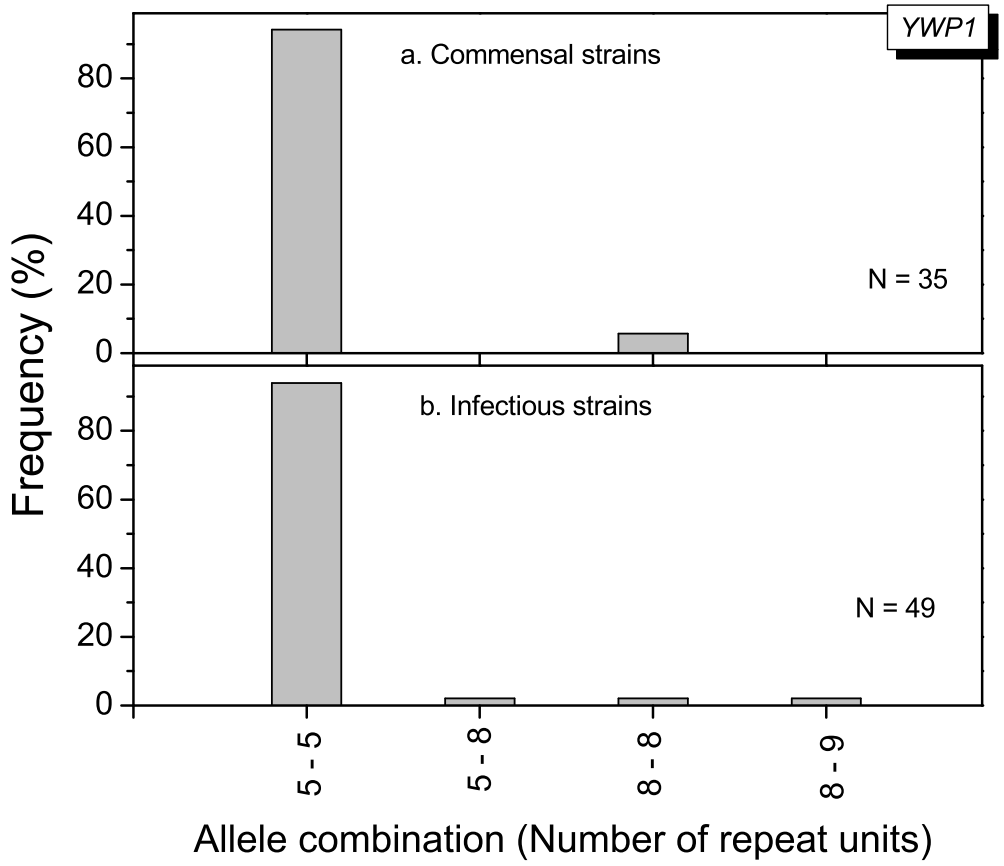

Figure 3.11: Distribution of allele combinations of *YWP1* for: a. commensal strains, and b. infection strains. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 5 and 5 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

Table 3.4: Frequency of allele combinations of *YWP1* in commensal strains. For comparison, the data from infection strains have been added. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeats.

No	Allele Combination		Frequency		
	Length (bp)	No. of repeats	Commensal strains	Infection strains	Total
1	275-275	5-5	33	46	79
2	275-293	5-8	0	1	1
3	293-293	8-8	2	1	3
4	293-299	8-9	0	1	1
Total			35	49	84

commensal strains is 94%, the same as in infection strains. The result from the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between commensal and infection strains ($p > 0.05$). Similar results were also observed for strains isolated from oral sites as shown in Figure 3.12. These results show that commensal and infection strains have the same predominant allele combinations.

3.3.3 Diversity of Alleles of the *YWP1* Gene

In this section, the diversity of alleles and allele combinations of *YWP1* is analysed. The intension was to compare the diversity of GPG and non-GPG strains, and the diversity of commensal and infection strains. A comparison of the allele diversity of GPG and non-GPG strains may provide insights into the influence of the genetic diversity on allele diversity of the gene (GPG strains have smaller genetic diversity than non-GPG strains). A comparison of the allele diversity of commensal and infection strains may provide insights into the influence of the immune status of the host on allele diversity.

The diversity is quantified by the index of diversity λ which is defined as,

$$\lambda = 1 - \frac{1}{N(N-1)} \sum_{i=1}^{k} n_i(n_i - 1), \tag{3.1}$$

where k is the number of different alleles or allele combinations, N is the number

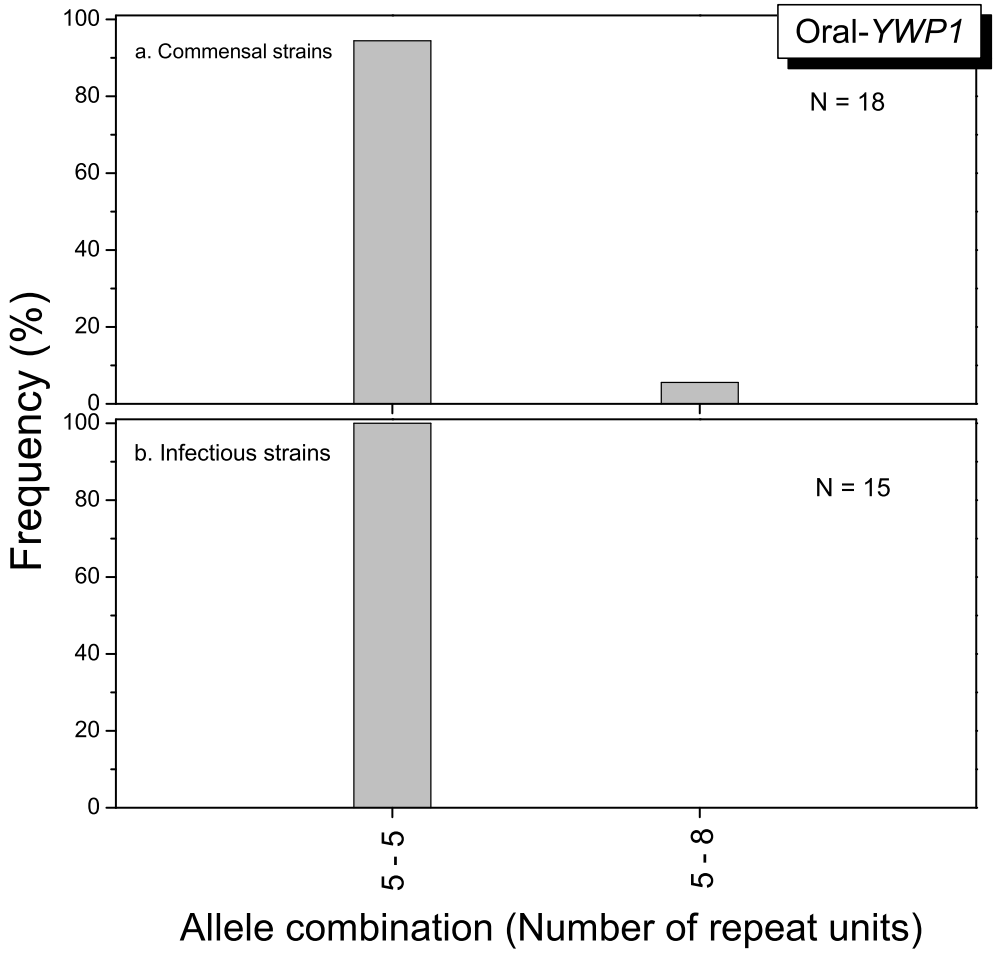

Figure 3.12: Distribution of allele combinations of *YWP1* for: a. commensal strains, and b. infection strains, where all strains were isolated from oral sites. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are number of repeat units. Allele combinations containing 5 and 5 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

of strains, and n_i is the number of strains having ith allele or allele combination. Equation 3.1 is the Simpson's index of diversity [72, 73]. The values of λ range from a minimum of 0 when all strains have the same alleles or allele combinations to 1 when every strain has different alleles or allele combinations. In determining λ for allele combinations, N is considered as the number of strains, while in determining λ for alleles, N is considered as the number of strains multiplied by 2, as there are two alleles per strain.

Table 3.5: Index of diversity λ of the alleles and the allele combinations of *YWP1* for GPG, non-GPG (infection), and (GPG) commensal strains.

Strains	Index of Diversity	
	Allele	Allele Combination
GPG	0.10	0.12
Non-GPG	0.69	0.82
Commensal (GPG)	0.11	0.11

The λ values for both the alleles and the allele combinations of *YWP1* for GPG, non-GPG, and commensal strains are summarized in Table 3.5, and presented graphically in Figure 3.13.

The results show that the diversity of the alleles and the allele combinations is influenced by the genetic diversity of the strains: GPG strains with less genetic diversity have smaller allele diversity than non-GPG strains with large genetic diversity. In contrast, infection and commensal strains have similar diversity, i.e. allele diversity is independent of whether the strains exhibit a commensal or an infection phenotype.

3.3.4 Analysis of the Combination of the Two Alleles of *YWP1* in an Individual Strain

In this section, the combination of the two alleles of *YWP1* in an individual strain among strains of interest is analysed. The section starts with a description of the distribution of the differences in the number of repeats between the two alleles, from

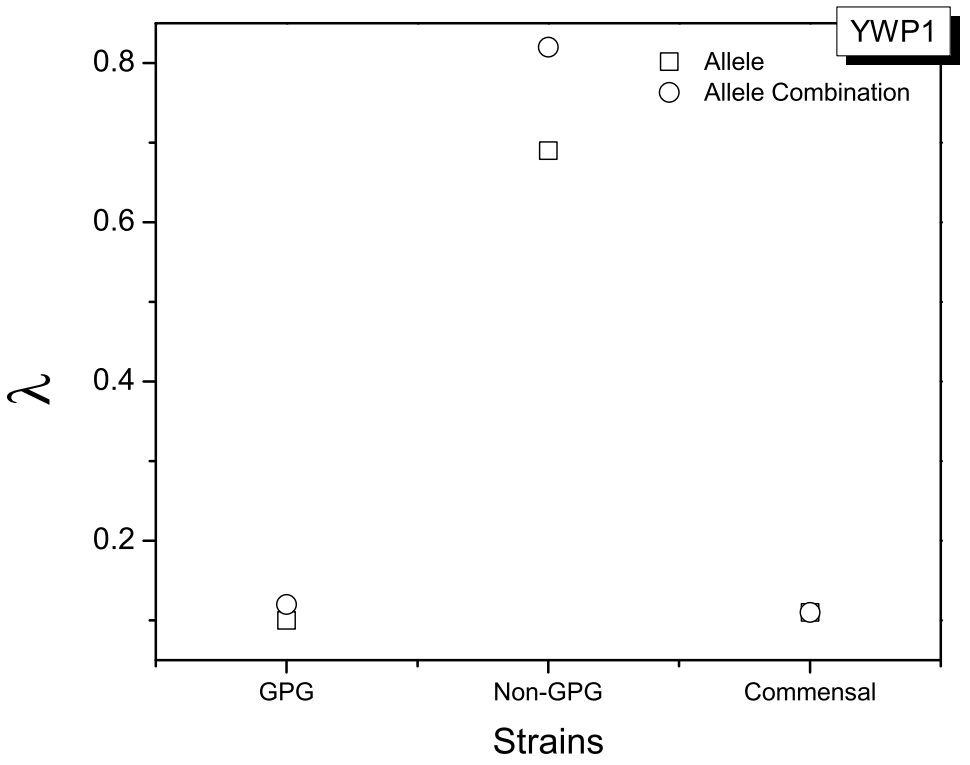

Figure 3.13: Index of diversity λ of the alleles and the allele combinations of *YWP1* for GPG, non-GPG (infection), and (GPG) commensal strains.

which the proportion of homozygous and heterozygous alleles is identified: zero difference denotes homozygous alleles, while a difference indicates heterozygous alleles. The section ends with a calculation of the mean of the differences in the number of repeats between the two alleles in an individual strain among strains of interest.

Figure 3.14 shows the distribution of differences in the number of repeats between the two alleles of *YWP1* in an individual strain for infection and commensal strains. The percentages of homozygous alleles of *YWP1* found in GPG, non-GPG, and commensal strains are 96%, 32%, and 100%, respectively.

The number of homozygous alleles in each group was compared to those expected if the combination of the alleles was random, taking allele frequency into consideration. For this purpose, a matrix with a size equal to the square of 2 times the number of strains analysed was generated. With this scenario, there are 9604 allele combinations of 49 GPG strains, 8836 of 47 non-GPG strains, and 4900 of 35 commensal strains. The number of homozygous and heterozygous alleles from random data was identified, and the Chi-square test for contingency tables was used to determine whether the number of homozygous or heterozygous alleles between the observed and random data is significantly different or not. The frequency of homozygotes and heterozygotes of *YWP1* for the data observed and those expected from random alleles is summarized in Table 3.6, and is presented graphically in Figure 3.15. The figure shows no significant difference between the number of homozygotes observed and those expected by chance for GPG and non-GPG strains (the Chi-square test for contingency tables , $p>0.1$). However, for commensal strains, the number of homozygotes is significantly larger than those expected by chance (the Chi-square test for contingency tables, $p<0.05$), i.e. *YWP1* exhibits significant excesses of homozygotes in commensal strains.

The mean ($\pm$ standard deviation) of the difference in the number of repeats between the two alleles in an individual strain was determined to be 0.08 ± 0.45, 2.49 ± 2.16, and 0 ± 0, for GPG, non-GPG, and commensal strains, respectively. The result of the t-test between the two means of GPG and non-GPG strains showed that the two alleles in a GPG strain are significantly more similar to each other than the alleles in a non-GPG strain ($p < 0.001$). The result of t-test between the two means of

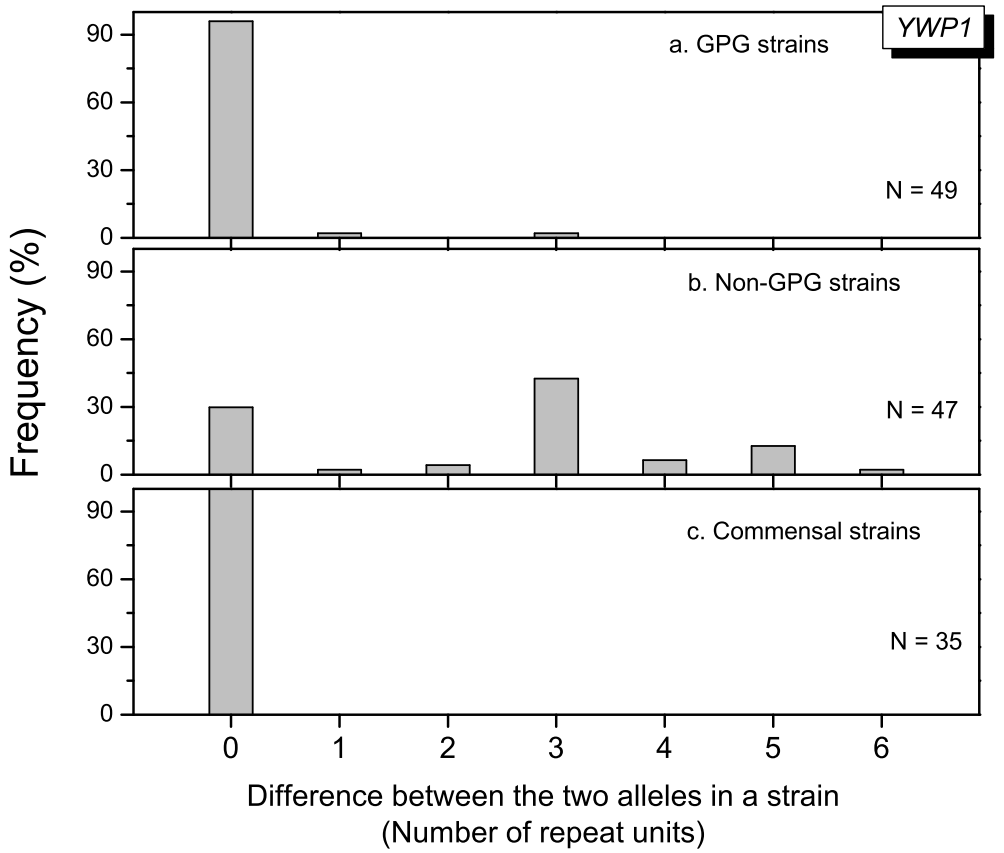

Figure 3.14: Distribution of differences in the number of repeats between the two alleles of *YWP1* in an individual strain for: a. GPG strains, b. Non-GPG strains, and c. Commensal strains. N represents the total number of strains in a particular group.

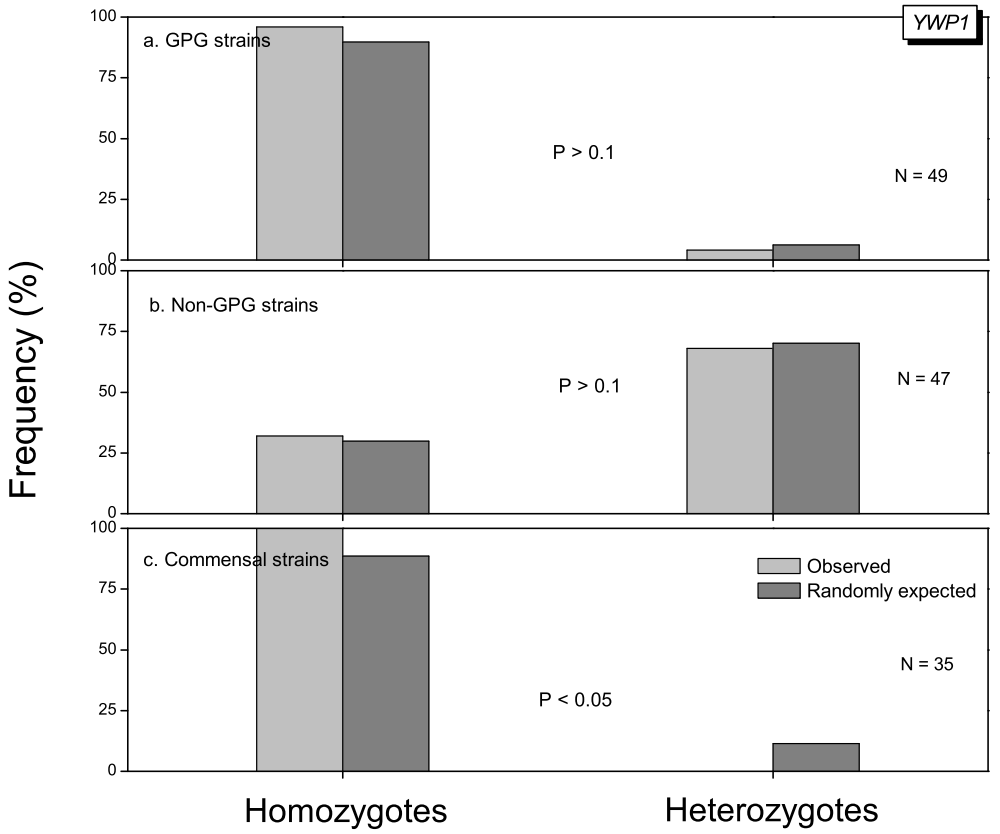

Figure 3.15: Frequency of homozygotes and heterozygotes of *YWP1* for GPG, non-GPG, and commensal strains. N represents the number of strains in a particular group. There is no significant difference between the number of homozygotes observed and those expected by chance (the Chi-square test for contingency tables , $p>0.1$) for GPG and non-GPG strains. However, for commensal strains, the number of homozygotes is significantly larger than those expected by chance (the Chi-square test for contingency tables, $p<0.005$), i.e. *YWP1* exhibits significant excesses of homozygotes in commensal strains.

Table 3.6: The number of homozygous alleles in GPG, non-GPG, and commensal strains compared to the expected number by chance. N indicates the number of data.

Strains	Number of homozygous alleles		Significance (p)
	Observed data (N)	Expected data	
GPG	47 (49)	44	$p > 0.1$
Non-GPG	15 (47)	14	$p > 0.1$
Commensal	35 (35)	31	$p < 0.05$

(GPG) commensal and infection strains showed no significant difference between the two means ($p > 0.1$); the two alleles in a commensal strain did not differ significantly from the two alleles in an infection strain.

To determine whether the difference in the number of repeats between the two alleles in an individual strain is random or non-random, the means of the differences between the two alleles in an individual strain observed among strains of interest were compared to the means expected if all alleles could be combined randomly, taking allele frequency into consideration. As explained at the beginning of this section, a matrix with the size equal to square of 2 times the number of strains analysed was generated.

For GPG strains, the mean (± standard deviation) of the differences expected from random alleles was determined to be 0.31 ± 0.95. The result of t-test between this mean and the mean from observed data of 0.08 ± 0.45 showed that the two alleles in a GPG strain are significantly more similar to each other than expected by chance ($p < 0.05$). For non-GPG strains, the mean (± standard deviation) of the differences expected from random alleles was determined to be 2.23 ± 2.08. The result of t-test between this mean and the mean from observed data of 2.40 ± 1.81 showed that the two alleles in a non-GPG strain did not differ significantly from those expected by chance ($p > 0.10$). For commensal strains, the mean (± standard deviation) of the differences expected from random alleles was determined to be 0.32 ± 0.96. The result of t-test between this mean and the mean from observed data of 0 ± 0 concludes that the two alleles in a commensal strain are significantly more similar to each other than

expected by chance ($p < 0.01$). The mean (± standard deviation) of the differences from observed data, data expected from random alleles, and their significance are summarized in Table 3.7.

Table 3.7: The mean (± standard deviation) of the differences in the number of repeats between the two alleles of the *YWP1* gene in an individual strain from observed data, data expected from random alleles, and their significance. N indicates the number of data.

Strains	Mean Differences (± SD)		Significance (p)
	Observed Differences (N)	Expected Differences (N)	
GPG	0.08 ± 0.45 (49)	0.31 ± 0.95 (9604)	$p < 0.05$
Non-GPG	2.40 ± 1.81 (47)	2.23 ± 2.08 (8836)	$p > 0.10$
Commensal	0 ± 0 (35)	0.32 ± 0.96 (4900)	$p < 0.01$

3.3.5 Allelic Distribution of *YWP1* in Strains Isolated From Different Sites of the Humans Body

In this section, *C. albicans* strains were grouped according to the sites of the humans body they were isolated from, and the allelic distributions of *YWP1* in strains from each site are compared one to another. This comparison is to observe whether particular alleles or allele combinations of the *YWP1* gene might confer selective advantages to *C. albicans* in the various environmental conditions unique to different sites of the humans body.

Figure. 3.16 shows the distribution of alleles of *YWP1* in strains isolated from different sites of the humans body. The distributions of alleles at all sites appear similar; alleles containing 5 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). Result from the Chi-square test for contingency tables showed no significant difference in the distribution of alleles containing 5 repeat units between any two sites for all sites sampled ($p > 0.05$). In addition, alleles containing 8 repeat units existed at all sites.

Figure. 3.17 shows the distribution of allele combinations of *YWP1* in strains isolated from different sites of the humans body. The distributions of allele combinations

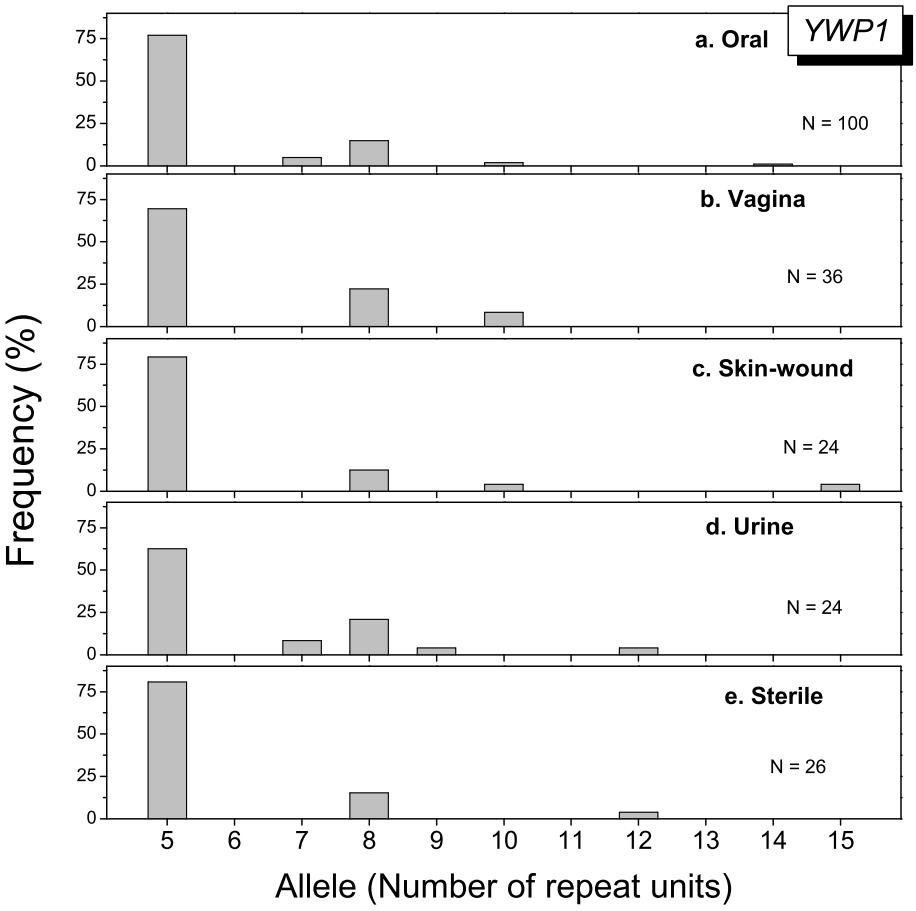

Figure 3.16: Distribution of alleles of *YWP1* in strains isolated from different sites of the humans body. N represents 2 times the number of strains, as there are 2 alleles per strain. Allele containing 5 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of alleles containing 5 repeat units between any two sites for all sites sampled (the Chi-square test for contingency tables, $p > 0.05$). In addition, alleles containing 8 repeat units existed at all sites.

at all sites appear similar; allele combinations containing 5 and 5 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). Result from the Chi-square test for contingency tables showed no significant difference in the distribution of allele combinations containing 5 and 5 repeat units between any two sites for all sites sampled ($p > 0.05$).

These results show that strains isolated from different sites of the humans body have the same predominant alleles and allele combinations, which means that there is no specific allele or allele combination of the *YWP1* gene selected for at specific sites.

3.4 Discussion

The aim of this particular study of the *YWP1* gene was to find out whether *YWP1* is a contingency gene, i.e. whether *YWP1* has a role in adaptation by changing the number of repeat units within the coding sequences. For this purpose, allelic distribution of *YWP1* in GPG and non-GPG strains, in commensal and infection strains, and in strains isolated from different sites of the humans body was examined and compared one with another.

3.4.1 Variability in the number of *YWP1* repeats

This study reveals that there is a variability in the number of *YWP1* repeats observed from different strains. From 96 infection strains, eight different alleles were detected, which formed 13 different allele combinations. Likewise, two different alleles were detected in 35 commensal strains (GPG) which formed two different allele combinations. From total 131 strains, there were eight different alleles detected which form 13 different allele combinations. The allele differences are caused by the addition or removal of a repeat unit or units.

A variability in the number of repeats of other genes has also been reported. The genes include *EF3* [47], *CEK1*, *HYR1*, *HYR2* [31], *Rlm1* [48], *ALS* [49], *ALS7* family [32], *ALS3* [27], *ALS5*, *ALS6* [33, 50], *PNG2* [51], and *SSR1* [52]. In addition to these genes, *HWP1* and *EAP1* are described in the following two chapters. The number of

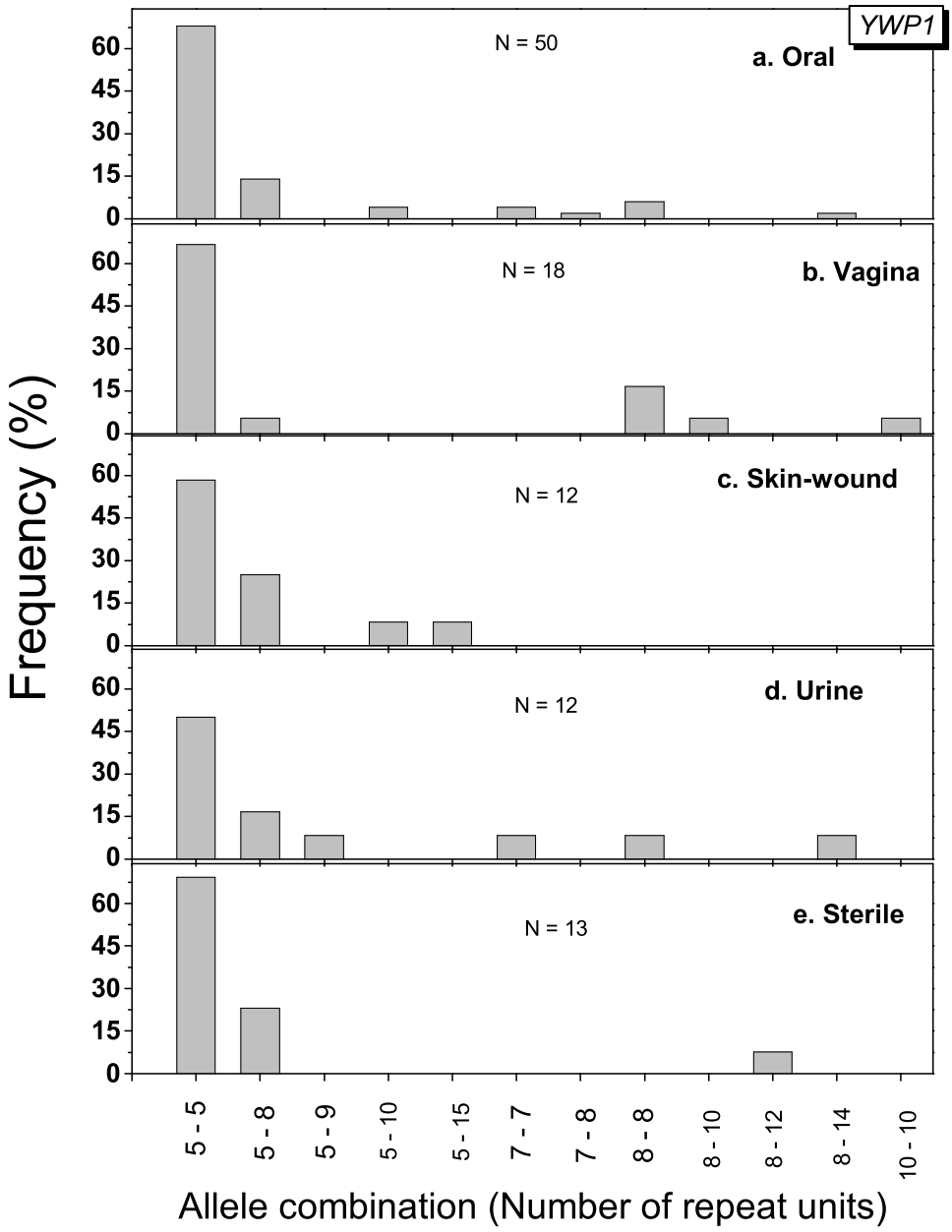

Figure 3.17: Distribution of allele combinations of *YWP1* in strains isolated from different sites of the humans body. N represents the number of strains for a particular site. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 5 and 5 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between any two sites for all sites sampled (The Chi-square test for contingency tables, $p > 0.05$).

different alleles detected varies from gene to gene, and is summarized in Table 3.8. The variation in the number of repeats may relate to the degree to which a gene is involved in adaptation; a pool of alleles is produced in such a way that the most suitable alleles are selected for adaptation purposes [55, 74, 75]. Therefore, the variation in the number of repeats may relate to the rate at which the a gene mutates; the more variation in the alleles, the higher the mutation rate. A previous study revealed a relationship between the diversity in bacterial population and the mutation rate of a contingency gene; increasing the mutation rate three times increased the diversity of the bacterial population by nine times [76].

If the cells are well adapted, most of the diverse alleles resulting from a high mutation rate are likely to be deleterious due to their low fitness, i.e. once the cells are well adapted, the advantages of high mutation rates are lost; this suggests that the well-adapted cells tend to have low mutation rates [77], resulting in less allele diversity. The fact that only 8 different alleles were observed in *YWP1* may indicate that the gene has a relatively low mutation rate, consequently may have a role in adaptation. The relationship between the mutation rate of *YWP1* and its allele diversity will be discussed in Chapter 7, while the role of *YWP1* in adaptation will be discussed in the following section.

Variability in the number of repeats in some of *C. albicans* genes appears to relate to the ability of *C. albicans* to adhere to the host. For example, it has been shown that the number of repeats in *ALS1*, *ALS3*, and *ALS5* correlates to the adherence properties of *C. albicans* [27, 28, 29]. Using non-adherent *S. cerevisiae* cells to express the truncated *ALS1* gene lacking 15 of the 20 repeat units, it was found that the ability of the cell to adhere decreased by 50% [28]. Studies on *ALS3* with either 9 or 12 repeat units showed that proteins with 12 repeat units allow cells to adhere more strongly than those with 9 repeat units [27]. Likewise, using non-adherent *S. cerevisiae* to express Als5p with 0 to 6 repeats also showed an evidence that adherent ability increases in proportion to the number of repeats [29].

In relation to the adherent ability, experimental studies on the *YWP1* gene showed that the product of this gene inhibits the adhesion of the yeast form of *C. albicans*

Table 3.8: Number of different alleles detected in some *C.albicans* genes. The table is arranged from the lowest to the highest numbers of different alleles observed. N indicates the number of strains tested. * represents results from this study, i.e. *YWP1* in this chapter, *HWP1* and *EAP1* in the two following chapters. *EAP1*(1) and *EAP1*(2) indicate repeat regions 1 and 2 of *EAP1* respectively, as there are two repeat regions of *EAP1* analysed.

Gene	# of different alleles (N)	Reference
HWP1	5 (131)	*
ALS6	7 (372)	[33]
YWP1	8 (131)	*
CEF3	8 (60)	[47]
ALS5	9 (277)	[33]
ERK1	12 (113)	[31]
EAP1(1)	14 (131)	*
ALS3	14	[27]
CEF3	16 (112)	[31]
PNG2	17 (80	[51]
ALS1	24 (111)	[31]
SSR1	24 (131)	[52]
EAP1(2)	24 (131)	*
ALS7	60 (66)	[32]

[60]. For example, when alleles of *YWP1* were disrupted, the adherent ability of the yeast cells increased, and when exogenous YWP1p was applied to cells, the adhesion of the cells to plastic is inhibited. It was also found that the expression of *YWP1* is downregulated upon filamentation [60, 61]. It seems that YWP1p may have a role in promoting the dispersal ability of the *C. albicans* in the yeast form. Since the adherent ability of *C. albicans* appears to increase with increasing the number of repeats [27, 28, 29], the observation that alleles containing 5 repeats, the minimum number of *YWP1* repeats observed, predominated in GPG and non-GPG strains, is consistent with the role YWP1p may play in promoting the dispersal ability of the *C. albicans*.

3.4.2 GPG and non-GPG strains have distinct predominant allele combinations, but share the same predominant alleles of *YWP1*

Although GPG and non-GPG strains have distinct predominant allele combinations, they share the same predominant alleles, but additional predominant alleles were observed in non-GPG strains, different from those of GPG strains; allele combinations containing 5 and 5 repeat units predominated in GPG strains, while alleles combinations containing 5 and 8 repeat units predominated in non-GPG strains (see Figure 3.5 on page 39 and Figure 3.7 on page 41).

The percentages of the predominant allele combinations in GPG strains are larger than in non-GPG strains. Consequently, the allele variation shown by the index of diversity is systematically smaller in GPG strains compared to non-GPG strains (see Figure 3.13 on page 51). The difference in the percentages of predominant allele combinations in GPG and non-GPG strains may be due to the genetic diversity of the strains, i.e. GPG strains with smaller genetic diversity (the mean genetic distance between any two isolates 0.19 [56]) have a larger percentage of the predominant allele combinations than non-GPG strains with larger genetic diversity (the mean genetic distance between any two isolates 0.37 [56]).

The results of the predominant allele combinations of other *C. albicans* genes in GPG strains have been shown previously by studies of *ALS7*, *PNG2* and *SSR1*; a particular allele combination was overrepresented in GPG compared to non-GPG strains, 76 vs 25% for *ALS7* [32], 80 vs 8% for *PNG2* [51], and 30 vs 2.2% and 28 vs 0% for *SSR1* [52] (*SSR1* has two predominant allele combinations). However, the predominant allele combinations in non-GPG strains were found only in *YWP1* of this study, and the percentage of the predominant allele combination of *YWP1* of GPG strains (94%) is higher than those of other genes.

It is not yet clear what causes a difference in the predominant allele combinations between GPG and non-GPG strains. Since the two groups have different genetic backgrounds, it is possible that each genetic background in *C. albicans* selects specific

allele combinations leading to the selection of distinct predominant allele combinations in each group. Fortunately, in non-GPG strains, there are two major branches specified as B and C [56], which can be used to observe the impact of genetic background. It was found that 94% of non-GPG strains which possess predominant allele combinations were specified as branch C strains. In fact, allele combinations containing 5 and 8 repeat units predominated in branch C strains, while from a total of nine strains of branch B, three have allele combinations containing 8 and 8 repeat units, and six strains each have different allele combinations. With an awareness that the number of strains in branch B is small, the distribution of the allele combinations of branches B and C strains shown in Figure 3.18 suggests that the predominant allele combinations are caused by a difference in the genetic background. An impact of the genetic background on the predominant allele combinations was also suggested by a previous study on *PNG2*, which found that the genetic background of GPG strains has a strong impact on the selection of the predominant allele combinations [51]. However, the *PNG2* study cannot provide a proof that other genetic backgrounds also select for specific allele combinations, as there were no obvious predominant allele combinations observed in non-GPG strains. In this study of *YWP1*, although the mean distance between any two strains in branches B was 0.33± 0.09 and C was 0.35 ±0.09, larger than 0.19±0.06 of GPG strains, these three major branches of strains each have distinct predominant allele combinations. It seems that the percentage of the predominant alleles and allele combinations is influenced by the genetic diversity of the strains tested, i.e. GPG strains with less genetic diversity have larger percentage than non-GPG strains with higher genetic diversity.

The biological significance of the presence of predominant allele combinations of *YWP1* in non-GPG strains and the absence of predominant allele combinations of other genes mentioned above in non-GPG strains requires more explanation. In Chapter 6, the possibility that the predominant allele combinations may relate to the gene interconnectedness will be discussed, and in Chapter 7, the possibility of the impact of the mutation rate on the predominant allele combinations will be discussed.

Beside distinct predominant allele combinations of *YWP1* in GPG and non-GPG

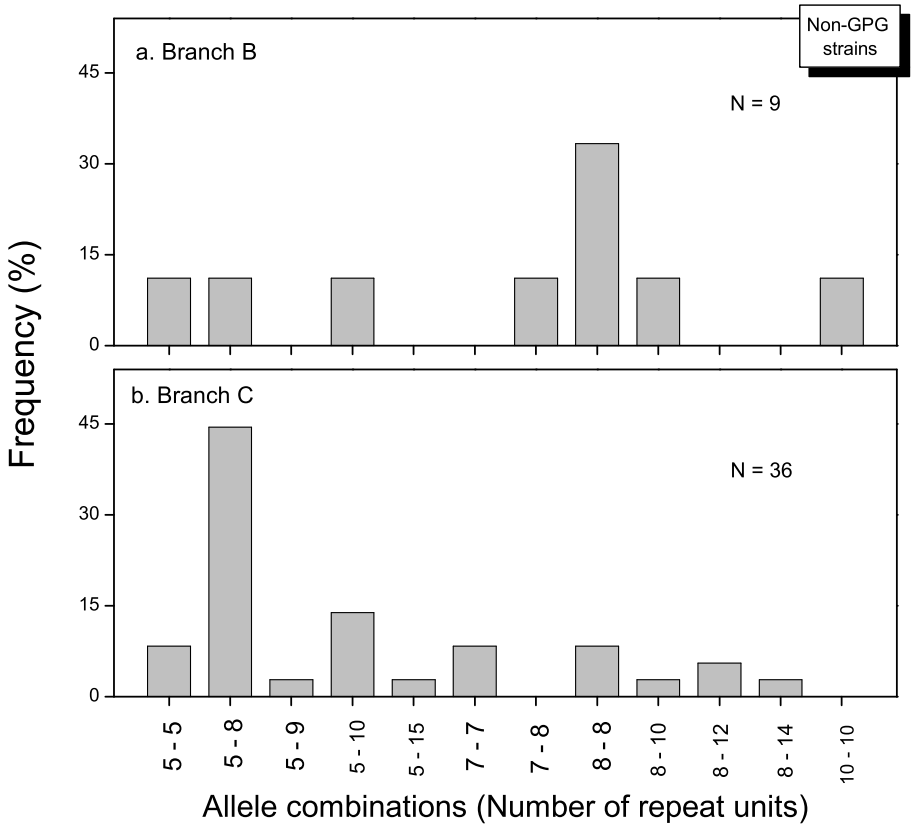

Figure 3.18: Distribution of allele combinations of *YWP1* of non-GPG strains of branches B and C.

strains selected by the genetic background, these two groups share the same predominant alleles, alleles containing 5 repeat units. This may be an indication of adaptation, which suggests that *YWP1* is a contingency gene. This indicates that alleles containing 5 repeat units are the optimal alleles, which two different groups of strains select. This result suggests that *YWP1* exhibits a property of a contingency gene, i.e. *YWP1* has a role in adaptation by changing the number of repeat units in the coding sequences.

GPG strains, which belong to clade 1 of five clades in *C. albicans*, have been proposed to be better adapted as humans epithelial colonizer and invader compared to other clades [78, 79], are more virulent than non-GPG strains in young candidemia patients [80], and are 10 to 100 times more likely to cause disease than non-GPG strains [56]. Based on this evidence, it is reasonable to assume that non-GPG strains select the predominant alleles of *YWP1*, which are the same as the predominant alleles of *YWP1* selected by GPG strains, i.e. optimal alleles. With this argument, alleles containing 5 repeat units are suspected to be the optimal alleles.

The fact that GPG and non-GPG strains select the same predominant alleles, i.e. alleles containing 5 repeat units, may indicate that alleles of 5 repeat units play an important role to optimize the function of ywp1 protein in *C. albicans* pathogenicity. Experimental studies on the *YWP1* gene showed that the product of this gene inhibits the adhesion of the yeast form of *C. albicans* [60]. As explained in the previous section, YWP1p seems to have a role in promoting the dispersal ability of the *C.albicans* in yeast form. Other genes such as *ALS5*, *ALS6*, and *ALS7* also share this property of *YWP1*. For example, deletion of *ALS5*, *ALS6*, and *ALS7* increases adhesion of *C. albicans* to humans endothelial and epithelial cells [81]. Interestingly, allelic distributions of *ALS5* and *ALS6* in strains from five different clades of *C. albicans* indicated that, although there is a clade specificity of allele combinations observed, different clades share the same predominant alleles in addition to distinct predominant alleles unique to each clade for *ALS5* or some clades for *ALS6* [33], i.e. strains in all clades select predominant alleles of *ALS5* and *ALS6* containing 4 repeat units. This is similar to the results of GPG and non-GPG strains, where strains from these two

groups select predominant alleles of *YWP1* containing 5 repeat units. The similar pattern of the allelic distribution of *YWP1*, *ALS5*, and *ALS6* observed may relate to the same property these gene share, i.e. adhesion inhibitor. This argument is with an exception for *ALS7*, where the predominant alleles were observed in GPG strains, but there were no obvious predominant alleles in non-GPG strains [32]. It seems that for YWP1p, ALS5p, and ALS6p to play a role in promoting the dispersal ability of the *C.albicans* in yeast form, one particular allele or number of repeats in their associated genes is vital; in *YWP1*, it is an allele of 5 repeat units, while in *ALS5* and *ALS6* they are both alleles of 4 repeat units.

The fact that alleles containing 5 repeat units occurred twice (allele combination 5-5) in GPG strains as often as in non-GPG strains (allele combination 5-8) may also indicate a gene dosage effect. The effect may relate to the facts that GPG strains are 10 to 100 times more likely to cause disease than non-GPG strains [56], and that higher mortality is observed in young patients of candidemia associated with GPG strains than non-GPG strains [80].

3.4.3 The number of *YWP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic

The comparison of the allelic distribution of *YWP1* between commensal and infection strains showed no significant difference between the distributions; both have the same predominant alleles and allele combinations, and their percentages are similar (see Figure 3.9 on page 45 and Figure 3.11 on page 47). In relation to that, the variation of allele combinations shown by the index of diversity is similar in both commensal and infection strains (see Figure 3.13 on page 51). The same predominant alleles and allele combinations of *YWP1* in commensal and infection strains indicate no association between the allele length or number of repeats and the immune status of the host. This suggests that the number of *YWP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic, i.e. *YWP1* may not act as a contingency gene when the immune status of the host changes. This is in agreement

with previous studies of other *C. albicans* genes. For example, results from studies on microsatellite of *CEF3* [82], *HWP1*, *ZNF1*, and *MNT1* [83] showed no variability in the number of repeats in commensal and infection strains. Similarly, studies on short sequence repeats found no variability in the number of repeats in commensal and infection strains in most cases [84]. It seems that for the *YWP1* gene, *C. albicans* select particular allele combinations when growing within a host in both commensal and pathogenic states. The presence of particular YWP1p on the yeast surface may be necessary in both commensal and pathogenic states and may have the same function in both states to allow optimal interaction with the host. The observation that YWP1p plays a role in *C. albicans* dispersal [60], an ability needed in both commensal colonization in a healthy host and invasion of tissue in an immunocompromised host, may explain why there is no significant change in the number of *YWP1* repeats when *C. albicans* state changes from commensal pathogenic.

In contrast to the results of the *YWP1* gene, studies on *SSR1* [52] and *EFG1* [83] of *C. albicans* showed a difference in the number of repeats in commensal and infection strains. These indicate that whether or not a change in the number of repeats of the alleles is required when *C. albicans* state changes from commensal to pathogenic is gene-dependent.

3.4.4 The number of *YWP1* repeats does not alter when *C. albicans* moves to particular sites of the humans body

This study showed no significant difference in the distribution of alleles and allele combinations of *YWP1* for any humans body sites sampled (see Figure 3.16 on page 57 and Figure 3.17 on page 59). The results suggest that a change in the number of *YWP1* repeats is not required when *C.albicans* moves to particular sites of the humans body. This is in agreement with the results of studies on *ALS7* [32], *PNG2* [51], *SSR1* [52], and a study on bloodstream and non-bloodstream isolates [85].

This result shows that there is no influence of the different environmental conditions (provided by different sites of the humans body) on the *YWP1* allelic distributions. This is also supported by the evidence that there is a similarity in the *YWP1* allelic distributions in strains isolated from oral sites compared to all strains from different sites. For example, the distribution of allele combinations of *YWP1* in GPG and non-GPG strains isolated from oral sites (see Figure 3.8 on page 43) is similar to the distribution of allele combinations of *YWP1* in GPG and non-GPG strains from all infection strains tested (see Figure 3.7 on page 41). The same is true when comparing allele combinations of *YWP1* in commensal and infection strains (see Figures 3.12 and 3.11 on pages 49 and 47).

Since *C.albicans* colonizes different sites of the humans body, the presence of a variability in the number of repeats in its genes is frequently associated with adaptation to a changing environment [27, 74]. However, the results of this particular study of the *YWP1* gene argue against this scenario, and indicate that there is no requirement to alter the number of *YWP1* repeats when *C. albicans* adapts to particular sites of the humans body.

3.4.5 The homozygous and heterozygous alleles of *YWP1*

The analysis of the combination of the two alleles of *YWP1* in an individual strain showed that *YWP1* exhibits excesses of homozygotes in GPG and non-GPG strains, but the excess of homozygotes in these two groups is not significant (see Table 3.6 on page 55). In addition, a significant excesses of homozygotes was observed in commensal strains. The results of significant excesses of homozygotes have been reported previously for *ALS5* and *ALS6* [33].

Loss of heterozygosity has been found to be a common event in *C. albicans* both in commensal and pathogenic states [86]. This may be important in evolution of *C. albicans* strains. It may be beneficial since in homozygosity, the number of gene product of selected allele is twice as many as in heterozygosity. Another possibility of the importance of this phenomenon is condominance of alleles as suggested for *TAC1*, a *C. albicans* gene that plays a role in azole resistance of this fungus. It has

been suggested that heterodimer is formed when two different alleles of *TAC1* are expressed in a cell while homodimer is formed when the same alleles of *TAC1* are expressed in a cell. A correct formation is achieved when the homodimer binds to another element which leads to a strong azole resistance. In contrast, an incorrect formation can be achieved when the heterodimer binds to the element leading to a weak azole resistance [87]. In this case, homozygosity is beneficial to *C. albicans* to have a strong antibiotic resistance.

The analysis of the combination of the two alleles in an individual strain showed that the two alleles of *YWP1* in a GPG strain are significantly more similar to each other than those expected by chance (see Table 3.7 on page 56). However, the analysis showed that the two alleles of *YWP1* in a non-GPG strain did not differ significantly from those expected by chance. In addition, the analysis showed that the two alleles of *YWP1* in a commensal strain are significantly more similar to each other than those expected by chance. The result that the two alleles are significantly more similar to each other than those expected by chance is similar to the results for *ALS5*, *ALS6*, and *ALS7* [32, 33]. However, this result is in contrast to the result for *ALS3* which showed the observed difference is significantly larger than the expected difference by chance [27, 33]. This study found that alleles with short length tend to pair with allele with long length in the same strain, and suggested that short and long alleles may encode proteins with different functions.

Chapter 4

Alleles of the *HWP1* Gene

This chapter describes the results of the allelic characterization of the *HWP1* gene in strains of interest, which is aimed to determine whether *HWP1* is a contingency gene, i.e. whether *HWP1* has a role in adaptation by changing the number of repeat units within the coding sequences. For this purpose, allelic distribution of *HWP1* in GPG and non-GPG strains, in commensal and infection strains, and in strains isolated from different sites of the human body will be compared one to another.

The chapter begins with a description of the identification of the repeat units in the *HWP1* gene. This is followed by the results of the allelic characterization for GPG and non-GPG strains (infection strains), two group of strains with different genetic backgrounds, and the results of allelic characterization for the commensal strains, for comparison with the infection strains (both GPG strains). Analysis of the diversity of the alleles and allele combinations, and analysis of combination of the two alleles in an individual strain for all strains of interest are explained before description of the results of allelic characterization for strains isolated from different sites of the human body. The chapter ends with a discussion of the biological implications of the results.

4.1 Identification of the Repeat Units in the *HWP1* Gene

The repeat units in the *HWP1* gene were identified using MacVector software (MacVector Inc, www.macvector.com) to analyse the DNA sequence of *HWP1* from strain SC5314 downloaded from www.candidagenome.org, i.e. the same procedure used to identify repeats in *YWP1*. The result of this analysis is shown in Figure 4.1.

Primers HWP1F and HWP1R (see Table 2.5 on page 24) were designed to amplify the region from nucleotide 65 - 584 (520 bp) which is a region containing repeat units.

4.2 Allelic Characterization of the *HWP1* Gene in GPG and Non-GPG Strains (Infection Strains)

Alleles of the *HWP1* gene from 49 GPG strains and 47 non-GPG strains were characterized by length using genotyping. As indicated previously, since *C. albicans* is diploid, the results are described for both alleles and allele combinations.

Based on allele length, seven different alleles were found in GPG strains and nine in non-GPG strains. In total, there were nine different alleles found in infection strains. In order to determine the number of repeat units for each fragment length, the repeat region of the *HWP1* gene of different fragment lengths were sequenced. Primers HWP1BF and HWP1BR (see Table 2.5 on page 24) were used for sequencing. The repeat unit identified in DNA sequencing was 30 bp in length which encodes the 10 amino acid sequence IPCDNPPQPD. The results of the sequencing showed that the difference between the fragment lengths observed is caused by a removal or addition of the repeat unit/s. However, it was also found that some different alleles in the fragment length have the same number of repeats. For example, alleles of 335 and 345 bp contain 2 repeat units, 365 and 375 bp contain 3 repeat units, and 395 and 405 contains 4 repeat units. For this case, the difference between the fragment lengths was not caused by a removal or addition of the repeat unit/s, but by changes of some parts outside the repeat units. The frequency of alleles of *HWP1* found in GPG and

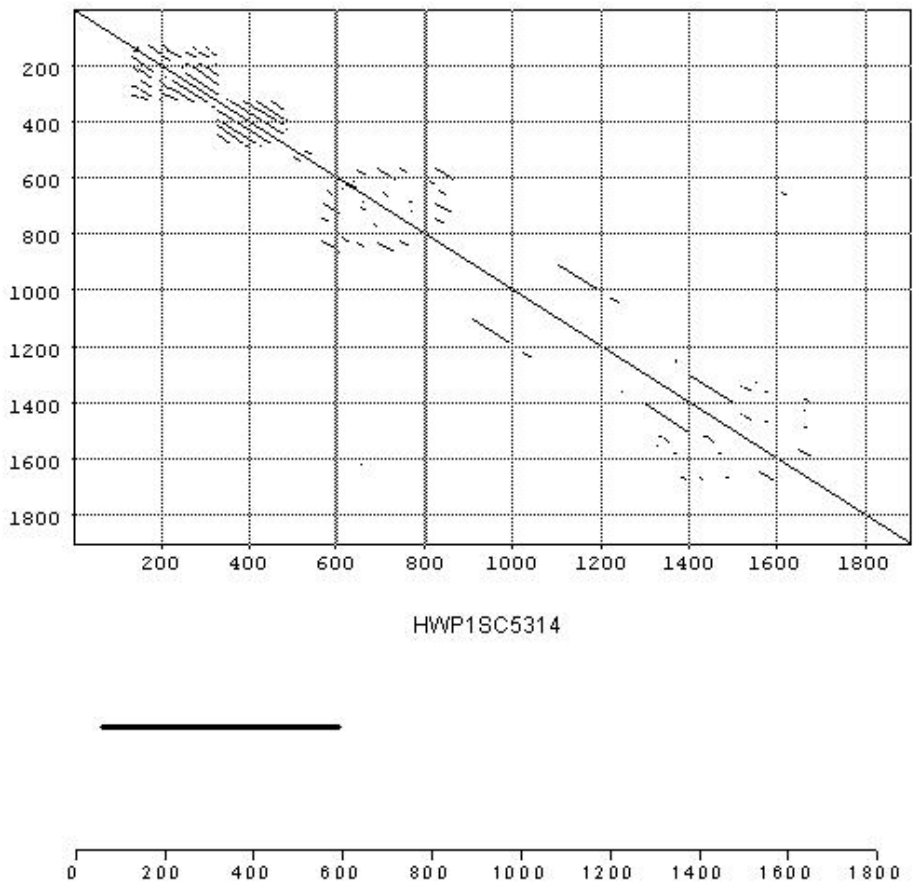

Figure 4.1: Identification of the repeat unit in the *HWP1* gene using Macvector software (top figure). The identification was performed on the DNA sequence from *HWP1* of strain SC5314 downloaded from www.candidagenome.org. The black thick line indicates the region amplified, i.e. nucleotide 65 - 584 (520 bp) (bottom figure).

non-GPG strains is summarized in Table 4.1, and presented graphically in Figure 4.2. The number of *HWP1* alleles for GPG strains was reduced from seven by fragment length to four by number of repeats, and for non-GPG strains from nine to five.

Figure 4.2 shows that alleles containing 4 and 5 repeat units predominated in both GPG and non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). The Chi-square test for contingency tables showed no significant difference in the distribution of these alleles between GPG and non-GPG strains ($p > 0.05$). Similar results were also observed for strains isolated from oral sites as shown in Figure 4.3. These results show that GPG and non-GPG strains have the same predominant alleles.

Table 4.1: Frequency of alleles of the *HWP1* gene in GPG and non-GPG strains. Alleles are represented by the fragment length and the number of repeat units.

No	Allele		Frequency		
	Length (bp)	No. of repeats	GPG Strains	Non-GPG Strains	Total
1	335	2	1	2	3
2	345	2	2	1	3
3	365	3	1	1	2
4	375	3	7	6	13
5	395	4	3	14	17
6	405	4	34	26	70
7	425	5	0	1	1
8	435	5	50	42	92
9	465	6	0	1	1
Total			98	94	192

Based on allele length, the number of different allele combinations found in GPG strains was 11, compared to 14 combinations in non-GPG strains. In total there were 17 different allele combinations found in infection strains. As some alleles that differ in length have the same number of repeats, the number of different allele combinations reduced from 11 by fragment length to eight by number of repeat units for GPG strains, from 14 to eight for non-GPG strains, and from 17 to 10 for all infection strains. The frequency of allele combinations of *HWP1* found in GPG and non-GPG strains is summarized in Table 4.2, and presented graphically in Figures 4.4.

Figure 4.4 shows that allele combinations of *HWP1* containing 4 and 4, 4 and

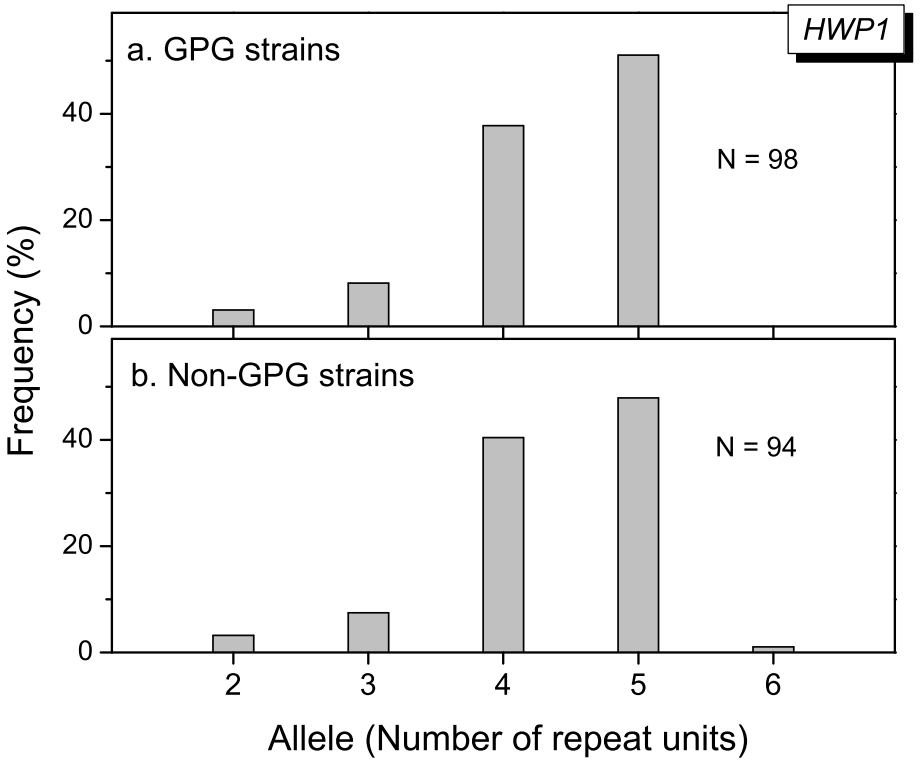

Figure 4.2: Distribution of alleles of *HWP1* for: a. GPG strains, and b. non-GPG strains. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 4 and 5 repeat units predominated in both GPG and non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p > 0.05$).

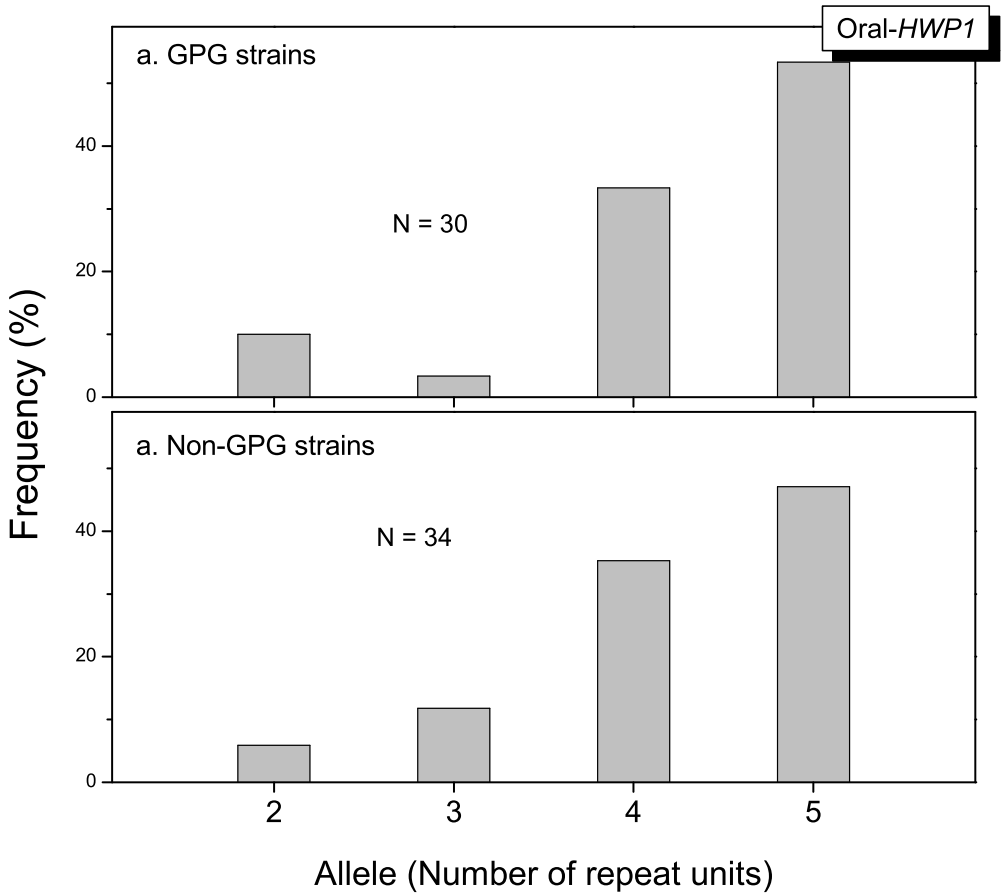

Figure 4.3: Distribution of alleles of *HWP1* for: a. GPG strains, and b. non-GPG strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 4 and 5 repeat units predominated in both GPG and non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p > 0.05$).

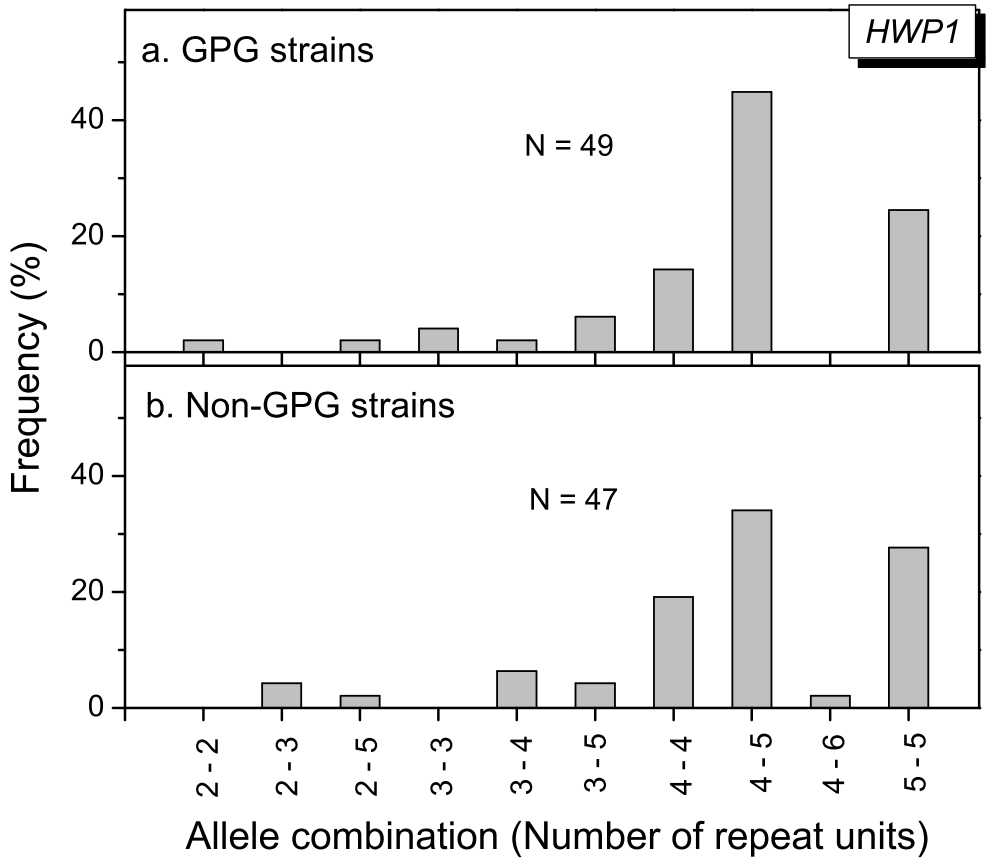

Figure 4.4: Distribution of allele combinations of *HWP1* for: a. GPG strains, and b. non-GPG strains. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 4 and 4, 4 and 5, and 5 and 5 repeat units predominated in both GPG and non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p > 0.05$).

Table 4.2: Frequency of allele combinations of *HWP1* in GPG and non-GPG strains. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

No	Allele Combination		Frequency		
	Length (bp)	No. of repeats	GPG Strains	Non-GPG Strains	Total
1	335-375	2-3	0	1	1
2	335-435	2-5	1	1	2
3	345-345	2-2	1	0	1
4	345-365	2-3	0	1	1
5	365-375	3-3	1	0	1
6	375-375	3-3	1	0	1
7	375-395	3-4	0	1	1
8	375-405	3-4	1	2	3
9	375-435	3-5	3	2	5
10	395-395	4-4	1	1	2
11	395-405	4-4	0	5	5
12	395-435	4-5	1	5	6
13	395-465	4-6	0	1	1
14	405-405	4-4	6	5	11
15	405-435	4-5	21	9	30
16	425-435	5-5	0	1	1
17	435-435	5-5	12	12	24
Total			49	47	96

5, and 5 and 5 repeat units predominated in both GPG and non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). The results of the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations of *HWP1* between GPG and non-GPG strains ($p > 0.05$). Similar results were also observed for strains isolated from oral sites as shown in Figure 4.5. These results show that GPG and non-GPG strains have the same predominant allele combinations of *HWP1*.

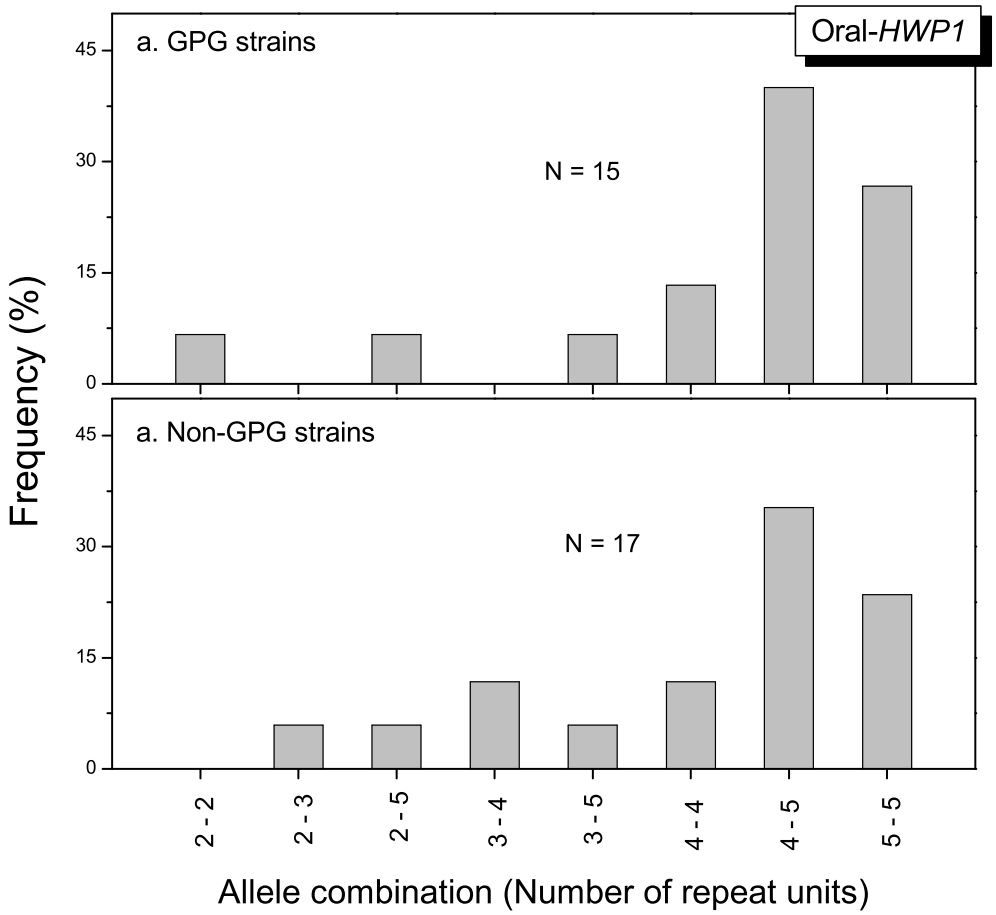

Figure 4.5: Distribution of allele combinations of *HWP1* for: a. GPG strains, and b. non-GPG strains, where all strains were isolated from oral sites. N represents total number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are number of repeat units. Allele combinations containing 4 and 4, 4 and 5, and 5 and 5 repeat units predominated in both GPG and non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p > 0.05$).

4.3 Allelic Characterization of the the *HWP1* Gene in Commensal Strains

Alleles of the *HWP1* gene from 35 commensal strains were characterized by length using genotyping. Based on allele length, five different alleles were found in commensal strains, compared to four in infection strains. The number of repeat units for each fragment length was determined using the sequencing results shown previously for the infection strains. The frequency of *HWP1* alleles found in commensal strains is summarized in Table 4.3, and presented graphically in Figure 4.6. For comparison, the data from infection strains have been added in the table and the figure.

Figure 4.6 shows the same pattern of the distribution of *HWP1* alleles in both commensal and infection strains, where alleles containing 4 and 5 repeat units predominated in each group (the Chi-square goodness of fit test, $p < 0.001$). The results of the Chi-square test for contingency tables showed no significant difference in the distribution of these alleles between commensal and infection strains ($p > 0.05$). Similar results were also observed for strains isolated from oral sites as shown in Figure 4.7. These results show that commensal and infection strains have the same predominant alleles of *HWP1*.

Table 4.3: Frequency of alleles of *HWP1* in commensal strains. Alleles are represented by the fragment length and the number of repeat units. For comparison, the data from infection strains have been added.

	Allele		Frequency		
No	Length (bp)	No. of repeats	Commensal Strains	Infection Strains	Total
1	335	2	2	1	3
2	345	2	0	2	2
3	365	3	0	1	1
4	375	3	6	7	13
5	395	4	0	3	3
6	405	4	27	34	61
7	435	5	34	50	84
8	465	6	1	0	1
Total			70	98	168

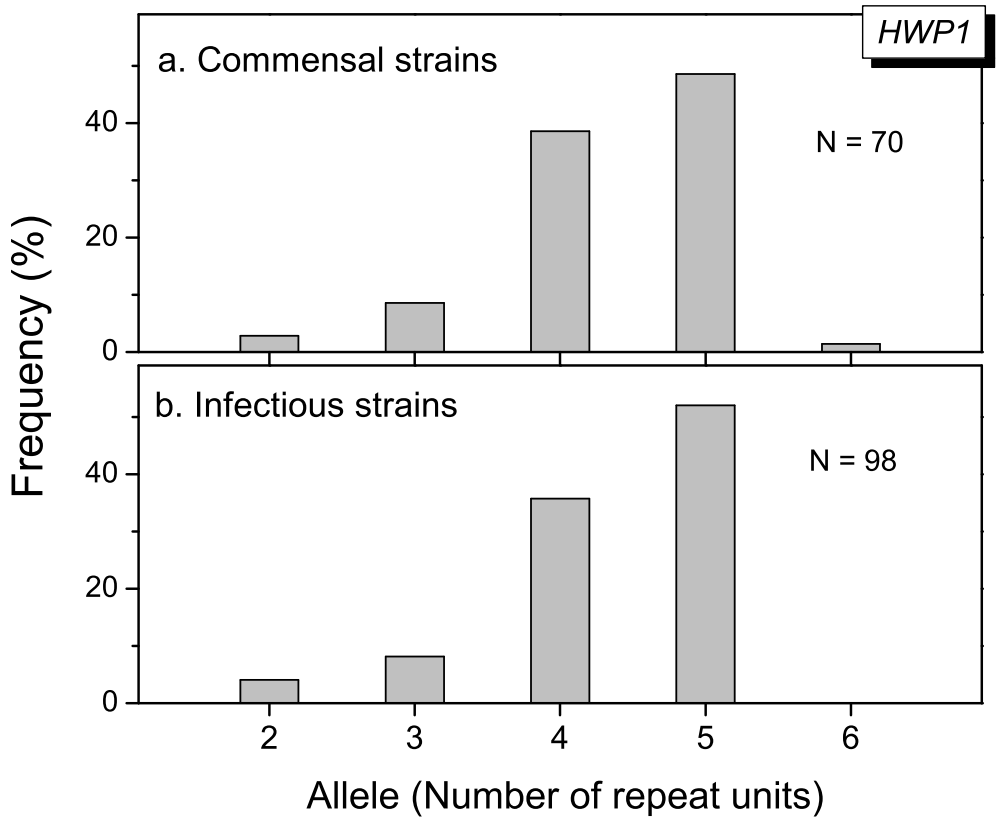

Figure 4.6: Distribution of alleles of *HWP1* for: a. Commensal strains, and b. Infection strains. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 4 and 5 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency table, $p > 0.05$).

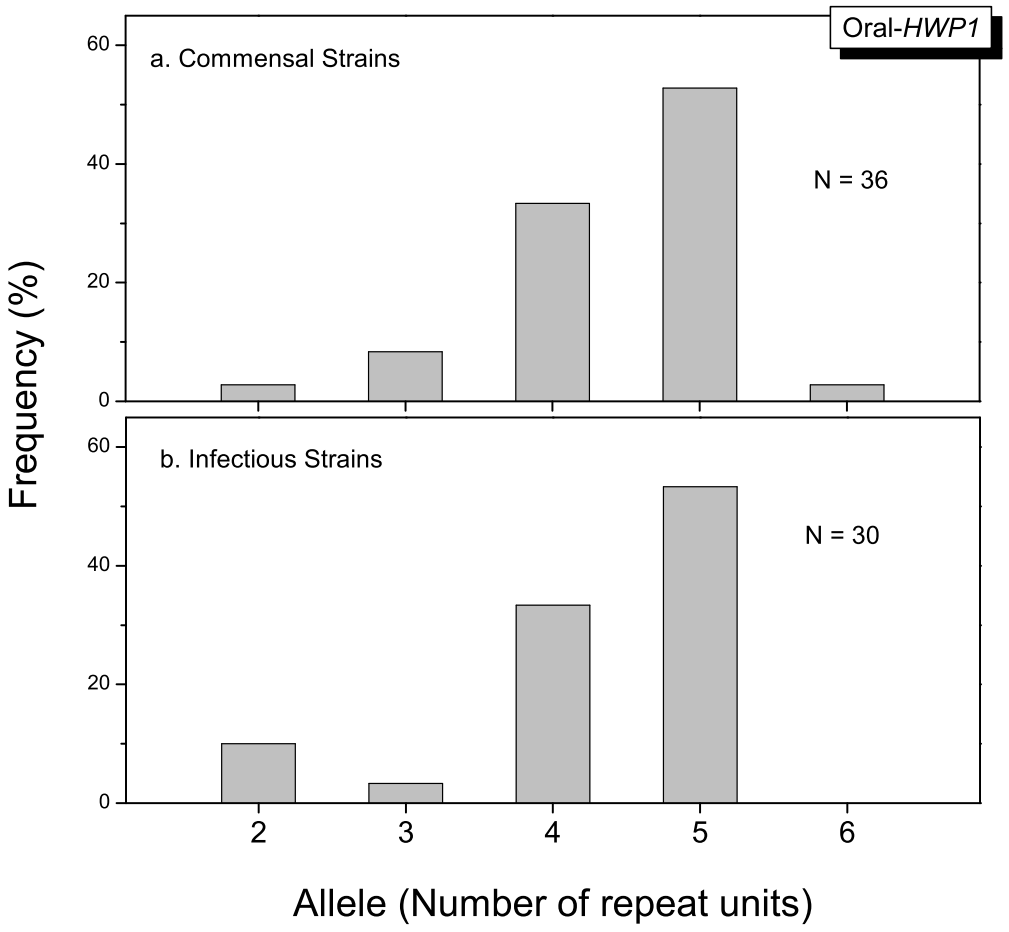

Figure 4.7: Distribution of alleles of *HWP1* for: a. Commensal strains, and b. Infection strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 4 and 5 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency table, $p > 0.05$).

Table 4.4: Frequency of allele combinations of *HWP1* in commensal strains. For comparison, the data from infection strains have been added. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

No	Allele Combination		Frequency		
	Length (bp)	No. of repeats	Commensal Strains	Infection Strains	Total
1	335-375	2-3	1	0	1
2	335-435	2-5	1	1	2
3	345-345	2-2	0	1	1
4	365-375	3-3	0	1	1
5	375-375	3-3	0	1	1
6	375-405	3-4	0	1	1
7	375-435	3-5	4	3	7
8	375-465	3-6	1	0	1
9	395-395	4-4	0	1	1
10	395-435	4-5	0	1	1
11	405-405	4-4	2	6	8
12	405-435	4-5	23	21	44
13	435-435	5-5	3	12	15
Total			35	49	84

The number of different allele combinations of *HWP1* found in commensal strains was seven compared to eight combinations in infection strains. The frequency of allele combinations of *HWP1* is summarized in Table 4.4, and presented graphically in Figure 4.8. For comparison, the data from infection strains have been added in the table and the figure.

Figure 4.8 shows that allele combinations of *HWP1* containing 4 and 5 repeat units predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$). Comparing the distribution of allele combinations between commensal and infection strains, there is a noticeable difference; moving from graphs for commensal to infection strains, there is a decrease in proportion of allele combinations containing 4 and 5 repeat units (from 66% to 45%), and a slightly increase in allele combinations containing 4 and 4, and 5 and 5 repeat units. However, the results of the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between commensal and infection strains ($p > 0.05$). Similar

results were also observed for strains isolated from oral sites as shown in Figure 4.9. These results show that commensal and infection strains have the same predominant allele combinations of *HWP1*.

4.4 Diversity of Alleles of the *HWP1* Gene

Allele diversity is quantified by the index of diversity λ which is defined by Equation 3.1 (see page 48). The λ values for both the alleles and the allele combinations of *HWP1* for GPG, non-GPG, and commensal strains are summarized in Table 4.5, and presented graphically in Figure 4.10.

Table 4.5: Index of diversity λ of the alleles and the allele combinations of *HWP1* for GPG, non-GPG (infection), and (GPG) commensal strains.

Strains	Index of Diversity	
	Allele	Allele Combination
GPG	0.60	0.73
Non-GPG	0.60	0.78
Commensal	0.61	0.56

The results show that the diversity of the alleles and the allele combinations of *HWP1* is not influenced by the genetic diversity of the strains; GPG strains with small genetic diversity and non-GPG strains with large genetic diversity have identical allele diversity. Although the diversity of alleles of commensal and infection strains is also similar, the diversity of allele combinations is different; the allele combinations of infection strains are more diverse than those of commensal strains.

4.5 Analysis of the Combination of the Two Alleles of the *HWP1* Gene in An Individual Strain

In this section, the combination of the two alleles of *HWP1* in an individual strain among strains of interest is analysed. Similar to the description for *YWP1* in the

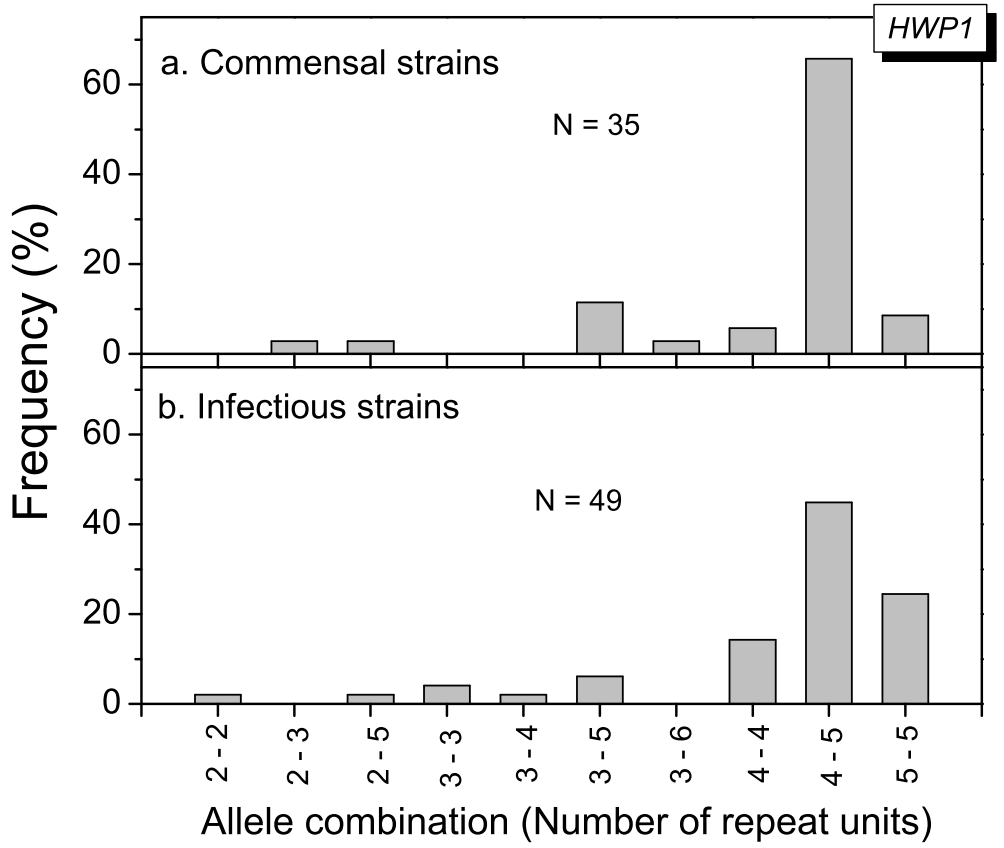

Figure 4.8: Distribution of allele combinations of *HWP1* for: a. Commensal strains, and b. Infection strains. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are number of repeat units. Allele combinations containing 4 and 5 repeat units predominated in commensal strains, while allele combinations containing 4 and 4, 4 and 5, and 5 and 5 repeat units predominated in infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in distribution of these allele combinations between commensal and infection strains. (the Chi-square test for contingency tables, $p < 0.05$).

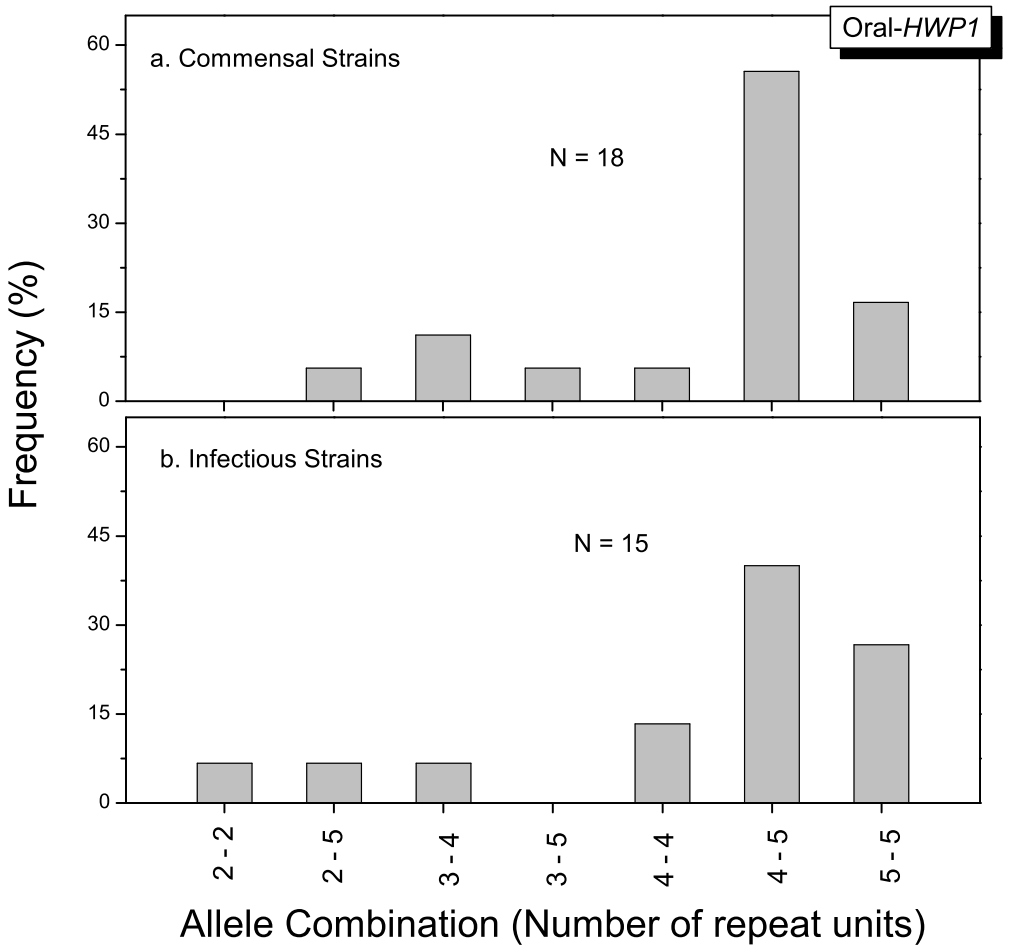

Figure 4.9: Distribution of allele combinations of *HWP1* for: a. Commensal strains, and b. Infection strains, where all strains were isolated from oral sites. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 4 and 5 repeat units predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), while allele combinations containing 4 and 5, and 5 and 5 repeat units predominated in infection strains. There is no significant difference in the distribution of these allele combinations between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

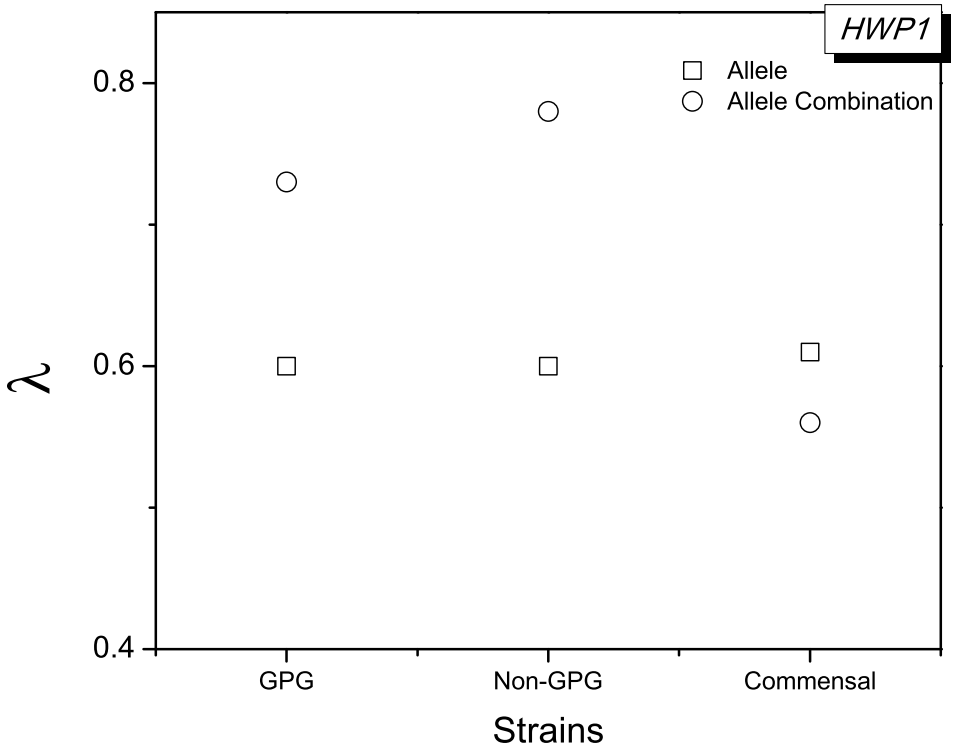

Figure 4.10: Index of diversity λ of the alleles and the allele combinations of *HWP1* for GPG, non-GPG (infection), and (GPG) commensal strains.

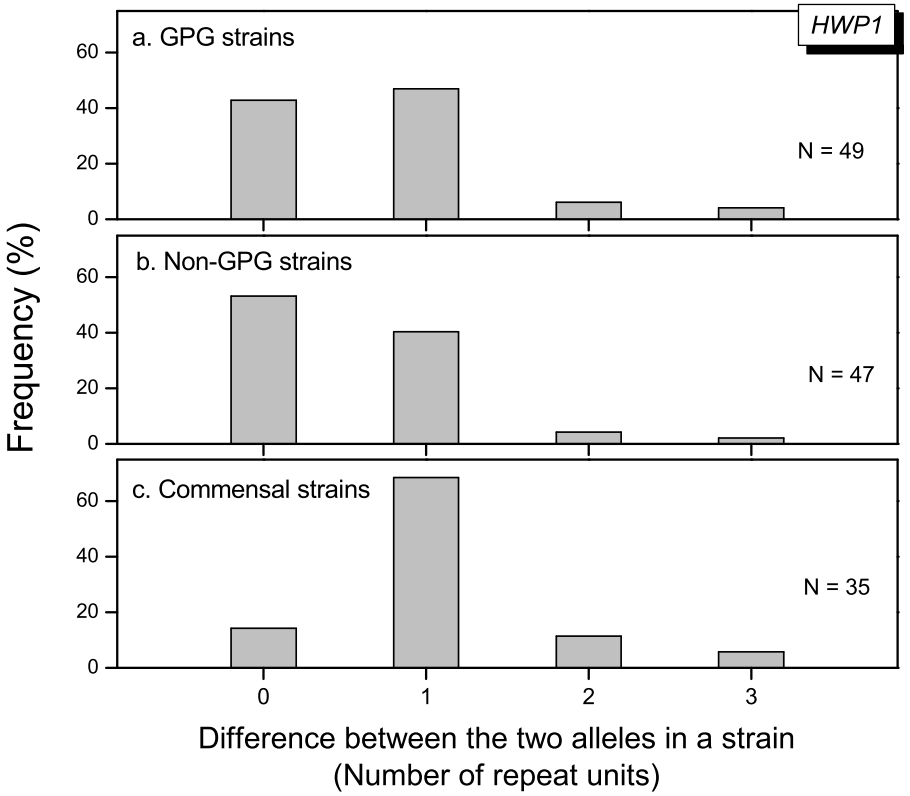

Figure 4.11: Distribution of differences in the number of repeat units between the two alleles of *HWP1* in an individual strain for: a. GPG strains, b. Non-GPG strain, and c. Commensal strains. N represents the total number of strains in a particular group.

previous chapter, the section starts with a description of the distribution of the differences between the two alleles in an individual strain, and ends with a calculation of the mean of the differences between the two alleles in an individual strain among strains of interest.

Figure 4.11 shows the distribution of differences in the number of repeat units between the two alleles of *HWP1* in an individual strain for GPG, non-GPG, and commensal strains. The percentages of homozygous alleles of *HWP1* found in GPG, non-GPG, and commensal strains are 43%, 53%, and 14%, respectively.

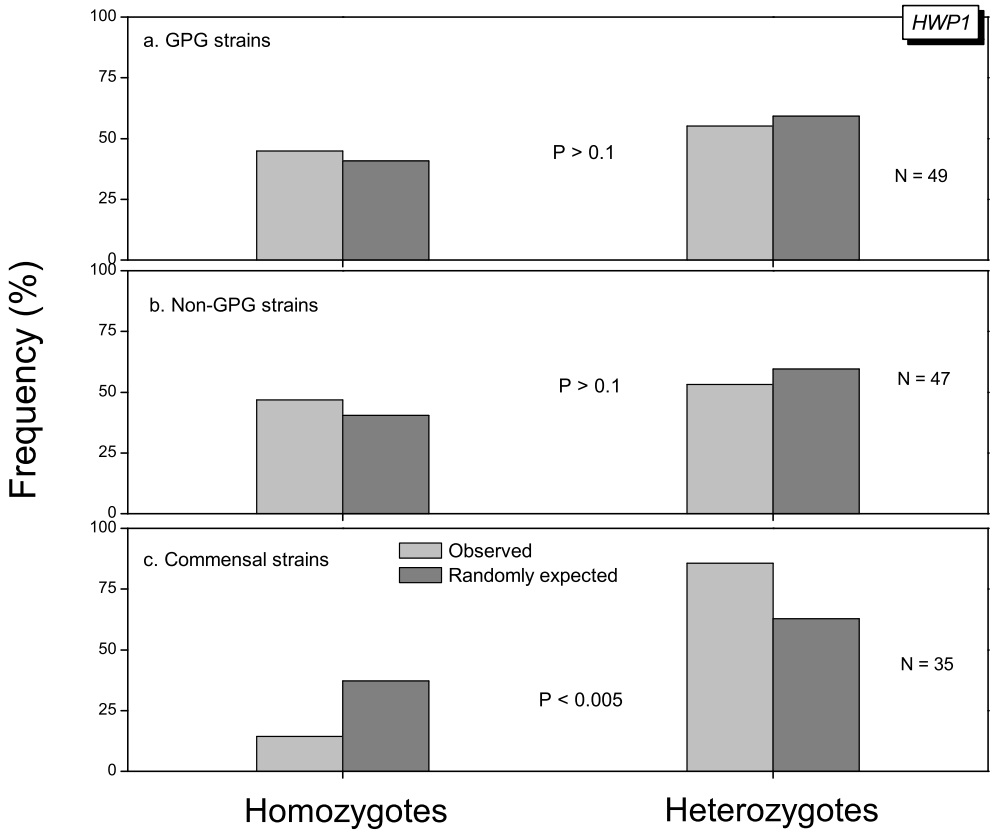

Figure 4.12: Frequency of homozygotes and heterozygotes of *HWP1* for GPG, non-GPG, and commensal strains. N represents the number of strains in a particular group. There is no significant difference between the number of homozygotes observed and those expected by chance for GPG and non-GPG strains (the Chi-square test for contingency tables , $p>0.1$). However, for commensal strains, the number of homozygotes is significantly smaller than those expected by chance (the Chi-square test for contingency tables, $p<0.005$), i.e. *HWP1* exhibits significant deficit of homozygotes in commensal strains.

The number of homozygous alleles in each group was compared to those expected if the combination of the alleles was random, taking allele frequency into consideration. For this purpose, a matrix with a size equal to the square of 2 times the number of strains analysed was generated, i.e. the same procedure used for the analysis for the *YWP1* gene in the previous chapter (see section 3.3.4, page 50). The frequency of homozygotes and heterozygotes of *YWP1* for the data observed and those expected from random alleles is summarized in Table 4.6, and is presented graphically in Figure 4.12. The figure shows no significant difference between the number of homozygotes observed and those expected by chance for GPG and non-GPG strains (the Chi-square test for contingency tables , p>0.1). However, for commensal strains, the number of homozygotes is significantly smaller than those expected by chance (the Chi-square test for contingency tables, p<0.05), i.e. *HWP1* exhibits significant deficit of homozygotes in commensal strains.

Table 4.6: The number of homozygous alleles in GPG, non-GPG, and commensal strains compared to the expected number by chance. N indicates the number of strains.

Strains	Number of homozygous alleles		Significance (p)
	Observed data (N)	Expected data	
GPG	22 (49)	20	$p > 0.1$
Non-GPG	22 (47)	19	$p > 0.1$
Commensal	5 (35)	13	$p < 0.005$

The mean (± standard deviation) of the differences in the number of repeat units between the two alleles in an individual strain was determined to be 0.65 ± 0.69, 0.64 ± 0.70, and 1.08 ± 0.69, for GPG, non-GPG, and commensal strains, respectively. The result of the t-test between the two means of GPG and non-GPG strains showed no significant difference between the two means ($p > 0.05$). The result of the t-test between the two means of (GPG) commensal and infection strains showed that that the two alleles in an infection strain are significantly more similar to each other than those in commensal strains ($p < 0.001$).

To determine whether the difference between the two alleles in an individual strain

is a random or non-random (i.e. is a selective pressure operating ?), the means of the differences of the alleles found were compared to the means expected if all alleles combined randomly, taking allele frequency into consideration. As explained previously, a matrix with a size equal to the square of 2 times the number of strains analysed was generated for this purpose.

For GPG strains, the mean (± standard deviation) of the differences in the number of repeat units between the two alleles of *HWP1* expected from random alleles was determined to be 0.76 ± 0.76. The result of t-test between this mean and the mean from observed data 0.65 ± 0.69 showed that the two alleles of *HWP1* in a GPG strain did not differ significantly from those expected by chance ($p > 0.10$). For non-GPG strains, the mean (± standard deviation) of the differences in number of repeat units between the two alleles of *HWP1* expected from random alleles was determined to be 0.77 ± 0.77. The result of the t-test between this mean and the mean from observed data 0.64 ± 0.69 showed that the two alleles of *HWP1* in a non-GPG strain did not differ significantly from those expected by chance ($p > 0.10$). For commensal strains, the mean (± standard deviation) of the differences in the number of repeat units between the two alleles of *HWP1* expected from random alleles was determined to be 0.76 ± 0.76. The result of the t-test between this mean and the mean from observed data 1.08 ± 0.69 showed that the two alleles of *HWP1* in a commensal strain are significantly more different to each other than expected by chance ($p < 0.01$). The mean (± standard deviation) of the differences in the number of repeat units between the two alleles of *HWP1* from observed data, data expected from random alleles, and the significance of the differences between the means are summarized in Table 4.7.

Table 4.7: The mean (± standard deviation) of the differences in the number of repeat units between the two alleles of *HWP1* in an individual strain from observed data, data expected from random alleles, and the significance of the difference between the two means for GPG, non-GPG, and commensal strains. N indicates the number of data.

Strains	Mean Differences (± SD)		Significance (p)
	Observed Differences (N)	Expected Differences (N)	
GPG	0.65 ± 0.69 (49)	0.76 ± 0.76 (9604)	$p > 0.10$
Non-GPG	0.64 ± 0.69 (47)	0.77 ± 0.77 (8836)	$p > 0.10$
Commensal	1.08 ± 0.69 (35)	0.76 ± 0.76 (4900)	$p < 0.01$

4.6 Allelic Distribution of the *HWP1* Gene in Strains Isolated From Different Sites of the Humans Body

Figure. 4.13 shows the distribution of alleles of *HWP1* in strains isolated from different sites of the humans body. The figure shows that the distributions of *HWP1* alleles at all sites appear similar; alleles containing 4 and 5 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). Result from the Chi-square test for contingency tables showed no significant difference in the distribution of these alleles between any two sites for all sites sampled ($p > 0.05$). These results show that strains isolated from different sites of the humans body have the same predominant alleles.

Figure. 4.14 shows the distribution of allele combinations of *HWP1* in strains isolated from different sites of the humans body. The figure shows that allele combinations containing 4 and 5 repeat units predominated at oral, vagina, and urine sites, but no obvious pattern of any predominant allele combinations observed at skin sites. There is no significant difference in the distribution of allele combinations containing 4 and 5 repeat units between oral and vagina sites, oral and urine sites, and vagina and urine sites, but there is a significant difference in the distribution of these allele combinations between vagina and skin sites, and vagina and sterile sites. The figure

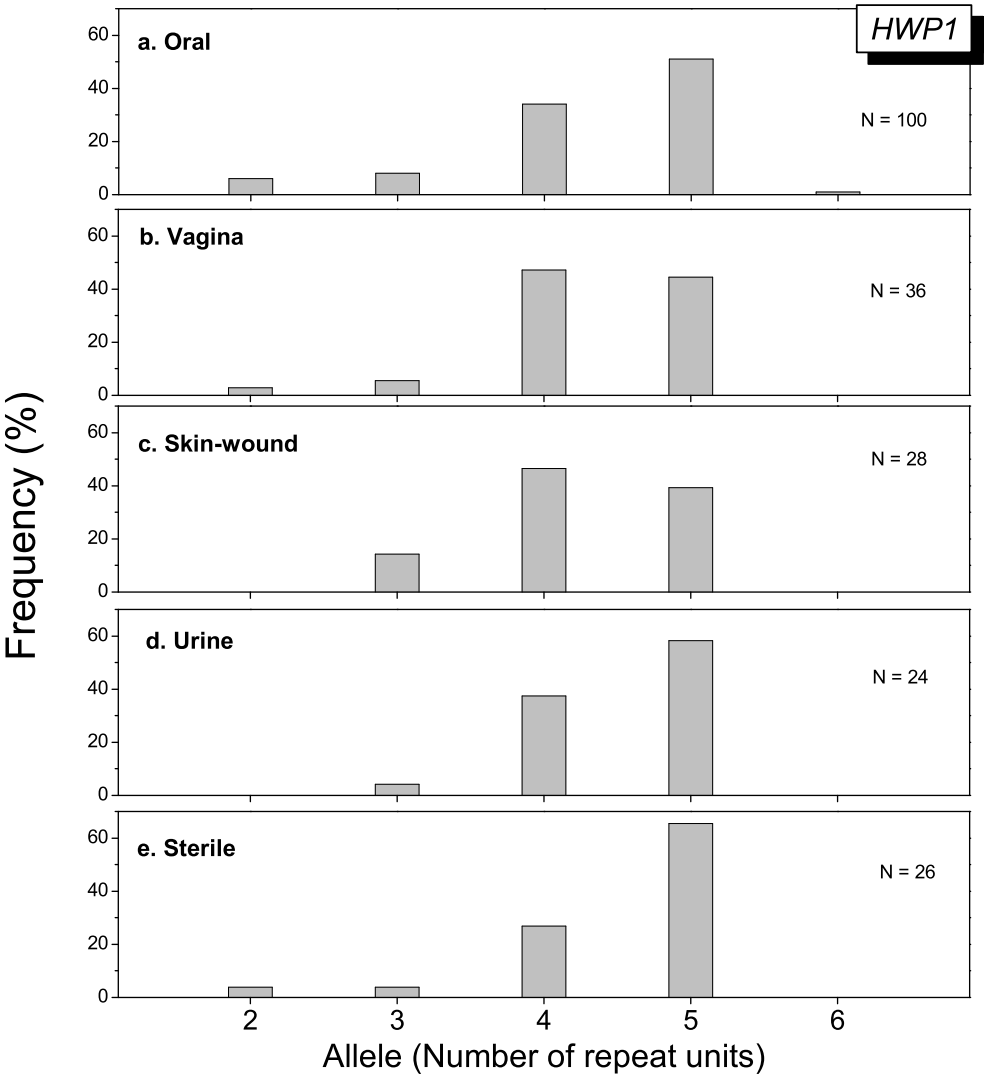

Figure 4.13: Distribution of alleles of *HWP1* in strains isolated from different sites of the humans body. N represents 2 times the number of strains at each site, as there are 2 alleles per strain. Allele containing 4 and 5 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between any two sites for all sites sampled (the Chi-square test for contingency table, $p > 0.05$).

also shows that allele combinations containing 5 and 5 repeat units predominated at sterile sites. There is a significant difference in the distribution of these allele combinations between sterile and oral sites, and sterile and vagina sites. In contrast, there is no significant difference in the distributions of these allele combinations between sterile and skin sites, and sterile and urine sites. These results show that although *C. albicans* strains isolated from different sites of the humans body have the same predominant alleles, they do not have the same predominant allele combinations.

4.7 Discussion

The aim of this particular study of the *HWP1* gene was to find out whether *HWP1* is a contingency gene, i.e. whether *HWP1* has a role in adaptation by changing the number of repeat units within the coding sequences. For this purpose, allelic distribution of *HWP1* in GPG and non-GPG strains, in commensal and infection strains, and in strains isolated from different sites of the humans body was examined and compared one with another.

4.7.1 Variability in the number of *HWP1* repeats

This study reveals that there is a variability in the number of *HWP1* repeats observed from different strains. From 96 infection strains, five different alleles were detected, which formed 10 different allele combinations. Likewise, five different alleles were detected in 35 commensal strains (GPG) which formed seven different allele combinations. From total 131 strains, there were five different alleles detected which form 11 different different allele combinations. The number of different *HWP1* alleles observed is less than that observed for *YWP1* (eight different alleles), and indeed it is the least compared to alleles observed for other genes summarized in Table 3.8 (see page 61). The least number of different *HWP1* alleles observed compared to other genes may indicate that *HWP1* has a relatively low mutation rate, consequently may have a role in adaptation. The mutation rate of *HWP1* will be discussed in Chapter 7, while the role of *HWP1* in adaptation will be discussed in the following section.

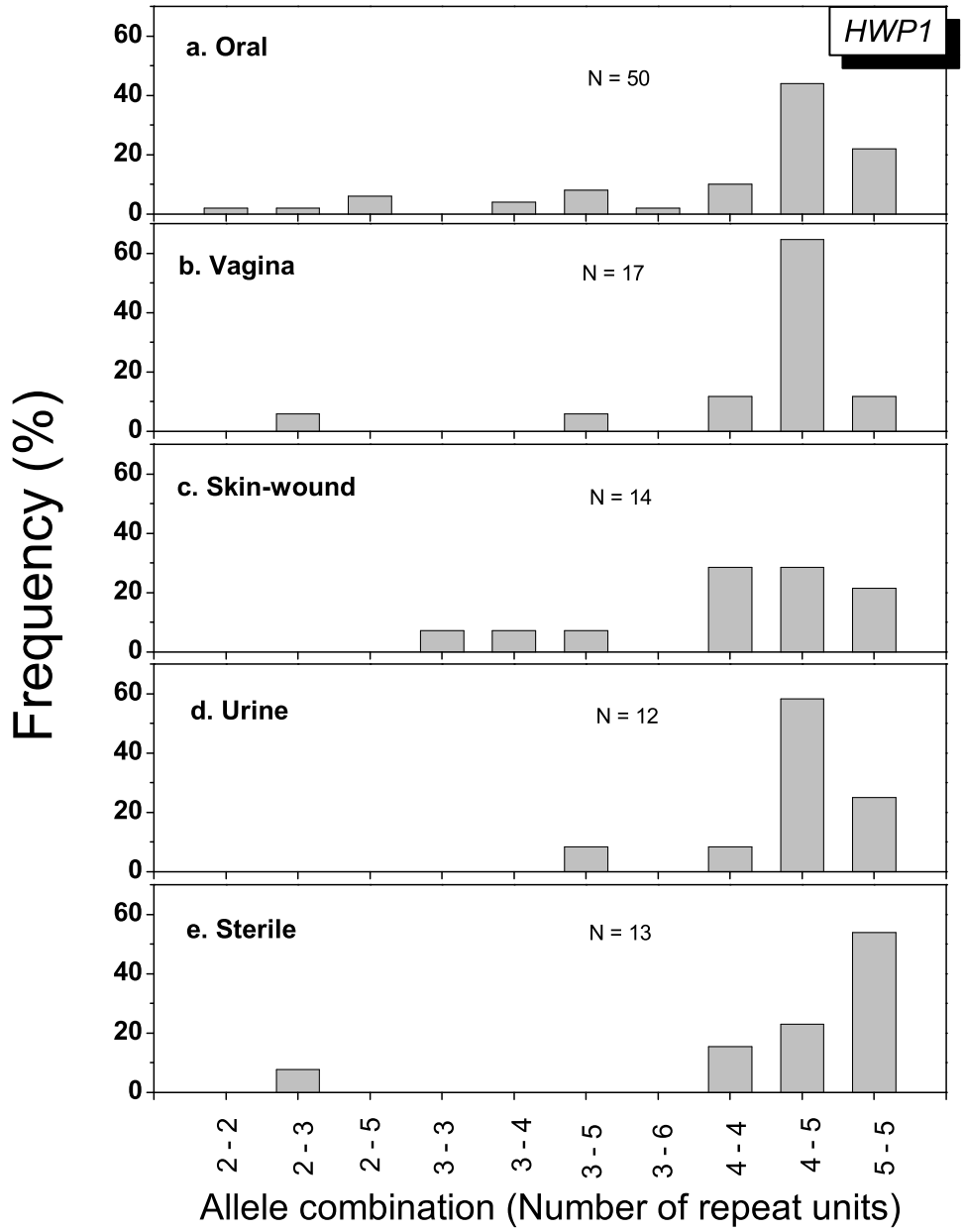

Figure 4.14: Distribution of allele combinations of *HWP1* in strains isolated from different sites of the humans body. N represents the number of strains at each site. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Alleles containing 4 and 5 repeat units predominated at oral, vagina, and urine sites, while alleles containing 5 and 5 repeat units predominated at sterile sites (the Chi-square goodness of fit test, $p < 0.001$); no obvious pattern of any predominant alleles observed at skin sites.

The *HWP1* allele variability has been suggested to play an important role in the function of HWP1p. It has been found that *HWP1*-2, a variant of *HWP1*, lacks some of the repeat regions including the repeat region investigated in this study, which affects the conformation of the protein and the interaction of the protein with binding partners. The homozygous strain of this variant had decreased biofilm production, hyphal formation, and the *HWP1*-2 gene expression. This differs from the homozygous strain of *HWP1*-1 variant, a variant of *HWP1*, which has the repeat regions lost in *HWP1*-2 [88]. These results for variants of *HWP1* indicate that the repeat regions of *HWP1* play an important role in the function of this gene.

As has already been discussed in Chapter 3, the repeat variability of some genes of *C. albicans* appears to relate to the ability of the *C.albicans* cells to adhere to the host [27, 28, 29]. In relation to that, it has been suggested that HWP1p play a role in adhesion. For example, the HWP1p has been found to be an adhesin required for covalent attachment to host epithelial cells [62]. In addition, experimental study showed that deleting *HWP1* resulted in a significant reduction (80%) in adherent ability of the fungus to oral cells as compared to wild types [63]. Moreover, it has been found that HWP1p binds to ALS1p and ALS3p which indicates that *HWP1* facilitates of hypha adhesion one with another, an important factor for biofilm formation [64].

4.7.2 GPG and non-GPG strains have the same predominant alleles and allele combinations of *HWP1*

The results of allelic characterization of the *HWP1* gene for GPG and non-GPG strains showed that these two different groups of infection strains have the same predominant alleles and allele combinations (see Figure 4.2 on page 74 and Figure 4.4 on page 76). The allele combinations 4-4, 4-5, and 5-5 predominated in both GPG and non-GPG strains, where the number of allele combinations 4-5 was the largest. There was no significant difference in the distribution of these alleles and allele combinations in the two groups of strains. These results suggest that *HWP1* is a contingency gene, i.e. *HWP1* has a role in adaptation.

There is a clear difference between the allelic distribution of *HWP1* and *YWP1*. First, GPG and non-GPG strains have different predominant allele combinations of *YWP1*, but the same predominant allele combinations of *HWP1*; the results of *HWP1* show that alleles containing 4 and 5 repeat units are important, but the results of *YWP1* show that only alleles containing 5 repeat units are important. Second, although GPG strains with similar genetic backgrounds (less diversity), and non-GPG strains with diverse genetic backgrounds (high diversity), the percentages of the predominant allele combinations of *HWP1* in GPG and non-GPG strains are similar. In contrast, the percentage of the predominant allele combinations of *YWP1* depends on the genetic diversity of a group of strains. The allelic characterization results of the *HWP1* gene are also different from the results of other repeat containing genes of *C. albicans* such as *ALS7* [32], *PNG2* [51], and *SSR1* [52], which showed predominant allele combinations in GPG strains, but no obvious predominant allele combinations in non-GPG strains. As far as the author's concern, the same predominant allele combinations of the *C. albicans* genes shown by GPG and non-GPG strains have not been observed: *HWP1* of this study is the first repeat-containing gene to show this phenomenon.

The allelic distributions of *HWP1* in GPG and non-GPG strains indicate a strong selective pressure on the gene. It has been suggested that if there is a selective pressure on the gene, the influence of the genetic background is diminished [57]. The allelic distributions of *YWP1* in GPG and non-GPG strains also indicate a selective pressure on the gene, but not as strong as it is on *HWP1*. However, the allelic distributions of *ALS7* [32], *PNG2* [51], and *SSR1* [52] in GPG and non-GPG strains showed no indication of a selective pressure on these genes.

The HWP1 protein has been found to be an adhesin required for covalent attachment to host epithelial cells. It serves as a substrate for mammalian transglutaminases, which permits the covalent cross-linking of *C. albicans* to epithelial cells [62]. The repeats of the *HWP1* protein of this study are part of N-terminal region that is predicted to be at the protein surface. This region functions as a substrate for mammalian transglutaminases that links it to other proteins on host epithelial cells.

The correct formation of the covalent cross-linking between the cells may associate with a certain number of repeat units. This may explain why alleles containing 4 and 5 repeats of *HWP1* were selected by two different groups of strains.

4.7.3 The number of *HWP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic

The comparison of the allelic distribution of the *HWP1* gene between commensal and infection strains showed that there is no significant difference between the distributions; both have the same predominant alleles (see Figure 4.6 on page 80). However, there is a noticeable difference in the distribution of allele combinations between commensal and infection strains (see Figure 4.8 on page 84); there is a decrease in proportion of allele combinations 4-5, and an increase in allele combinations 4-4, and 5-5 as one moves from the distribution in commensal strains to the distribution in infection strains. In spite of this, the results of the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between commensal and infection strains. These results showed that commensal and infection strains have the same predominant allele combinations.

The same predominant alleles and allele combinations of the *HWP1* gene in commensal and infection strains indicate no association between the *HWP1* alleles and the immune status of the host. This suggests that the number of *HWP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic. For example, Figure 4.9 (on page 85) showed the case for strains isolated from oral sites, where there was no significant difference in the distribution of allele combinations between commensal and infection strains, i.e. at the same site of the humans body, when the immune status of the host changes, there is no requirement for the alleles to alter. It seems that no allele change is required in order to achieve the correct optimal binding between *C. albicans* hyphal cells and cells at oral sites, although the *C. albicans* state has changed from commensal to pathogenic. Expression of HWP1 mRNA was found to be equivalent in both commensal and pathogenic states [89]. Since HWP1p is only

found in the hyphal form, not in the yeast form, the presence of HWP1 mRNA in a commensal state implies the presence of hyphal form in that state. This indicates that a HWP1p plays a role in both colonization of healthy individuals and in the invasion of immunocompromised individuals.

That commensal and infection strains have the same predominant allele combinations of the *HWP1* gene is in agreement with that of the *YWP1* gene. In addition, the diversity of the alleles of the *HWP1* gene is similar in commensal and infection strains, similar to those in the *YWP1* gene (see Table 4.5 on page 83 and Figure 4.10 on page 86). However, the diversity of allele combination of the *HWP1* gene is different in commensal and infection strains, i.e. infection strains have more diverse allele combinations compared to commensal strains. This is in contrast to the results of the *YWP1* gene, where the diversity of allele combinations is similar in both commensal and infection strains.

4.7.4 The number of *HWP1* repeats may require to alter when *C. albicans* moves to particular sites of the humans body

This study showed no significant difference in the distribution of *HWP1* alleles for any humans body sites sampled, i.e. alleles containing 4 and 5 repeat units predominated at all sites, and there was no significant difference in the distribution of these alleles for any humans sites sampled (see Figure 4.13 on page 92). This suggests that the number of *HWP1* repeats may not be required to alter when *C. albicans* moves to particular sites of the humans body. However, the comparison of the distribution of allele combination of *HWP1* showed that strains from some particular sites have specific allele combinations. The obvious pattern is that allele combinations containing 4 and 5 repeat units are specific to oral, vagina, and urine sites, while allele combinations containing 5 and 5 repeat units are specific to sterile sites. These results may indicate that no alteration of the *HWP1* allele is required when the cells move from one site to another, except when the cell reach the sterile sites of the humans

body. From oral and vagina sites, the cells reach sterile sites via the blood stream. At oral sites, expression of alleles containing 4 and 5 repeat units might promote optimal proper binding between *C. albicans* hyphal cells and epithelial cells, while expression of alleles containing 5 repeat units might promote optimal proper binding between *C. albicans* hyphal cells and epithelial cells at sterile sites. The results are in contrast to the results of studies on bloodstream and non-bloodstream isolates which showed no difference in the distribution of allele combinations between bloodstream and non-bloodstream strains [85].

4.7.5 The homozygous and heterozygous alleles of *HWP1*

The analysis of the combination of the two alleles of *HWP1* in an individual strain showed that *HWP1* exhibits excesses of homozygotes in GPG and non-GPG strains, but the excesses of homozygotes is not significant (see Table 4.6 on page 89). However, the number of homozygotes in commensal strains is significantly smaller than those expected by chance. In other words, commensal strains had a significant excess of heterozygotes. Excess of heterozygotes of some loci in *C. albicans* has been found and was maintained between clades. Heterozygosity may be selected for some loci to maintain genetic diversity that confers advantages to *C. albicans*. The results for the commensal strains are in contrast to the results for the *YWP1* gene, where commensal strains had a significant excess of homozygotes.

The analysis also showed that the two alleles of *HWP1* in a GPG strain and a non-GPG strain did not differ significantly from those expected by chance. The result for the GPG strains is in contrast to the result for *YWP1*, where the two alleles in a GPG strain are significantly more similar to each other than those expected by chance, but it is similar to the result for non-GPG strains, i.e. the two alleles of *YWP1* in a non-GPG strain did not differ significantly from those expected by chance. In addition, the two alleles of *HWP1* in a commensal strain are significantly more different to each other than those expected by chance. This result is in contrast to the result for *YWP1*, where the two alleles of *YWP1* are significantly more similar to each other than those expected by chance.

Chapter 5

Alleles of the *EAP1* Gene

This chapter describes the results of the allelic characterization of the *EAP1* gene in strains of interest, which is aimed to determine whether *EAP1* is a contingency gene, i.e. whether *EAP1* has a role in adaptation by changing the number of repeat units within the coding sequences. For this purpose, allelic distribution of *EAP1* in GPG and non-GPG strains, in commensal and infection strains, and in strains isolated from different sites of the human body will be compared one to another. For a contingency gene, GPG and non-GPG strains, two groups of strains with different genetic backgrounds, should select the same alleles, which are advantageous alleles. A comparison of the allelic distributions of *EAP1* in commensal and infection strains, and in strains isolated from different sites of the human body, respectively, is to observe whether *EAP1* acts as a contingency gene, when *C. albicans* changes from a commensal state to a pathogenic state, and when it moves to particular sites of the human body.

The chapter begins with a description of the identification of the repeat units in *EAP1*, where two regions in the gene were identified to contain repeat units. This is followed by two sections containing the results of allelic characterization of *EAP1* for each repeat region. Each of these two sections contains the results of allelic characterization for GPG and non-GPG strains (infection strains), and for the commensal strains, for comparison with the infection strains (both GPG strains). Analysis of the diversity of the alleles and allele combinations, and analysis of combination of

the two alleles in an individual strain for all strains of interest are explained before description of the results of allelic characterization of *EAP1* for strains isolated from different sites of the human body. The chapter ends with a discussion of the biological implications of the results.

5.1 Identification of the Repeat Units in the *EAP1* Gene

The repeat units in *EAP1* were identified using MacVector software (MacVector Inc, www.macvector.com) to analyse the DNA sequence of *EAP1* from strain SC5314 downloaded from www.candidagenome.org, i.e. the same procedure used to identify repeats in *YWP1* and *HWP1*. The result of this analysis is shown in Figure 5.1.

Primers EAP1AF and EAP1AR (see Table 2.5 on page 24) were designed to amplify the region from nucleotide 858 - 1725 (868 bp) which is a region containing repeat units, and for convenience is called repeat region 1. Primers EAP1BF and EAP1BR (see Table 2.5 on page 24) were designed to amplify the region from nucleotide 239 - 710 (472 bp) which is a region containing repeat units, and for convenience is called repeat region 2.

5.2 Results of Allelic Characterization of the *EAP1* Gene For Repeat Region 1

Repeat region 1 of *EAP1* from each strain of interest was amplified by PCR, and the amplicons were characterized by length using genotyping. As indicated previously, since *C. albicans* is diploid, the results are described for both alleles and allele combinations.

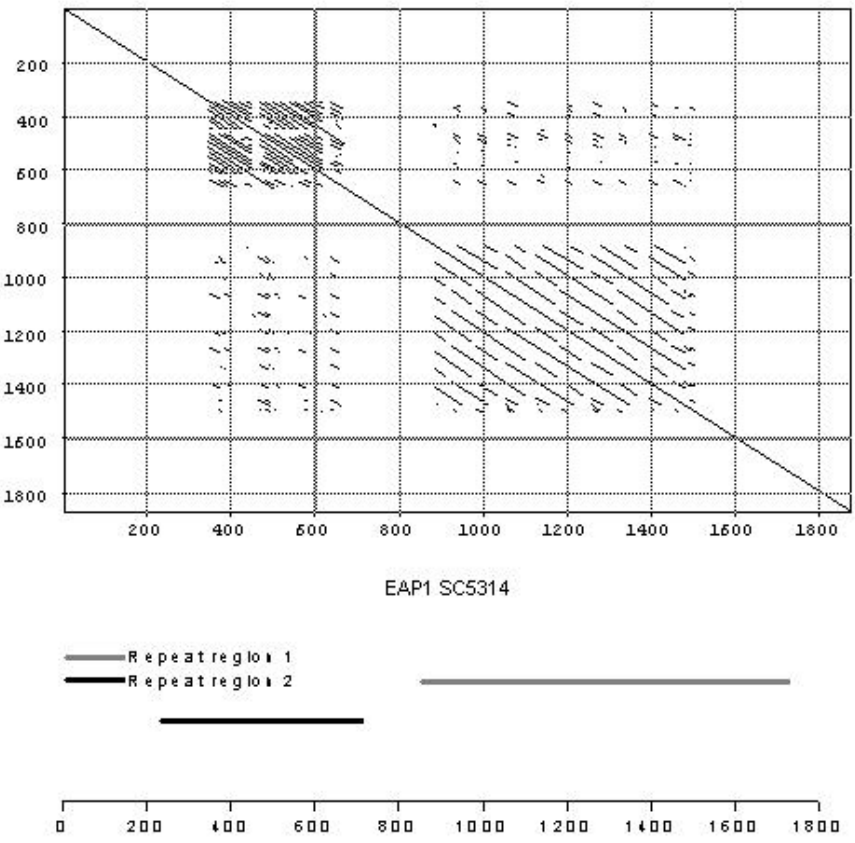

Figure 5.1: Identification of the repeat unit in *EAP1* using MacVector software (top figure). The identification was performed on the DNA sequence of *EAP1* from strain SC5314 downloaded from www.candidagenome.org. The grey and black thick lines indicate regions amplified, i.e. nucleotide 858 - 1725 (868 bp) for region 1 and nucleotide 239 - 710 (472 bp) for region 2 (bottom figure).

5.2.1 Allelic Characterization of the *EAP1* Gene For Repeat Region 1 in GPG and Non-GPG Strains (Infection Strains)

Based on allele length, 13 different alleles were found in GPG strains, and 28 in non-GPG strains. In total, there were 34 different alleles found in infection strains. In order to determine the number of repeat units for each fragment length, ideally all different fragments would be sequenced. However, most of the repeat region lengths observed were larger than 700 bp, which according to the manufacturer's instructions is much larger than the length that the instrument used for this measurement can sequence. For this reason, only some strains were sequenced. Primers EAP1AF and EAP1AR (see Table 2.5 on page 24) were designed for the sequencing. The repeat unit identified in DNA sequencing was 60 bp in length which encodes the 20 amino acid sequence TPAAPGTPVESQPVIPGTET. The number of repeat units in each fragment is determined by the ratio of the length of the repeat region (i.e. fragment that consists of repeat units) to the length of a repeat unit (60 bp). The length of the repeat region is calculated by taking the difference between total fragment length (determined by genotyping) and the part of each fragment that does not consist of repeat units. The part of each fragment that did not consist of repeat units was identified from the DNA sequencing and is assumed to be constant. Thus, the number of repeat units N_{RU} was calculated from the fragment length F_l as,

$$N_{RU} = \frac{F_l - L_o}{L_u}, \tag{5.1}$$

where L_o and L_u are the length of the fragment that did not consist of repeat units, and the length of a repeat unit, respectively. The number of repeat units N_{RU} in each fragment length was calculated using Equation 5.1, where the values of L_o and L_u bp were taken from the results of the sequencing. Not all calculated N_{RU} values were integers, and non-integer values were rounded to the nearest integer. Due to this rounding, there were several cases where alleles differed in fragment length were estimated to have the same number of repeat units. Consequently, the number of different alleles in GPG strains reduced from 13 by fragment length to eight by number

of repeat units, and in non-GPG strains from 28 to 14. Based on the number of repeat units, there were 14 different alleles found in infection strains. The frequency of alleles of *EAP1* for repeat region 1 in GPG and non-GPG strains is summarized in Table 5.1 and presented graphically in Figure 5.2.

Table 5.1: Frequency of alleles of *EAP1* for repeat region 1 in GPG and non-GPG strains. Alleles are represented by the fragment length and the number of repeat units.

No	Allele		Frequency		
	Length (bp)	No. of repeats	GPG strains	Non-GPG strains	Total
1	408	2	0	8	8
2	468	3	0	7	7
3	539	4	0	3	3
4	598	5	0	7	7
5	646	6	0	6	6
6	729	7	4	6	10
7	800	8	1	20	21
8	883	9	10	4	14
9	930	10	43	3	46
10	1000	11	35	15	50
11	1082	12	1	1	2
12	1123	13	3	10	13
13	1207	14	1	3	4
14	1307	15	0	1	1
Total			98	94	192

Figure 5.2 shows that alleles containing 10 and 11 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These alleles were overrepresented in GPG strains (43 and 35%), compared to non-GPG strains (3 and 15%), i.e. there is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). Alleles containing 8 and 11 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.01$), and alleles containing 8 repeat units were overrepresented in non-GPG strains, compared to GPG strains (the Chi-square test for contingency tables, $p < 0.001$). Similar results were also observed in strains isolated from oral sites as shown in Figure 5.3.

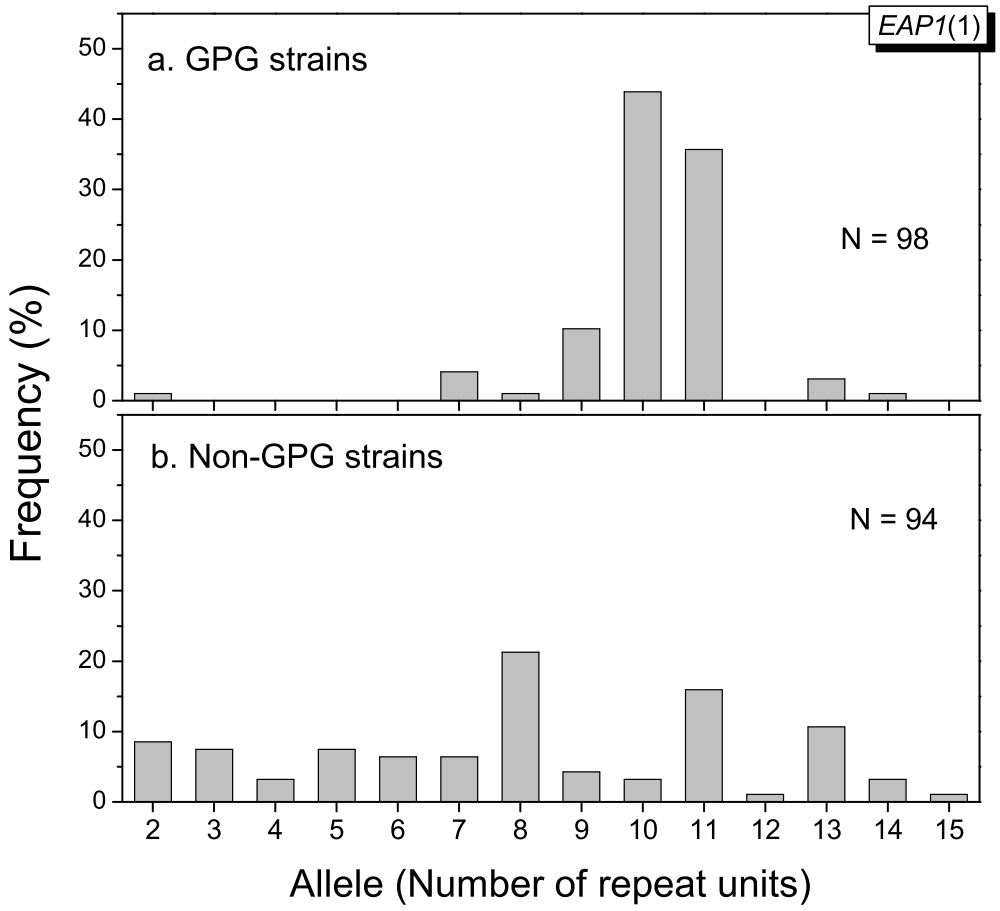

Figure 5.2: Distribution of alleles of *EAP1* for repeat region 1 in: a. GPG strains, and b. Non-GPG strains. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 10 and 11 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). Alleles containing 8 and 11 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.01$), and alleles containing 8 repeat units were overrepresented in non-GPG strains, compared to GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

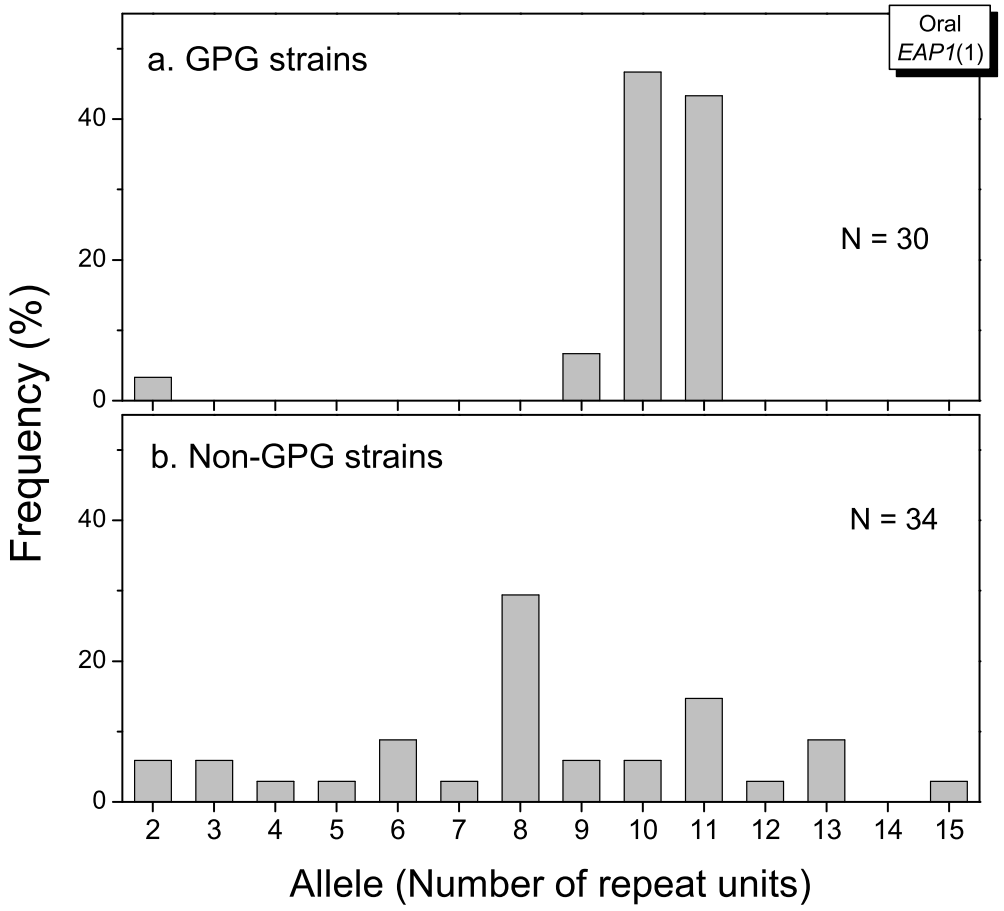

Figure 5.3: Distribution of alleles of *EAP1* for repeat region 1 for: a. GPG strains, and b. Non-GPG strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 10 and 11 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). Alleles containing 8 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.01$), and were over-represented in non-GPG strains, compared to GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

Based on allele length, the number of different allele combinations found in GPG strains was 15, compared to 34 combinations in non-GPG strains. In total, there were 46 different allele combinations found in infection strains. As some alleles that differ in length have the same number of repeats, the number of allele combinations by number of repeats were 12 and 28 for GPG and non-GPG strains, respectively. In total, there were 35 different allele combinations by number of repeats found in infection strains. The frequency of allele combinations of *EAP1* for repeat region 1 found in infection strains is summarized in Table 5.2, and presented graphically in Figure 5.4.

Figure 5.4 shows that allele combinations containing 10 and 11 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These allele combinations were overrepresented in GPG strains (59%), compared to non-GPG strains (2%), i.e. there is a significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant allele combinations observed in non-GPG strains. Similar results were also observed in strains isolated from oral sites as shown in Figure 5.5.

5.2.2 Allelic Characterization of the *EAP1* Gene For Repeat Region 1 in Commensal Strains

Based on allele length, eight different alleles were found in 29 commensal strains. The number of repeat units in each fragment was determined using Equation 5.1 (see page 103). Based on the number of repeat units, seven different alleles were found in commensal strains, compared to eight in infection strains. The frequency of alleles of *EAP1* for repeat region 1 found in commensal strains is summarized in Table 5.3, and presented graphically in Figure 5.6. For comparison, the data from infection strains have been added in the table and the figure.

Figure 5.6 shows that alleles containing 10 and 11 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p <$

Table 5.2: Frequency of allele combinations of *EAP1* for repeat region 1 for GPG and non-GPG strains. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

	Allele Combination		Frequency		
No	Length (bp)	No. of repeats	GPG Strains	Non-GPG Strains	Total
1	408-408	2-2	0	4	4
2	408-1000	2-11	1	0	1
3	468-729	3-7	0	1	1
4	473-883	3-9	0	1	1
5	460-1123	3-13	0	5	5
6	539-752	4-7	0	1	1
7	539-800	4-8	0	1	1
8	527-1000	4-11	0	1	1
9	598-598	5-5	0	1	1
10	598-800	5-8	0	3	3
11	611-1123	6-13	0	1	1
12	586-1206	5-14	0	1	1
13	646-646	6-6	0	2	2
14	646-788	6-8	0	1	1
15	660-1020	6-11	0	1	1
16	729-729	7-7	1	0	1
17	729-800	7-8	0	2	2
18	729-930	7-10	1	0	1
19	729-1000	7-11	1	0	1
20	741-1123	7-13	0	1	1
21	729-1207	7-14	0	1	1
22	800-800	8-8	0	5	5
23	800-860	8-9	0	1	1
24	800-1000	8-11	0	1	1
25	788-1082	8-12	0	1	1
26	800-1123	8-13	1	0	1
27	859-859	9-9	4	0	4
28	859-977	9-11	2	2	4
29	930-930	10-10	6	1	7
30	930-1000	10-11	29	1	30
31	930-1198	10-14	1	0	1
32	1000-1000	11-11	1	4	5
33	1000-1307	11-15	0	1	1
34	1127-1127	13-13	1	1	2
35	1151-1211	13-14	0	1	1
Total			49	47	96

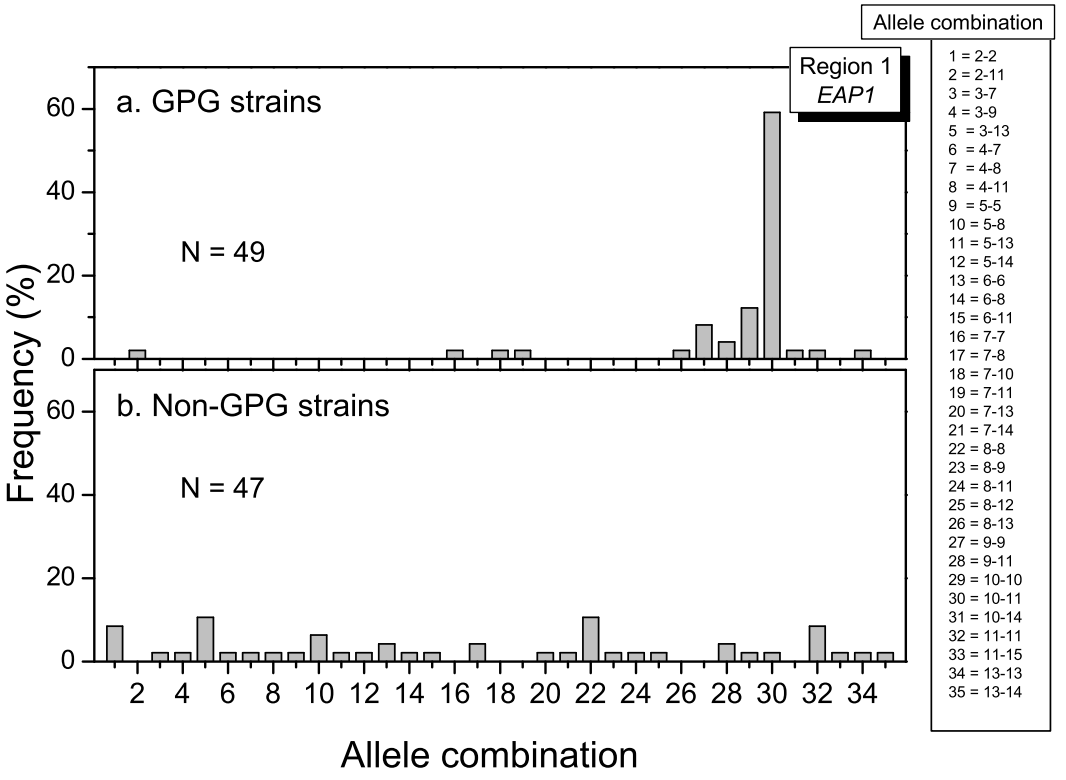

Figure 5.4: Distribution of allele combinations of the *EAP1* gene for repeat region 1 for: a. GPG strains, and b. Non-GPG strains. N represents the number of strains in a particular group. Due to the limitation of space in the x axis, allele combinations are written as numbers from 1 to 35 and the legend on the right shows the associated allele combinations. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 10 and 11 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These allele combinations were overrepresented in GPG strains compared to non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

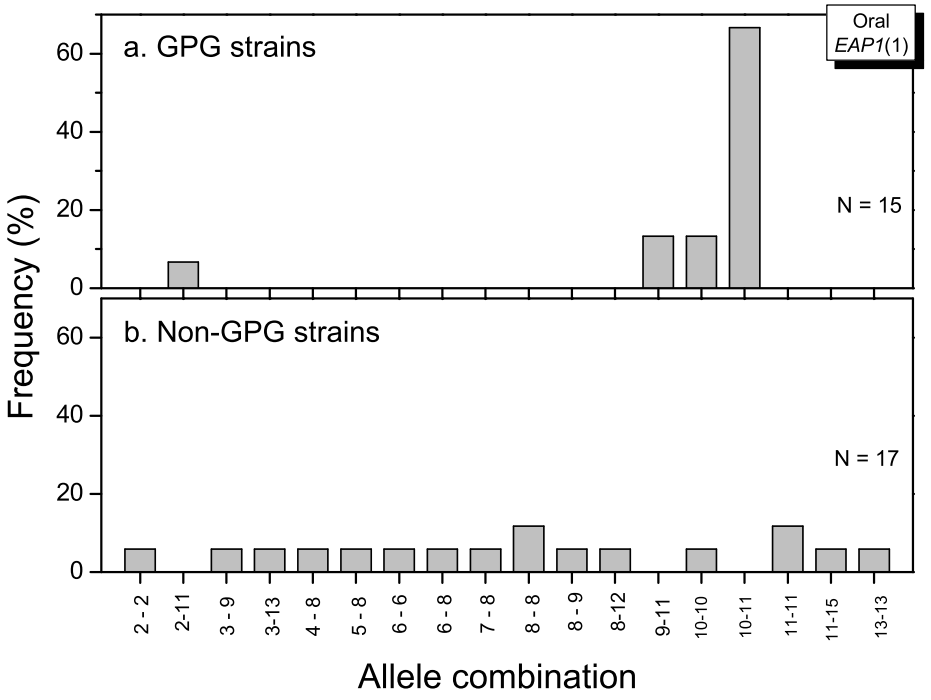

Figure 5.5: Distribution of allele combinations of *EAP1* for repeat region 1 for: a. GPG strains, and b. Non-GPG strains, where all strains were isolated from oral sites. N represents total number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 10 and 11 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These allele combinations were overrepresented in GPG strains compared to non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

0.001), and there is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$). Similar results were also observed in strains isolated from oral sites as shown in Figure 5.7. These results show that commensal and infection strains have the same predominant alleles.

Table 5.3: Frequency of alleles of *EAP1* for repeat region 1 in commensal strains. Alleles are represented by the fragment length and the number of repeats. For comparison, the data from infection strains have been added.

	Allele		Frequency		
No	Length (bp)	No. of repeats	Commensal Strains	Infection Strains	Total
1	408	2	0	1	1
2	729	7	1	4	5
3	800	8	1	1	2
4	883	9	2	10	12
5	930	10	25	43	68
6	1000	11	27	35	62
7	1082	12	1	0	1
8	1123	13	1	3	4
9	1207	14	0	1	1
Total			58	98	156

Based on allele length, the number of different allele combinations found in commensal strains was nine, compared to 15 in infection strains. However, based on the number of repeats, the number of different alleles in commensal strains was eight, compared to 12 in infection strains. The frequency of allele combinations of the *EAP1* gene for repeat region 1 found in commensal strains is summarized in Table 5.4, and presented graphically in Figure 5.8. For comparison, the data from infection strains have been added in the table and the figure.

Figures 5.8 shows that allele combinations containing 10 and 11 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). The result from the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between commensal and infection strains ($p > 0.05$). Similar results were also observed in strains

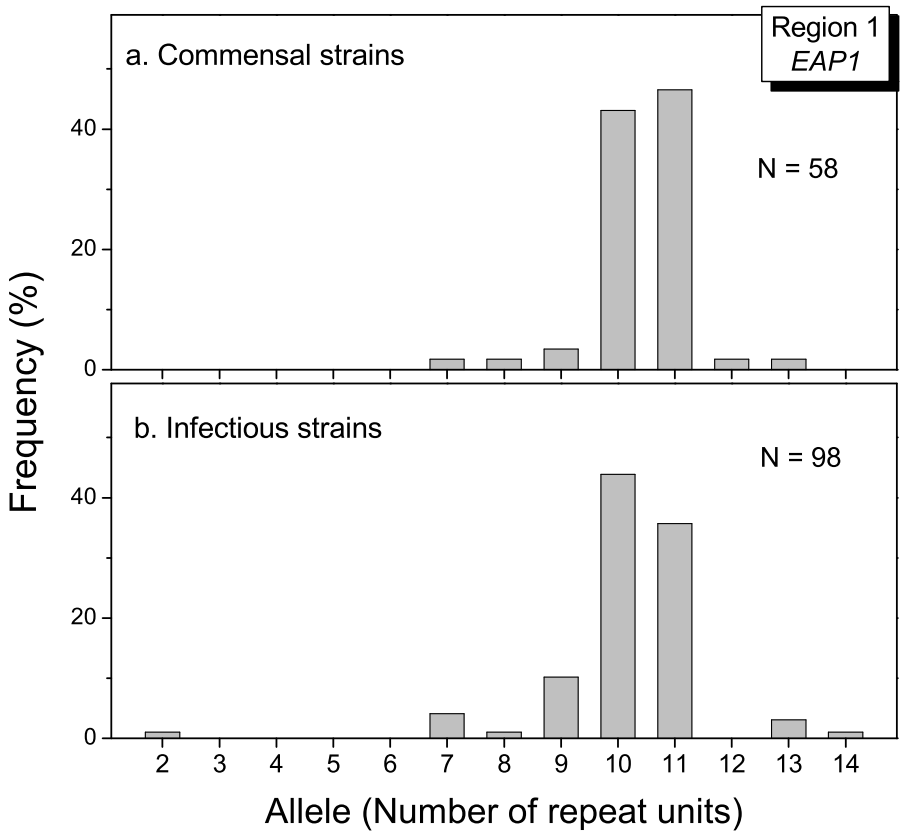

Figure 5.6: Distribution of alleles of *EAP1* for repeat region 1 for: a. Commensal strains, and b. Infection strains. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 10 and 11 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

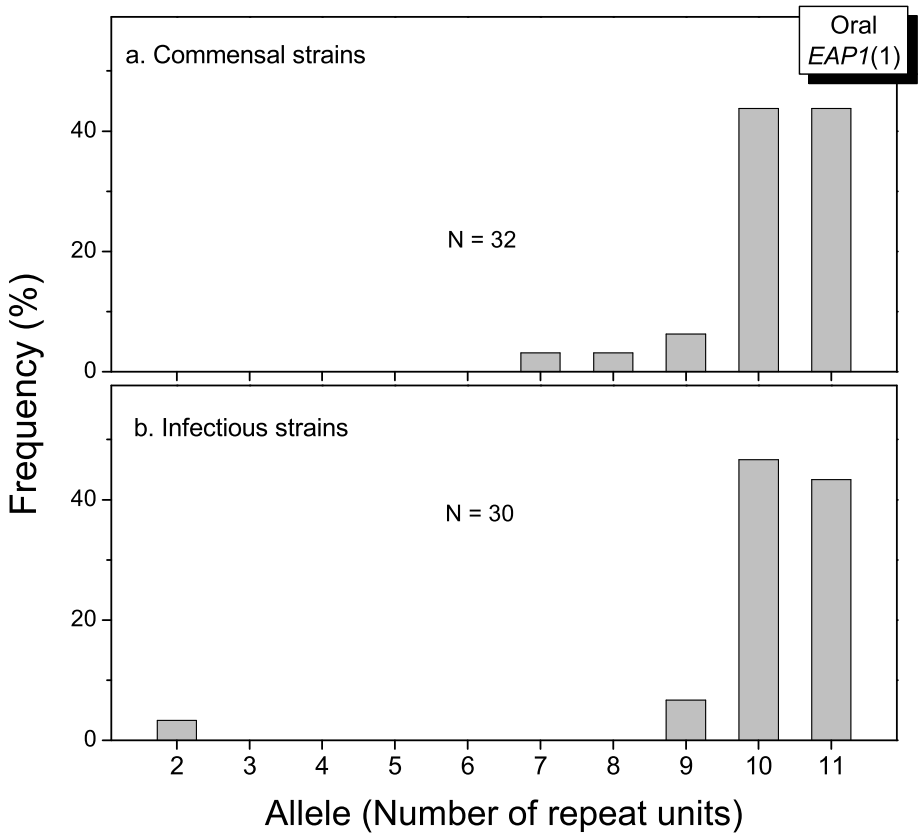

Figure 5.7: Distribution of alleles of *EAP1* for repeat region 1 for: a. Commensal strains, and b. Infection strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in a particular group, as there are 2 alleles per strain. Alleles containing 10 and 11 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

isolated from oral sites as shown in Figure 5.9. These results show that commensal and infection strains have the same predominant allele combinations.

Table 5.4: Frequency of allele combinations of *EAP1* for repeat region 1 in commensal strains. For comparison, the data from infection strains have been added. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

	Allele Combination		Frequency		
No	Length (bp)	No. of repeats	Commensal Strains	Infection Strains	Total
1	408-1000	2-11	0	1	1
2	729-729	7-7	0	1	1
3	729-930	7-10	0	1	1
4	729-1000	7-11	1	1	2
5	800-860	8-9	1	0	1
6	800-1123	8-13	0	1	1
7	859-859	9-9	0	4	4
8	859-930	9-10	1	0	1
9	859-977	9-11	0	2	2
10	930-930	10-10	4	6	10
11	930-1000	10-11	16	29	45
12	930-1198	10-14	0	1	1
13	1000-1000	11-11	4	1	5
14	1000-1060	11-12	1	0	1
15	1000-1199	11-13	1	0	1
16	1127-1127	13-13	0	1	1
Total			29	49	78

5.2.3 Diversity of Alleles of the *EAP1* Gene For Repeat Region 1

Allele diversity is quantified by the index of diversity λ which is defined by Equation 3.1 (see page 48). The λ values for both the alleles and the allele combinations of *EAP1* for repeat region 1 for GPG, non-GPG, and commensal strains are summarized in Table 5.5 and presented graphically in Figure 5.10.

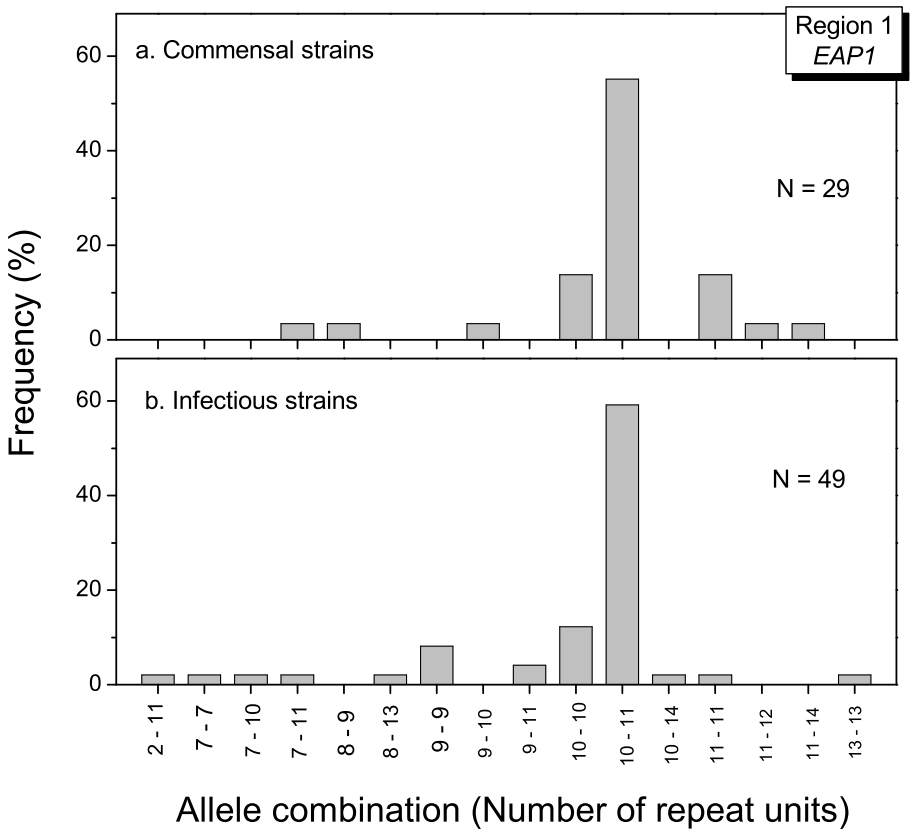

Figure 5.8: Distribution of allele combinations of *EAP1* for repeat region 1 for: a. Commensal strains, and b. Infection strains. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 10 and 11 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains. (the Chi-square test for contingency tables, $p > 0.05$).

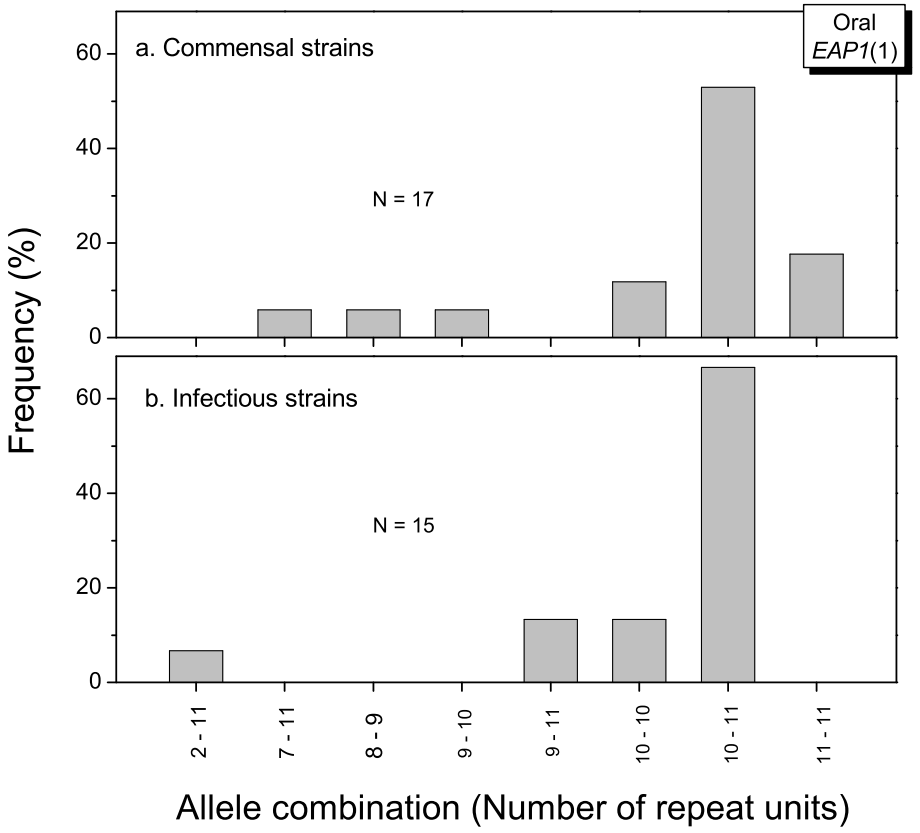

Figure 5.9: Distribution of allele combinations of *EAP1* for repeat region 1 for: a. Commensal strains, and b. Infection strains, where all strains were isolated from oral sites. N represents the number of strains in a particular group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 10 and 11 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains. (the Chi-square test for contingency tables, $p > 0.05$).

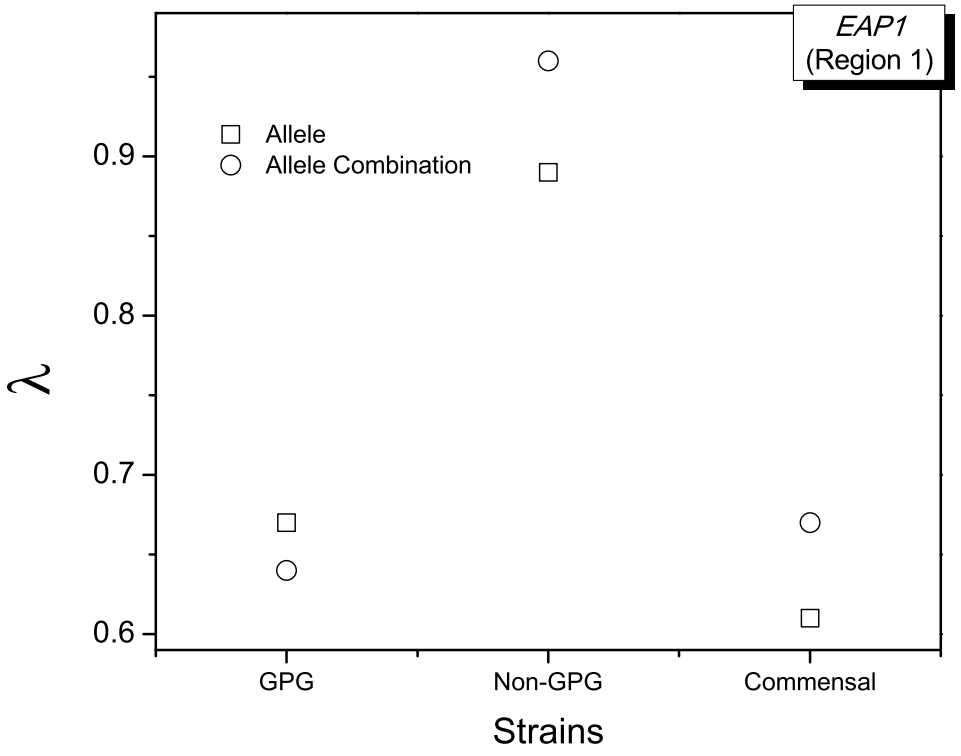

Figure 5.10: Index of diversity λ of the alleles and the allele combinations of *EAP1* for repeat region 1 for GPG, non-GPG, and commensal strains.

Table 5.5: Index of diversity of alleles and allele combinations of *EAP1* for repeat region 1 for GPG, non-GPG, and commensal strains.

Strains	Index of Diversity	
	Allele	Allele Combination
GPG	0.67	0.64
Non-GPG	0.89	0.96
Commensal	0.61	0.67

5.2.4 Analysis of the Combination of the Two Alleles of the *EAP1* Gene For Repeat Region 1 in An Individual Strain

In this section, the combination of the two alleles of *EAP1* for repeat region 1 in an individual strain among strains of interest is analysed. Similar to the description for *YWP1* and *HWP1* in the previous chapters, the section starts with a description of the distribution of the differences between the two alleles in an individual strain, and ends with a calculation of the mean of the differences between the two alleles in an individual strain among strains of interest.

Figure 5.11 shows the distribution of the differences in the number of repeat units between the two alleles of *EAP1* for repeat region 1 in an individual strain for GPG, non-GPG, and commensal strains. The percentages of homozygous strains found in GPG, non-GPG, and commensal strains are 26%, 38%, and 27%, respectively.

The number of homozygous alleles in each group was compared to those expected if the combination of the alleles was random, taking allele frequency into consideration. For this purpose, a matrix with a size equal to the square of 2 times the number of strains analysed was generated, i.e. the same procedure used for the analysis for the *YWP1* gene in the previous chapter (see section 3.3.4, page 50). The frequency of homozygotes and heterozygotes of *EAP1* (repeat region 1) for the data observed and those expected from random alleles is summarized in Table 5.6, and is presented graphically in Figure 5.12. The figure shows no significant difference between the number of homozygotes observed and those expected by chance for (GPG)

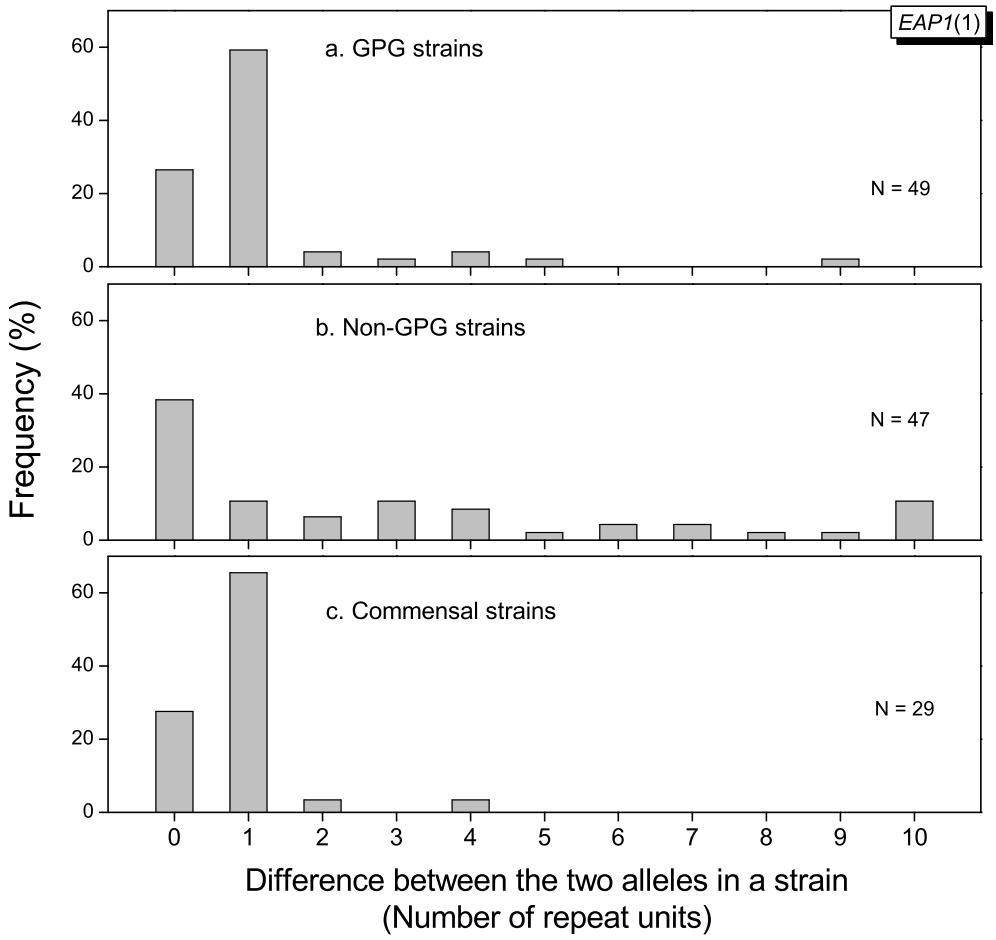

Figure 5.11: Distribution of differences in the number of repeat units between the two alleles of *EAP1* for repeat region 1 in an individual strain for: a. GPG strains, b. Non-GPG strains, and c. (GPG) Commensal strains. N represents the number of strains in a particular group.

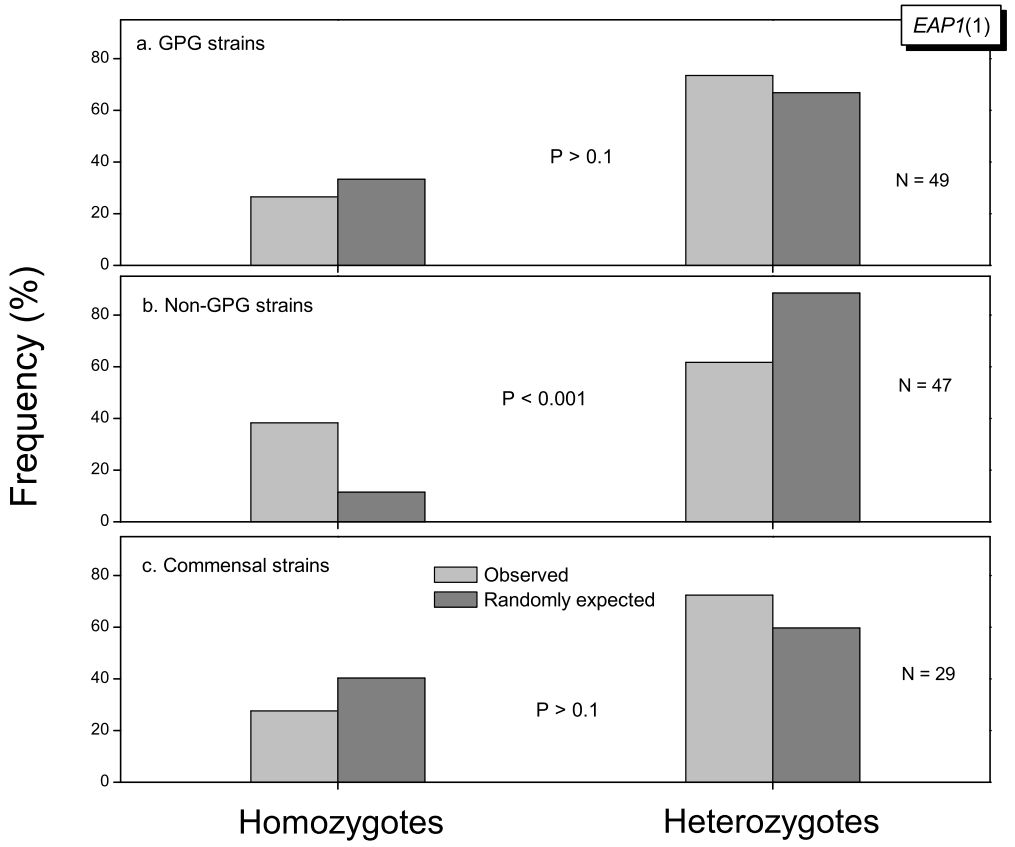

Figure 5.12: Frequency of homozygotes and heterozygotes of *EAP1* for repeat region 1 for GPG, non-GPG, and commensal strains. N represents the number of strains in a particular group. There is no significant difference between the number of homozygotes observed and those expected by chance (the Chi-square test for contingency tables , $p>0.1$) for (GPG) commensal and infection strains. However, for non-GPG strains, the number of homozygotes is significantly larger than those expected by chance (the Chi-square test for contingency tables, $p<0.001$), i.e. *EAP1*(1) exhibits significant excesses of homozygotes in non-GPG strains.

commensal and infection strains (the Chi-square test for contingency tables , p>0.1). However, for non-GPG strains, the number of homozygotes is significantly larger than those expected by chance (the Chi-square test for contingency tables, p<0.001), i.e. *EAP1*(1) in non-GPG strains exhibits significant excesses of homozygotes.

Table 5.6: The number of homozygous strains in GPG, non-GPG, and commensal strains compared to the expected number by chance. N indicates the number of strains.

Strains	Number of homozygotes		Significance (p)
	Observed data (N)	Expected data	
GPG	13 (49)	16	$p > 0.1$
Non-GPG	18 (47)	5	$p < 0.001$
Commensal	8 (35)	7	$p > 0.1$

The mean (± standard deviation) of the differences in the number of repeat units between the two alleles of the *EAP1* gene for repeat region 1 in an individual strain for GPG, non-GPG, and commensal strains, was determined to be 1.18±1.55, 2.98±3.45, and 0.86 ± 0.79, respectively. The result of the t-test between the means of GPG and non-GPG strains showed that alleles in a GPG strain are significantly more similar to each other than alleles in a non-GPG strain ($p < 0.001$). The result of the t-test between the means of (GPG) commensal and infection strains showed that the two alleles in a commensal strain did not differ significantly from those in an infection strain ($p > 0.1$).

To determine whether the difference between the two alleles in an individual strain is random or non-random (i.e. is a selective pressure operating ?), the means of the differences of the alleles found were compared to the means expected if all alleles combined randomly, taking allele frequency into consideration. As explained previously, a matrix with a size equal to the square of 2 times the number of strains analysed was generated.

For GPG strains, the mean (± standard deviation) of the differences in the number of repeat units between the two alleles expected from random alleles was determined to be 1.24 ± 2.32. The result of the t-test between this mean and the mean from

observed data of 1.18 ± 1.55 showed that the two alleles in an individual GPG strain did not differ significantly from those expected by chance ($p > 0.10$). For non-GPG strains, the mean (± standard deviation) of the differences in the number of repeat units between the two alleles expected from random alleles was determined to be 4.03 ± 2.96. The result of the t-test between this mean and the mean from observed data of 2.98 ± 3.45 showed that the two alleles in an individual non-GPG strain are significantly more similar to each other than those expected by chance ($p < 0.01$). For commensal strains, the mean (± standard deviation) of the differences in the number of repeat units between the two alleles expected from random alleles was determined to be 0.83 ± 0.91. The result of the t-test between this mean and the mean from observed data of 0.86 ± 0.79 showed that the two alleles in an individual commensal strain did not differ significantly from those expected by chance ($p > 0.10$),i.e. differences in number of repeats were random. The mean (± standard deviation) of the differences in the number of repeat units between the two alleles from observed data, data expected from random alleles, and the significance of the differences between the means are summarized in Table 5.7.

Table 5.7: The mean (± standard deviation) of the differences in the number of repeat units between the two alleles of the *EAP1* gene for repeat region 1 in an individual strain from observed data, data expected from random alleles, and the significance of the difference between the two means for infection and commensal strains. N indicates the number of strains.

Strains	Mean Differences (± SD)		Significance (p)
	Observed Differences (N)	Expected Differences (N)	
GPG	1.18 ± 1.55 (49)	1.24 ± 2.32 (9604)	$p > 0.10$
Non-GPG	2.98 ± 3.45 (47)	4.03 ± 2.96 (8836)	$p < 0.01$
Commensal	0.86 ± 0.79 (29)	0.83 ± 0.91 (3364)	$p > 0.1$

5.2.5 Allelic Distribution of the *EAP1* Gene For Repeat Region 1 in Strains Isolated From Different Sites of the Humans Body

Figure 5.13 shows the distribution of alleles of *EAP1* for repeat region 1 in strains isolated from different sites of the humans body. The distributions of alleles at all sites appear similar; alleles containing 10 and 11 repeat units predominated at each site (the Chi-square goodness of fit test, $p < 0.001$), except at sterile sites, where the percentage of alleles containing 9 repeat units is comparable to the percentages of alleles 10 and 11 repeat units. Result of the Chi-square test for contingency tables showed no significant difference in the distribution of alleles containing 10 and 11 repeat units (and also alleles containing 9 repeat units) between any two sites for all sites sampled ($p > 0.05$).

Figure 5.14 shows the distribution of allele combinations of *EAP1* for repeat region 1 in strains isolated from different sites of the humans body. The distributions of allele combinations at all sites appear similar; allele combinations containing 10 and 11 repeat units predominated at each site (the Chi-square goodness of fit test, $p < 0.001$), except at sterile sites, where the percentage of allele combinations containing 9 and 9 repeat units is comparable to that of 10 and 11 repeat units. The result of the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between any two sites for all sites sampled ($p > 0.05$).

5.3 Results of Allelic Characterization of the *EAP1* Gene For Repeat Region 2

Repeat region 2 of the *EAP1* gene from each strain of interest was amplified by PCR, and the amplicons were characterized by length using genotyping and gel electrophoresis. The use of gel electrophoresis in addition to genotyping was due to the fact that most of the repeat region lengths in non-GPG strains were larger than 1200

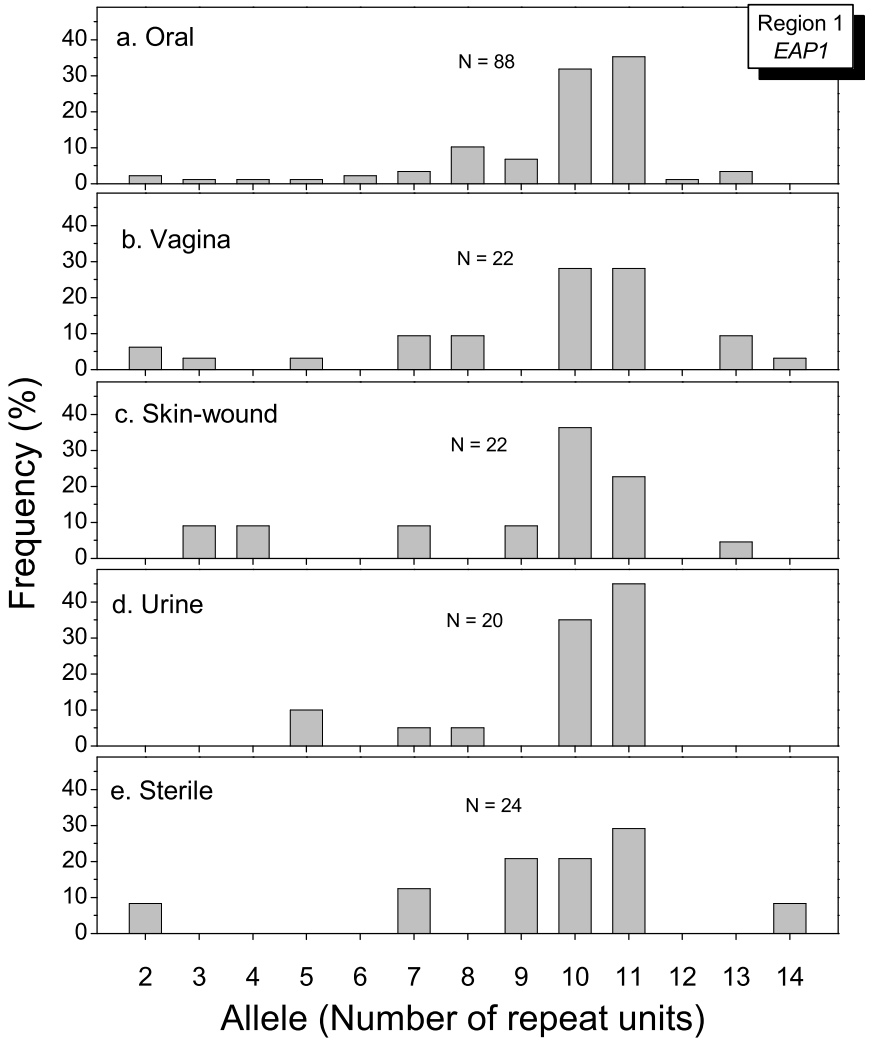

Figure 5.13: Distribution of alleles of *EAP1* for repeat region 1 in strains isolated from different sites of the humans body. N represents 2 times the number of strains, as there are 2 alleles per strain. Alleles containing 10 and 11 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$), except at sterile sites, where the percentage of alleles containing 9 repeat units is comparable to the percentages of alleles containing 10 and 11 repeat units. The Chi-square test for contingency tables showed no significant difference in the distribution of alleles containing 10 and 11 repeat units (and also alleles containing 9 repeat units) between any two sites for all sites sampled ($p > 0.05$).

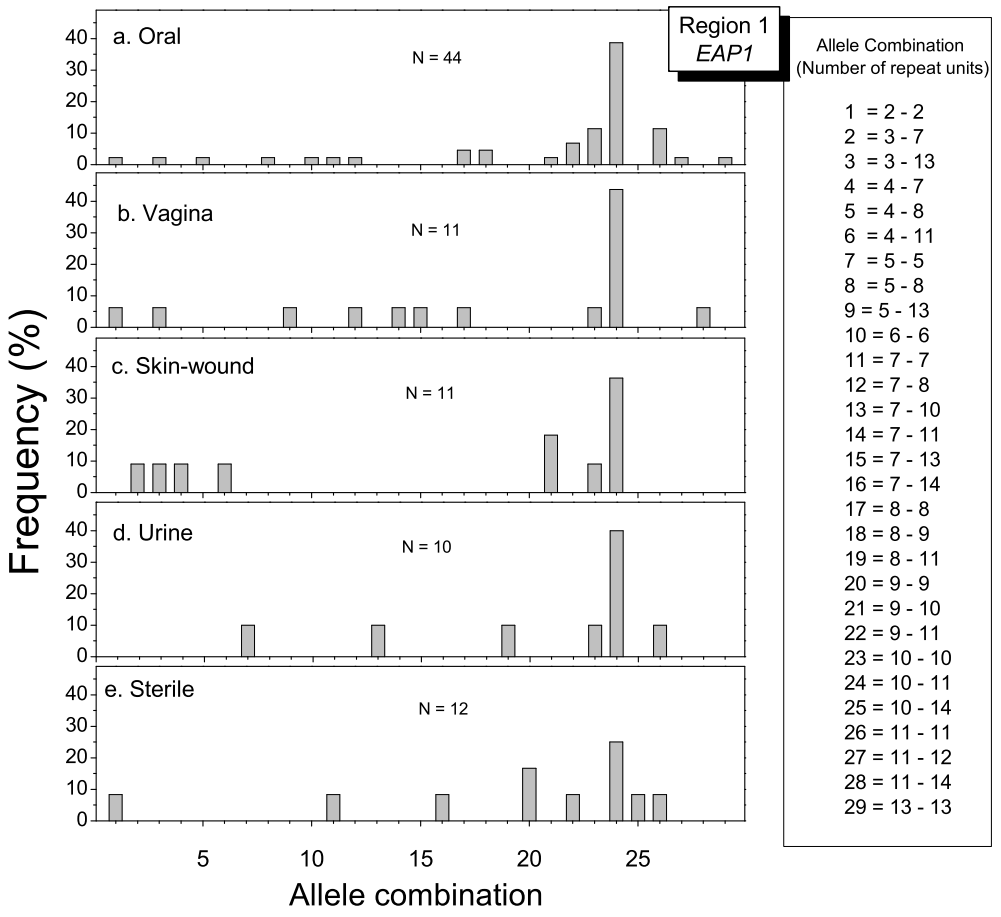

Figure 5.14: Distribution of allele combinations of *EAP1* for repeat region 1 in strains isolated from different sites of the humans body. N represents the number of strains for a particular site. Due to the limitation of space in the x axis, the allele combinations are written as numbers from 1 to 29 and the legend on the right shows the associated allele combinations. For convenience, allele combinations of x and y are written as x-y, where x and y are number of repeat units. Allele combinations containing 10 and 11 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$, except for sterile, $p < 0.05$). The Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between any two sites for all sites sampled ($p > 0.05$).

bp, which according to the manufacturer's instructions is much larger than the length that the genotyping instrument used in this study can quantify. As indicated previously, since *C. albicans* is diploid, the results are described for both alleles and allele combinations.

5.3.1 Allelic Characterization of the *EAP1* Gene For Repeat Region 2 in GPG and non-GPG Strains (Infection Strains)

Based on allele length, eight different alleles of *EAP1* for repeat region 2 were found in 48 GPG strains, and 22 in 46 non-GPG strains. In total, there were 25 different alleles found in 94 infection strains. In order to determine the number of repeat units for each fragment length, ideally all different fragments would be sequenced. However, most of the repeat region lengths observed were larger than 700 bp, which according to the manufacturers instructions is much larger than the length that the instrument used for this measurement can sequence. For this reason, only some strains were sequenced. Primers EAP1BF and EAP1BR (see Table 2.5 on page 24) were designed for the sequencing. The repeat unit identified in DNA sequencing was 18 bp in length which encodes the 6 amino acid sequence STPATE. The number of repeat units for each fragment length N_{RU} was calculated using Equation 5.1 (see page 103), where the values of L_o and L_u were taken from the results of the sequencing, and the non-integer values of N_{RU} were rounded to the nearest integer. It was found that the number of different alleles by the fragment length is the same as the number of different alleles by the number of repeat units. The frequency of alleles of *EAP1* for repeat region 2 found in GPG and non-GPG strains is summarized in Table 5.8, and presented graphically in Figure 5.15.

Figure 5.15 shows that alleles containing 33 and 48 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These alleles were overrepresented in GPG strains (44 and 35%), compared to non-GPG strains (2 and 2%) (the Chi-square test for contingency tables, $p < 0.001$). In addition, alleles

containing 83 and 102 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These alleles were overrepresented in non-GPG strains (23 and 27%), compared to GPG strains (2 and 0%) (the Chi-square test for contingency tables, $p < 0.001$). Similar results were also observed in strains isolated from oral sites as shown in Figure 5.16. These results show that GPG and non-GPG strains have distinct predominant alleles; alleles containing 33 and 48 repeat units are specific to GPG strains, while alleles containing 83 and 102 are specific to non-GPG strains.

Based on both allele length and the number of repeat units, eight different allele combinations were found in GPG strains and 22 in non-GPG strains. In total, there were 26 different allele combinations found in infection strains. The frequency of allele combinations of *EAP1* for repeat region 2 found in GPG and non-GPG strains is summarized in Table 5.9, and presented graphically in Figure 5.17.

Figure 5.17 shows that allele combinations containing 33 and 48 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These allele combinations were overrepresented in GPG strains (71%), compared to non-GPG strains (4%) (the Chi-square test for contingency tables, $p < 0.001$). In addition, allele combinations containing 83 and 83, and 102 and 102 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These allele combinations were overrepresented in non-GPG strains (41%), compared to GPG strains (2%) (the Chi-square test for contingency tables, $p < 0.001$). Similar results were also observed in strains isolated from oral sites as shown in Figure 5.18. These results show that GPG and non-GPG strains have distinct predominant allele combinations; allele combinations containing 33 and 48 repeat units are specific to GPG strains, while allele combinations containing 83 and 102 are specific to non-GPG strains.

Table 5.8: Frequency of alleles of *EAP1* for repeat region 2 in GPG and non-GPG strains. Alleles are represented by the fragment length and the number of repeat units.

	Allele		Frequency		
No	Length (bp)	No. of repeats	GPG Strains	Non-GPG Strains	Total
1	400	13	0	1	1
2	468	17	12	2	14
3	495	19	1	0	1
4	628	26	0	2	2
5	753	33	42	2	44
6	800	35	0	4	4
7	808	36	0	1	1
8	929	43	0	2	2
9	1000	47	0	3	3
10	1024	48	34	2	36
11	1100	52	0	2	2
12	1158	55	1	0	1
13	1300	63	0	2	2
14	1311	64	2	0	2
15	1325	65	0	4	4
16	1400	69	0	1	1
17	1650	83	2	21	23
18	2000	102	0	25	25
19	2058	105	0	1	1
20	2100	108	0	2	2
21	2200	113	2	2	4
22	2500	130	0	4	4
23	2600	135	0	2	2
24	3000	158	0	5	5
25	4072	228	0	2	2
Total			96	92	188

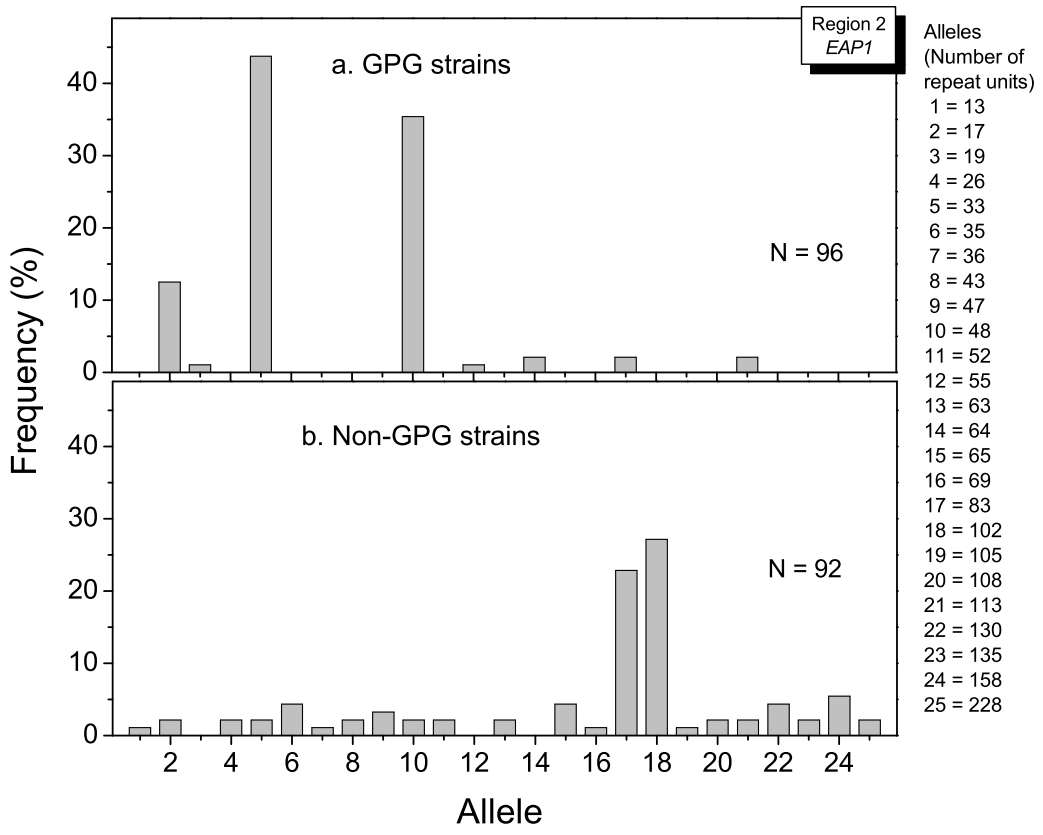

Figure 5.15: Distribution of alleles of *EAP1* for repeat region 2 for: a. GPG strains, and b. Non-GPG strains. N represents 2 times the number of strains in each group, as there are 2 alleles per strain. Due to the limitation of space in the x axis, alleles were written as numbers from 1 to 25 and the legend on the right shows the associated alleles. Alleles containing 33 and 48 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). In addition, alleles containing 83 and 102 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

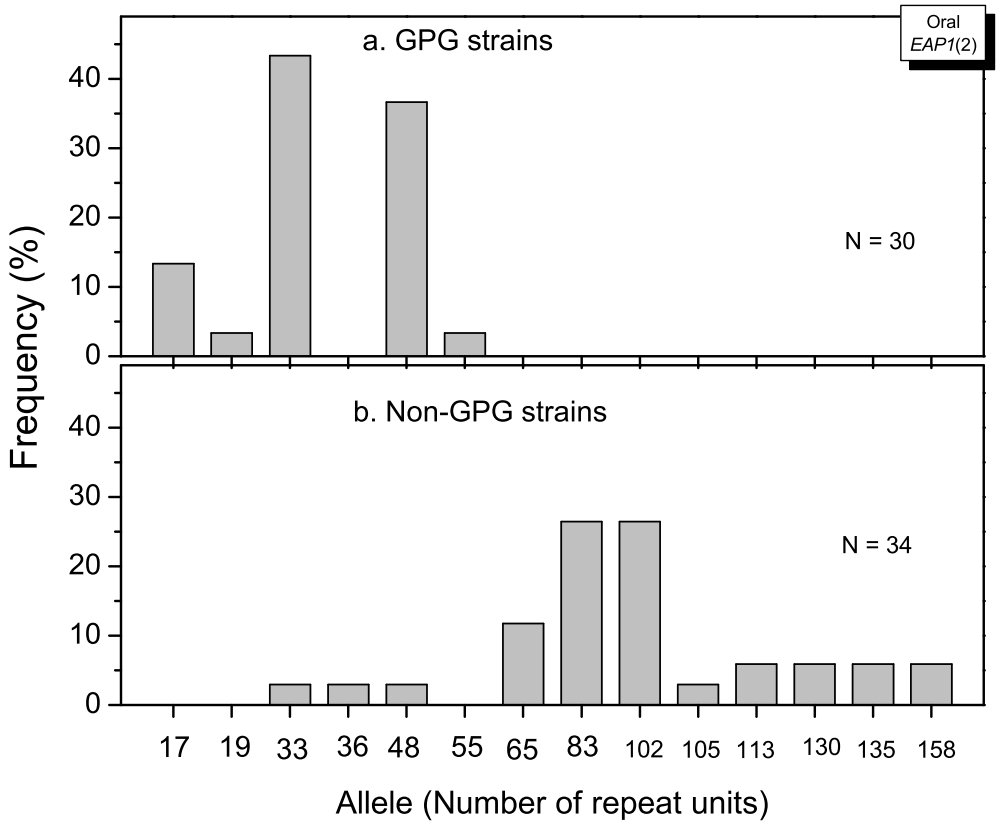

Figure 5.16: Distribution of alleles of *EAP1* for repeat region 2 for: a. GPG strains, and b. Non-GPG strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in each group, as there are 2 alleles per strain. Alleles containing 33 and 48 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). In addition, alleles containing 83 and 102 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these alleles between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

Table 5.9: Frequency of allele combinations of *EAP1* for repeat region 2 in GPG and non-GPG strains. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

No	Allele Combination		Frequency		
	Length (bp)	No. of repeats	GPG Strains	Non-GPG Strains	Total
1	400-1000	13-47	0	1	1
2	468-468	17-17	6	1	7
3	495-753	19-33	1	0	1
4	628-628	26-26	0	1	1
5	753-753	33-33	3	0	3
6	753-1024	33-48	34	2	36
7	753-1158	33-55	1	0	1
8	800-800	35-35	0	2	2
9	808-2058	36-105	0	1	1
10	929-929	43-43	0	1	1
11	1000-1000	47-47	0	1	1
12	1100-1100	52-52	0	1	1
13	1300-1300	63-63	0	1	1
14	1311-1311	64-64	1	0	1
15	1325-1325	65-65	0	2	2
16	1400-2000	69-102	0	1	1
17	1650-1650	83-83	1	9	10
18	1650-2000	83-102	0	3	3
19	2000-2000	102-102	0	10	10
20	2000-3000	102-158	0	1	1
21	2100-2100	108-108	0	1	1
22	2200-2200	113-113	1	1	2
23	2500-2500	130-130	0	2	2
24	2600-2600	135-135	0	1	1
25	3000-3000	158-158	0	2	2
26	4072-4072	228-228	0	1	1
Total			48	46	94

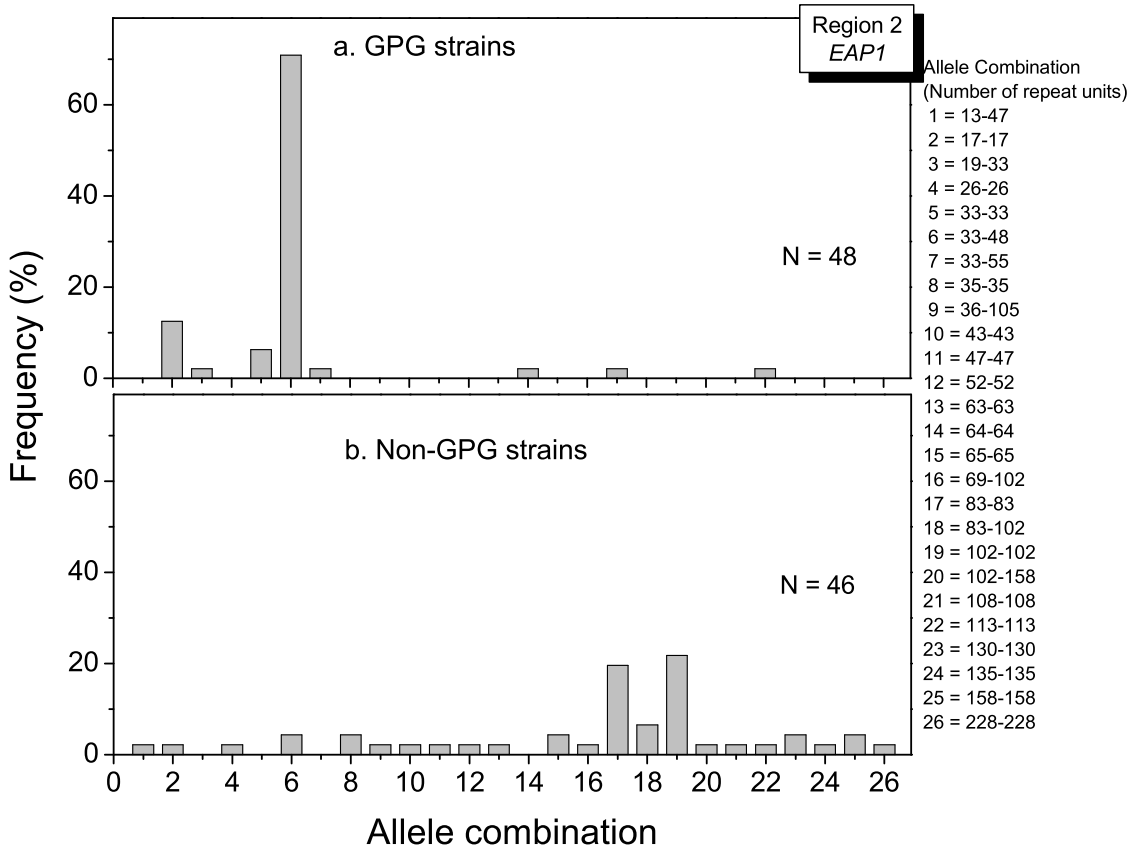

Figure 5.17: Distribution of allele combinations *EAP1* of repeat region 2 for: a. GPG strains, and b. Non-GPG strains. N represents the number of strains in each group. Due to the limitation of space in the x axis, allele combinations were written as numbers from 1 to 26 and the legend on the right shows the associated allele combinations. For convenience, allele combinations of x and y are written as x-y, where x and y are number of repeat units. Allele combinations containing 33 and 48 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). In addition, allele combinations containing 83 and 83, and 102 and 102 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

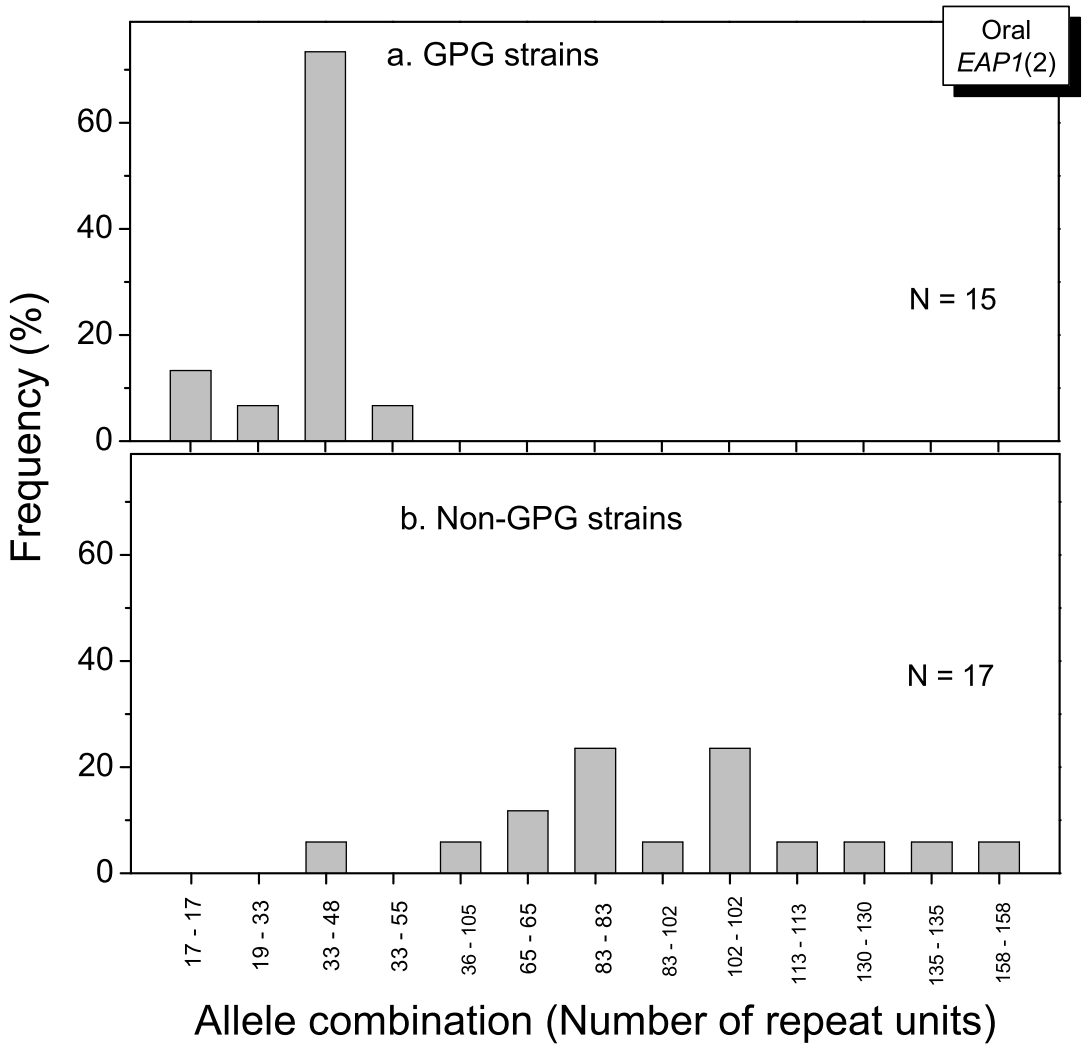

Figure 5.18: Distribution of allele combinations of *EAP1* for repeat region 2 for: a. GPG strains, and b. Non-GPG strains, where all strains were isolated from oral sites. N represents the number of strains in each group. For convenience, allele combinations of x and y are written as x-y, where x and y are number of repeat units. Allele combinations containing 33 and 48 repeat units predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). In addition, allele combinations containing 83 and 83, and 102 and 102 repeat units predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There is a significant difference in the distribution of these allele combinations between GPG and non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$).

Table 5.10: Frequency of alleles of *EAP1* for repeat region 2 in commensal strains. Alleles are represented by the fragment length and the number of repeat units. For comparison, the data from infection strains have been added.

	Allele		Frequency		
No	Length (bp)	No. of repeats	Commensal Strains	Infection Strains	Total
1	468	17	0	12	12
2	495	19	0	1	1
3	753	33	28	42	70
4	800	35	2	0	2
5	843	38	2	0	2
6	924	43	1	0	1
7	1024	48	27	34	61
8	1158	55	0	1	1
9	1311	64	0	2	2
10	1650	83	0	2	2
11	2200	113	2	2	4
Total			62	96	144

Table 5.11: Frequency of allele combinations of *EAP1* for repeat region 2 in commensal strains. For comparison, the data from infection strains have been added. For convenience, allele combinations of x and y are written as x-y, where x and y are the fragment length in bp or the number of repeat units.

	Allele Combination		Frequency		
No	Length (bp)	No. of repeats	Commensal Strains	Infection Strains	Total
1	468-468	17-17	0	6	6
2	495-753	19-33	0	1	1
3	753-753	33-33	3	3	6
4	753-1024	33-48	22	34	56
5	753-1158	33-55	0	1	1
6	800-800	35-35	1	0	1
7	843-843	38-38	1	0	1
8	929-1024	43-48	1	0	1
9	1024-1024	48-48	2	0	4
10	1311-1311	64-64	0	1	1
11	1650-1650	83-83	0	1	1
12	2200-2200	113-113	1	1	2
Total			31	48	80

5.3.2 Allelic Characterization of the *EAP1* Gene For Repeat Region 2 in Commensal Strains

Based on both allele length and the number of repeat units, six different alleles were found in commensal strains, compared to eight in infection strains. The frequency of alleles of *EAP1* for repeat region 2 found in commensal strains is summarized in Table 5.10, and presented graphically in Figure 5.19. For comparison, the data from infection strains have been added in the table and the figure.

Figure 5.19 shows that alleles containing 33 and 48 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square for contingency table, $p > 0.05$). Similar results were also observed in strains isolated from oral sites as shown in Figure 5.20. These results show that commensal and infection strains have the same predominant alleles.

Based on both allele length and the number of repeat units, seven different allele combinations were found in commensal strains, compared to eight combinations in infection strains. The frequency of allele combinations of the *EAP1* gene for repeat region 2 found in commensal strains is summarized in Table 5.11, and presented graphically in Figure 5.21. For comparison, the data from infection strains have been added in the table and the figure.

Figure 5.21 shows that allele combinations containing 33 and 48 repeat units predominated both in commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains (the Chi-square for contingency table, $p > 0.05$). Similar results were also observed in strains isolated from oral sites as shown in Figure 5.22. These results show that commensal and infection strains have the same predominant allele combinations.

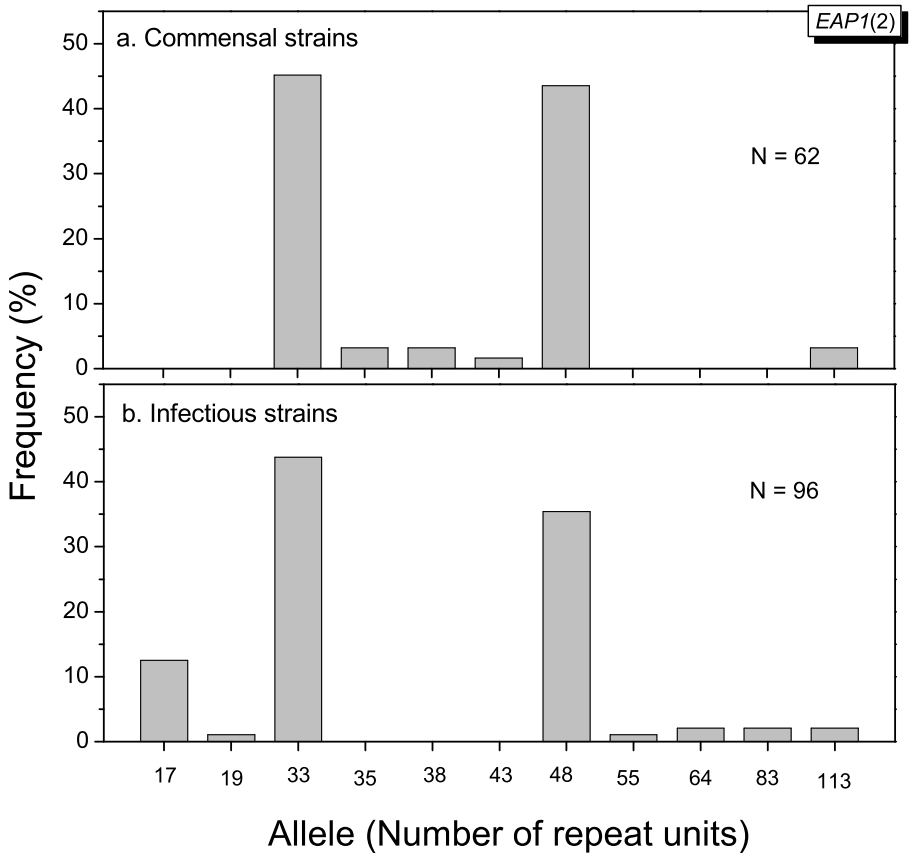

Figure 5.19: Distribution of alleles of the *EAP1* gene for repeat region 2 for: a. Commensal strains, and b. Infection strains. N represents 2 times the number of strains in each group, as there are 2 alleles per strain. Alleles containing 33 and 48 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

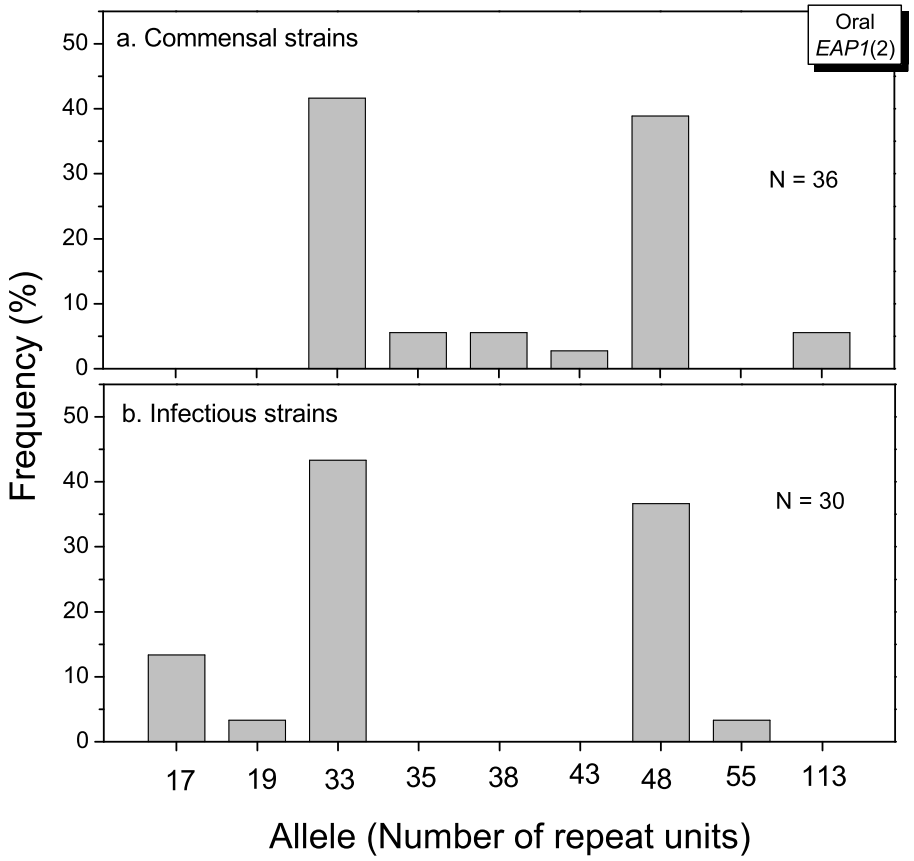

Figure 5.20: Distribution of alleles of *EAP1* for repeat region 2 by the number of repeat units for: a. Commensal strains, and b. Infection strains, where all strains were isolated from oral sites. N represents 2 times the number of strains in each group, as there are 2 alleles per strain. Alleles containing 33 and 48 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these alleles between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.05$).

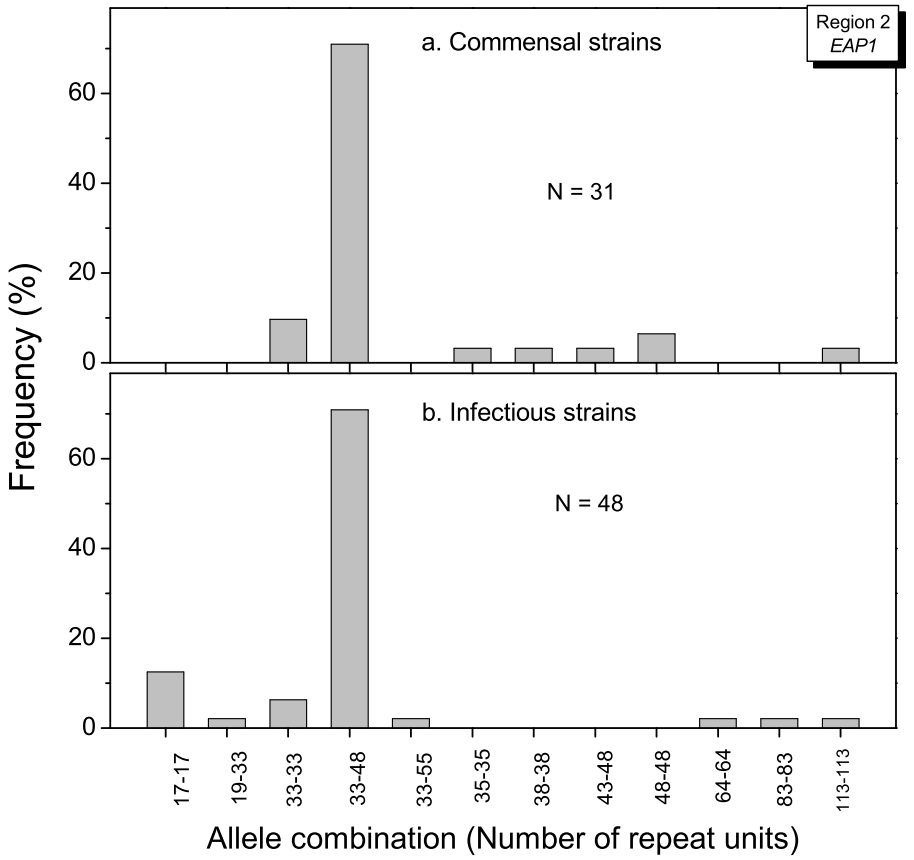

Figure 5.21: Distribution of allele combinations of *EAP1* for repeat region 2 for: a. Commensal strains, and b. Infection strains. N represents the number of strains in each group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 33 and 48 repeat units predominated in both commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains (the Chi-square for contingency tables, $p > 0.05$).

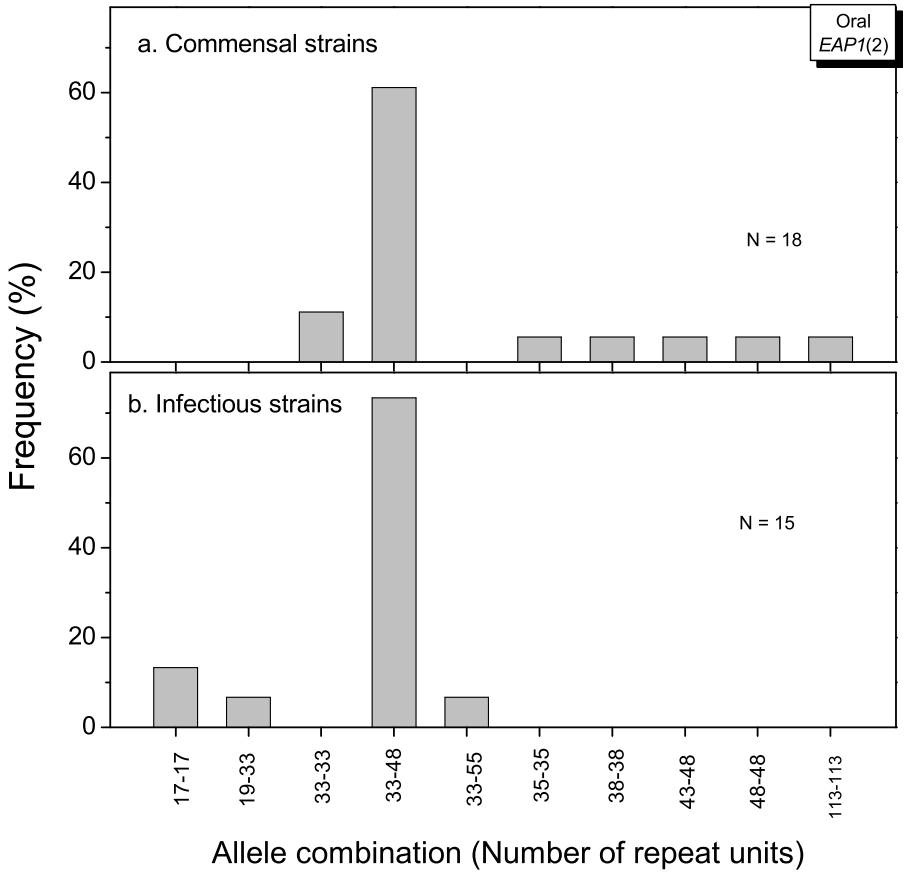

Figure 5.22: Distribution of allele combinations of *EAP1* for repeat region 2 for: a. Commensal strains, and b. Infection strains, where all strains were isolated from oral sites. N represents the number of strains in each group. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 33 and 48 repeat units predominated both in commensal and infection strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these allele combinations between commensal and infection strains (the Chi-square for contingency tables, $p > 0.05$).

5.3.3 Diversity of Alleles of the *EAP1* Gene For Repeat Region 2

Allele diversity is quantified by the index of diversity λ which is defined by Equation 3.1 (see page 48). The λ values of both the *EAP1* alleles and allele combinations of GPG, non-GPG, and commensal strains are summarized in Table 5.12, and presented graphically in Figure 5.23.

The results show that the diversity of alleles and allele combinations is influenced by the genetic diversity of the strains: GPG strains with less genetic diversity have smaller allele diversity than non-GPG strains with large genetic diversity. In contrast, infection and commensal strains have similar diversity i.e. allele diversity is independent of whether the strains exhibit a commensal or an infection phenotype.

Table 5.12: Index of diversity λ of alleles and allele combinations of *EAP1* for repeat region 2 in GPG, non-GPG, and commensal strains.

Strains	Index of Diversity	
	Allele	Allele Combination
GPG	0.67	0.49
Non-GPG	0.87	0.96
Commensal	0.61	0.49

5.3.4 Analysis of the Combination of the Two Alleles of the *EAP1* Gene of For Repeat Region 2 in An Individual Strain

Similar to the description for *YWP1* and *HWP1* in the previous chapters, the section starts with a description of the distribution of the differences between the two alleles in an individual strain, and ends with a calculation of the mean of the differences between the two alleles in an individual strain among strains of interest.

Figure 5.24 shows the distribution of differences in the number of repeat units between the two alleles of the *EAP1* gene for repeat region 2 in an individual strain

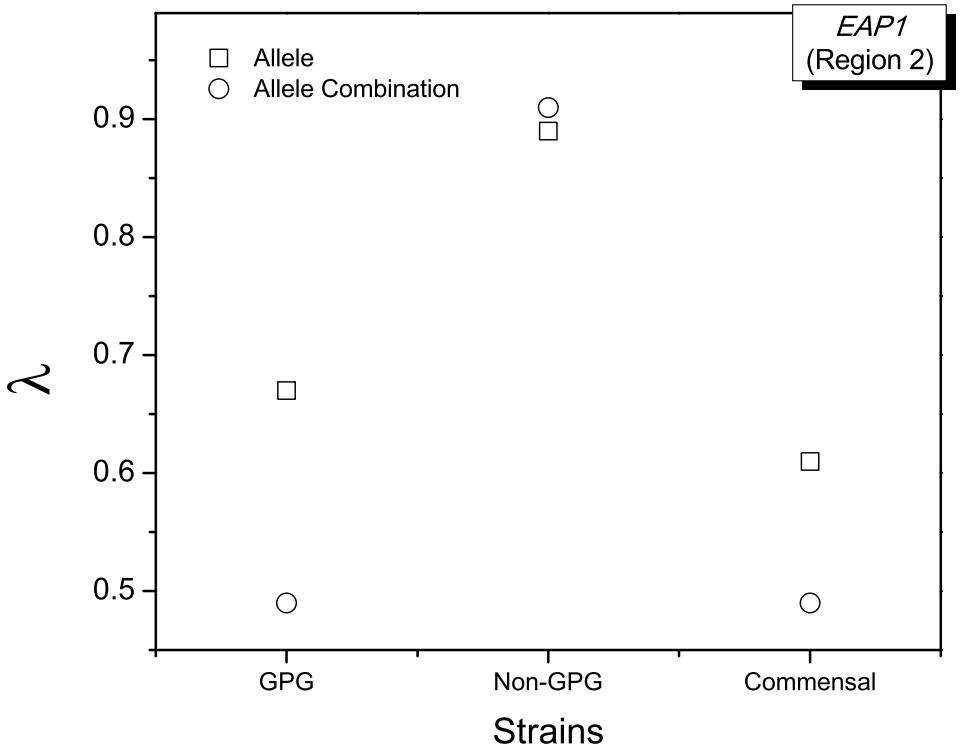

Figure 5.23: Index of diversity λ of the alleles and the allele combinations of *EAP1* for repeat region 2 in GPG, non-GPG, and commensal strains.

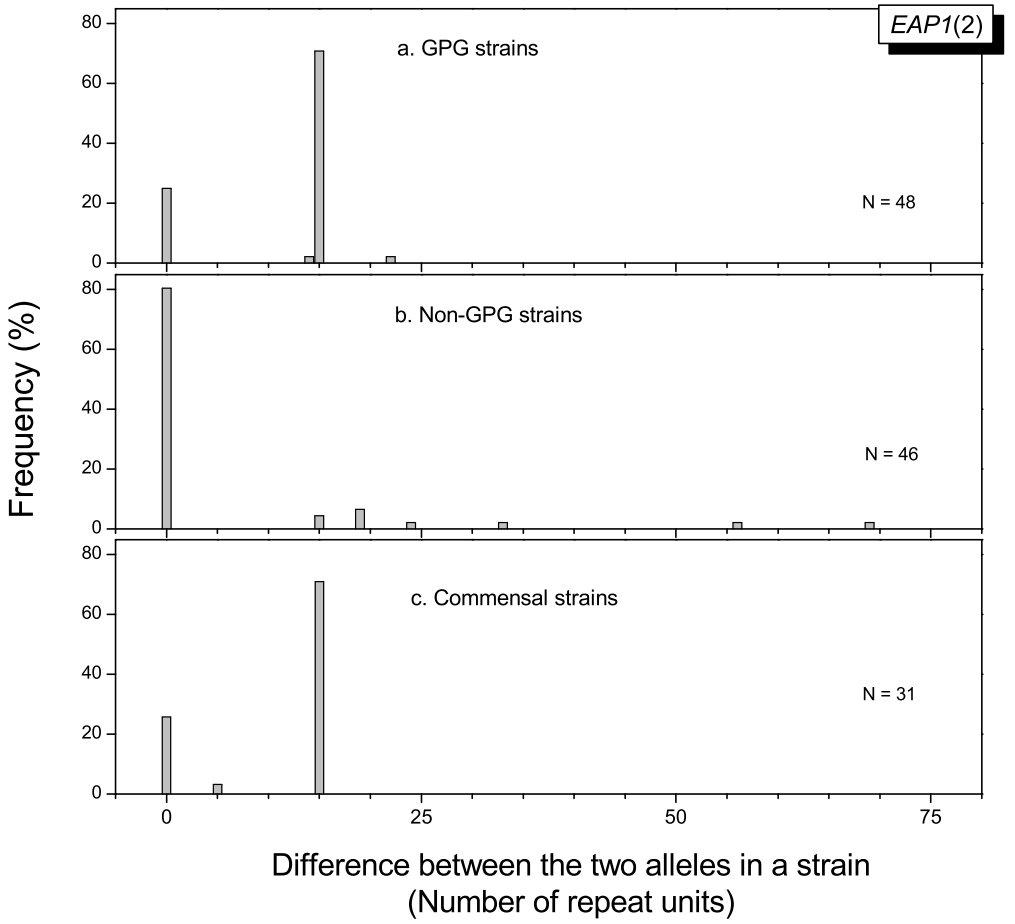

Figure 5.24: Distribution of the differences in the number of repeat units between the two alleles of *EAP1* for repeat region 2 in an individual strain for: a. GPG strains, b. Non-GPG strains, and c. Commensal strains. N represents the number of strains in each group.

for GPG, non-GPG, and commensal strains. The percentages of homozygous strains found in GPG, non-GPG, and commensal strains are 25%, 80%, and 26%, respectively.

The number of homozygous alleles in each group was compared to those expected if the combination of the alleles was random, taking allele frequency into consideration. For this purpose, a matrix with a size equal to the square of 2 times the number of strains analysed was generated, i.e. the same procedure used for the analysis for the *YWP1* gene in the previous chapter (see section 3.3.4, page 50). The frequency of homozygotes and heterozygotes of *EAP1* (repeat region 1) for the data observed and those expected from random alleles is summarized in Table 5.13, and is presented graphically in Figure 5.25. The figure shows no significant difference between the number of homozygotes observed and those expected by chance for (GPG) commensal (the Chi-square test for contingency tables , p>0.1) and infection strains (p>0.05). However, for non-GPG strains, the number of homozygotes is significantly larger than those expected by chance (the Chi-square test for contingency tables, p<0.001), i.e. *EAP1*(1) exhibits significant excesses of homozygotes.

Table 5.13: The number of homozygous strains in GPG, non-GPG, and commensal strains compared to the expected number by chance. N indicates the number of strains.

Strains	Number of homozygous strains		Significance (p)
	Observed (N)	Randomly expected	
GPG	12 (48)	16	$p > 0.1$
Non-GPG	37 (47)	7	$p < 0.001$
Commensal	8 (31)	12	$p > 0.05$

The mean (± standard deviation) of the differences in the number of repeat units between the two alleles in an individual strain was determined to be 11.4 ± 6.7, 6.1 ± 14.9, and 10.8 ± 6.7, for GPG, non-GPG, and commensal strains, respectively. The result of the t-test between the two means of GPG and non-GPG strains showed that alleles in a GPG strain are significantly more different to each other than alleles in a non-GPG strain ($p < 0.01$). The result of the t-test between the two means of (GPG) commensal and infection strains showed that alleles in a commensal strains

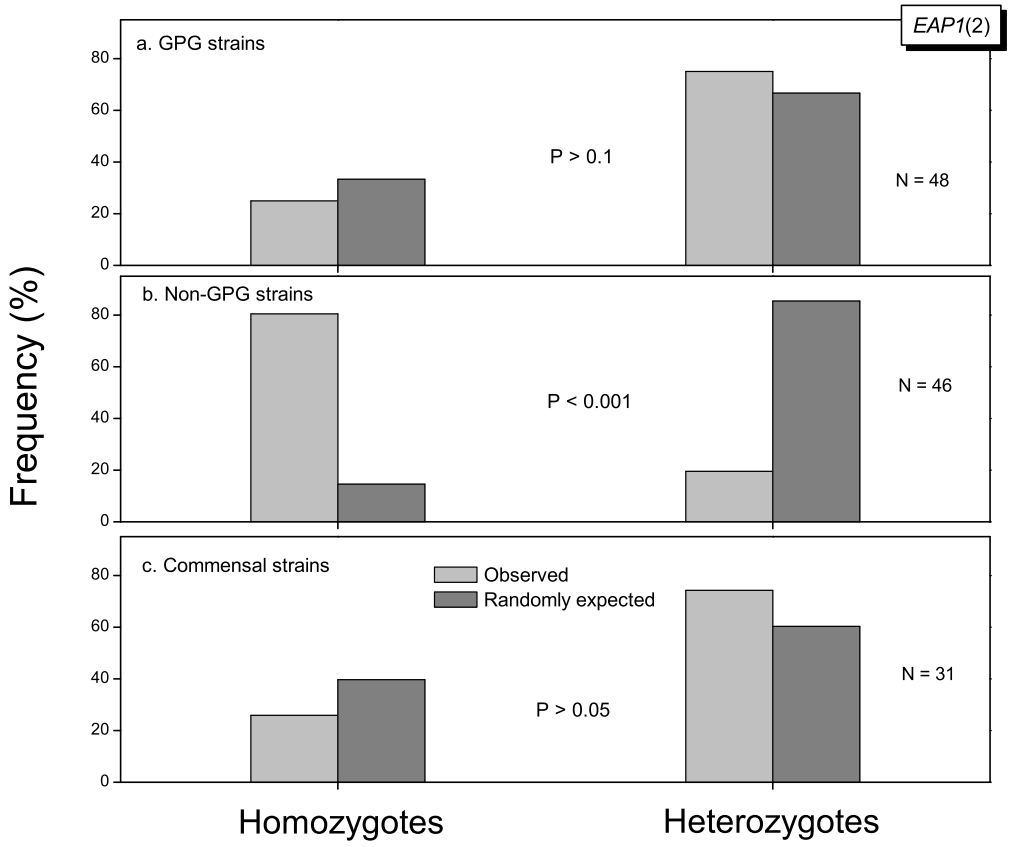

Figure 5.25: Frequency of homozygotes and heterozygotes of *EAP1* for repeat region 2 for GPG, non-GPG, and commensal strains. N represents the number of strains in a particular group. There is no significant difference between the number of homozygotes observed and those expected by chance for (GPG) commensal (the Chi-square test for contingency tables , $p>0.1$) and infection strains ($p>0.05$). However, for non-GPG strains, the number of homozygotes is significantly larger than those expected by chance (the Chi-square test for contingency tables, $p<0.001$), i.e. *EAP1*(2) exhibits significant excesses of homozygotes in non-GPG strains.

did not differ significantly from alleles in an infection strain ($p > 0.1$).

To determine whether the difference between the two alleles in an individual strain is random or non-random, the means of the differences of the alleles found were compared to the means expected if all alleles combined randomly, taking allele frequency into consideration. As explained previously, a matrix with the size equal to square of 2 times the number of strains analysed was generated for this purpose.

For GPG strains, the mean ($\pm$ standard deviation) of the differences in the number of repeat units between the two alleles expected from random alleles was determined to be 15.8 ± 17.9. The result of the t-test between this mean and the mean from observed data 11.4 ± 6.7 showed that alleles in a GPG strain are significantly more similar to each other than those expected by chance ($p < 0.05$). For non-GPG strains, the mean ($\pm$ standard deviation) of the differences in the number of repeat units between the two alleles expected from random alleles was determined to be 42.2 ± 37.0. The result of the t-test between this mean and the mean from observed data 6.1 ± 14.9 showed that alleles in a non-GPG strain are significantly more similar to each other than those expected by chance ($p < 0.005$). For commensal strains, the mean ($\pm$ standard deviation) of the differences in the number of repeat units between the two alleles expected from random alleles was determined to be 11.5 ± 17.3. The result of the t-test between this mean and the mean from observed data 10.8 ± 6.7 showed that alleles in a commensal strain did not differ significantly from those expected by chance ($p > 0.10$). The mean ($\pm$ standard deviation) of the differences in the number of repeat units between the two alleles from observed data, data expected from random alleles, and the significance of the differences between the means are summarized in Table 5.14.

Table 5.14: The mean (± standard deviation) of the differences in the number of repeat units between the two alleles of *EAP1* for repeat region 2 in an individual strain from observed data, data expected from random alleles, and the significance of the difference between the two means for GPG, non-GPG, and commensal strains. N indicates the number of data.

Strains	Mean Differences (± SD)		Significance (p)
	Observed Differences (N)	Expected Differences (N)	
GPG	11.4 ± 6.7 (48)	15.8 ± 17.5 (9216)	$p < 0.05$
Non-GPG	6.1 ± 14.9 (46)	42.2 ± 37.0 (8464)	$p < 0.005$
Commensal	10.8 ± 6.7 (31)	11.5 ± 17.3 (3844)	$p > 0.10$

5.3.5 Allelic Distribution of the *EAP1* Gene For Repeat Region 2 in Strains Isolated From Different Sites of the Humans Body

Figure 5.26 shows the distribution of alleles of *EAP1* for repeat region 2 in strains isolated from different sites of the humans body. Alleles containing 33 and 48 repeat units predominated at oral and vagina sites (the Chi-square goodness of fit test, $p < 0.001$), while at skin, urine, and sterile sites, no obvious pattern of any predominant alleles was observed. The result of the Chi-square test for contingency tables showed no significant difference in the distribution of these alleles between any two sites for all sites sampled ($p > 0.05$).

Figure 5.27 shows the distribution of allele combinations of *EAP1* for repeat region 2 in strains isolated from different sites of the humans body. The distributions of allele combinations at all sites appear similar; allele combination containing 33 and 48 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). The result of the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between any two sites for all sites sampled ($p > 0.05$).

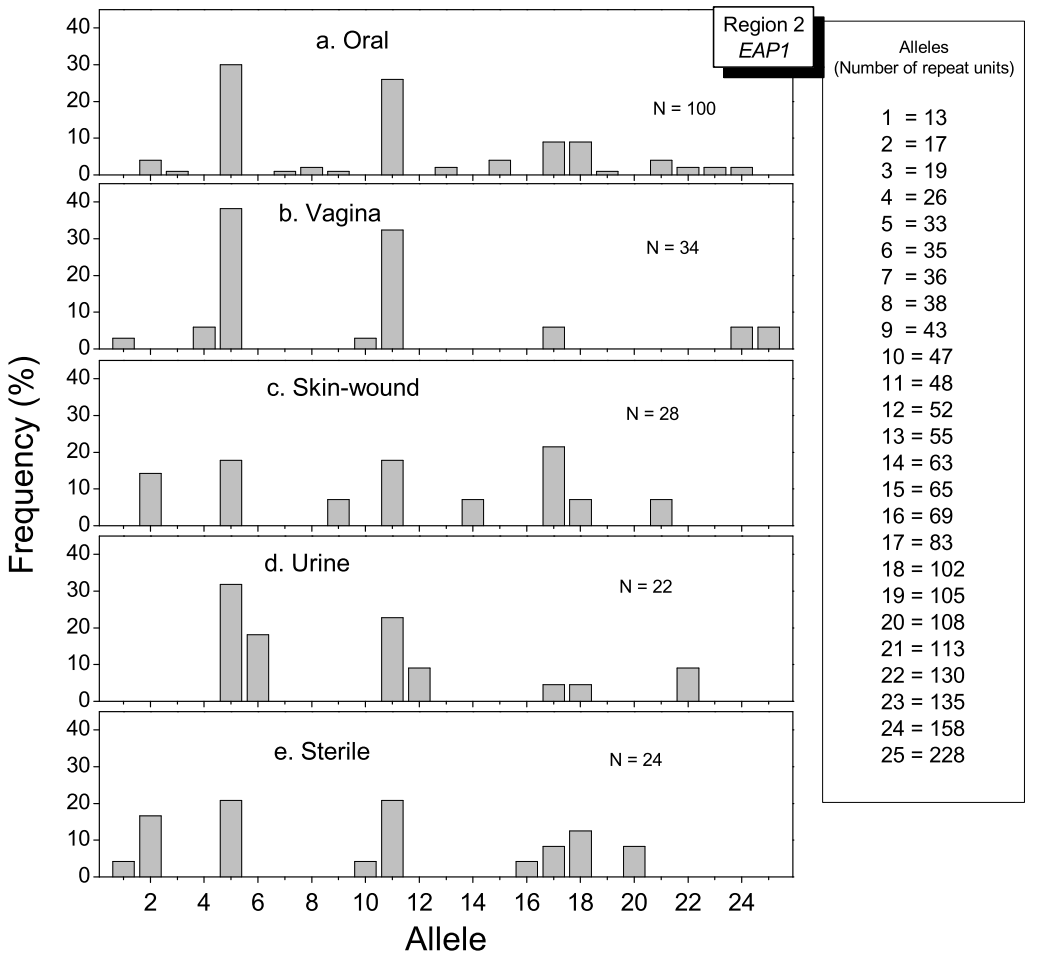

Figure 5.26: Distribution of alleles of *EAP1* for repeat region 2 in strains isolated from different sites of the humans body. N represents 2 times the number of strains at each site, as there are 2 alleles per strain. Due to the limitation of space in the x axis, alleles were written as numbers from 1 to 25, and the legend on the right represents the associated alleles. Alleles containing 33 and 48 repeat units predominated at oral and vagina sites (the Chi-square goodness of fit test, $p < 0.001$). At skin, urine, and sterile sites, no obvious pattern of any predominant alleles was observed. The result of the Chi-square test for contingency tables showed no significant difference in the distribution of these alleles between any two sites for all sites sampled ($p > 0.05$).

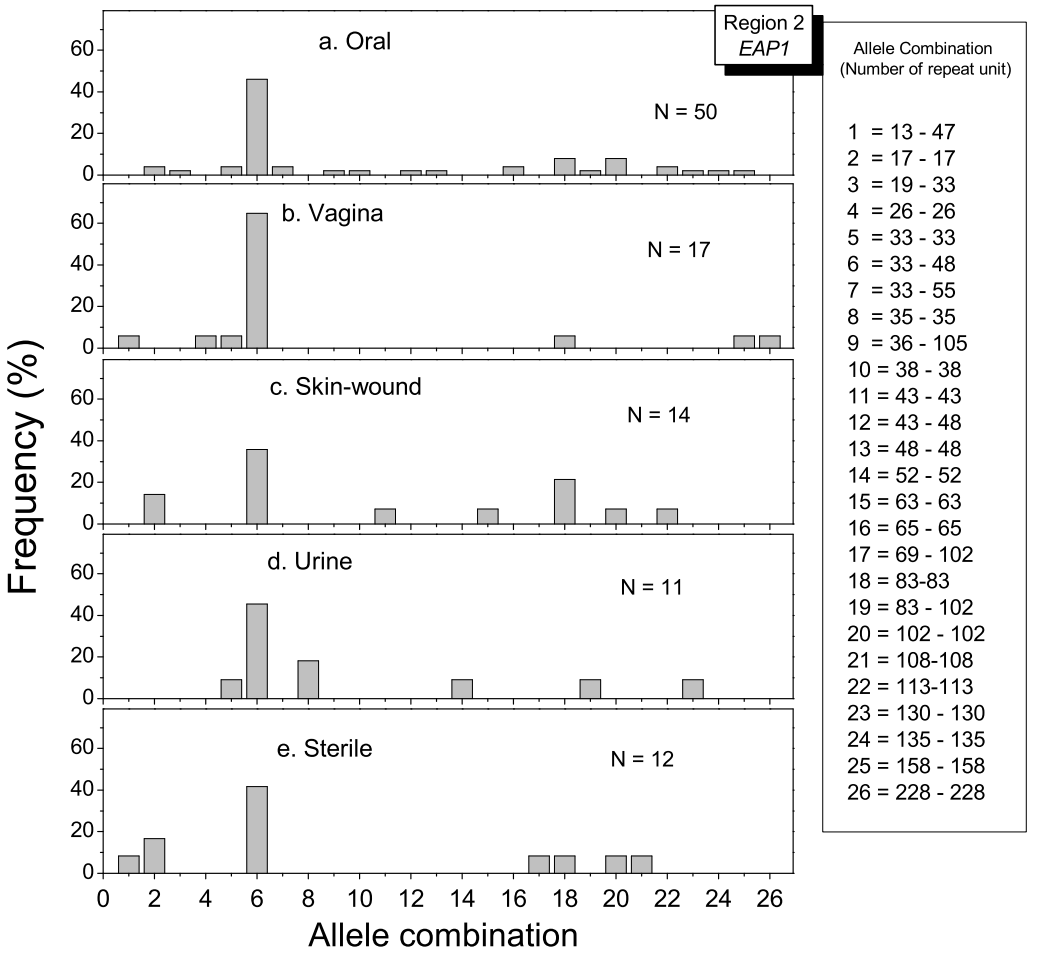

Figure 5.27: Distribution of allele combinations of *EAP1* for repeat region 2 in strains isolated from different sites of the humans body. N represents the number of strains at each site. Due to the limitation of space in the x axis, allele combinations were written as numbers from 1 to 26, and the legend on the right represents the associated allele combinations. For convenience, allele combinations of x and y are written as x-y, where x and y are the number of repeat units. Allele combinations containing 33 and 48 repeat units predominated at all sites (the Chi-square goodness of fit test, $p < 0.001$). The result of the Chi-square test for contingency tables showed no significant difference in the distribution of these allele combinations between any two sites for all sites sampled ($p > 0.05$).

5.4 Allele Combinations of the *EAP1* Gene For Pairs of Repeat Regions 1 and 2

In the previous sections, the results of allelic characterization of *EAP1* for repeat regions 1 and 2 are described separately. In this section, the repeat regions 1 and 2 are described as one representation of the *EAP1* gene. For convenience, allele combinations of *EAP1* for pairs of repeat regions 1 and 2 are referred to as genotypes of *EAP1*.

The genotypes of *EAP1* are described in a three dimensional graph, where each of the two horizontal axes represents allele combinations of each repeat region, and the vertical axis represents the frequency of the genotypes. For convenience, the genotypes are written as "p-q + r-s", where "p-q" and "r-s" associate with repeat regions 1 and 2, respectively, and p, q, r, s are the number of repeat units.

5.4.1 Genotypes of the *EAP1* Gene in GPG and Non-GPG Strains

Figures 5.28 and 5.29 show the distributions of genotypes of *EAP1* for GPG and non-GPG strains, respectively. Figure 5.28 indicates that there are 13 different genotypes in GPG strains, while Figure 5.29 indicates that there are 42 different genotypes in non-GPG strains. In total, there are 49 different genotypes found in infection strains.

Figure 5.28 shows that genotypes containing 33 and 48 repeats of repeat region 1, and 10 and 11 repeats of repeat region 2 predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains, compared to non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). Figure 5.29 shows no obvious pattern of any predominant genotypes in non-GPG strains.

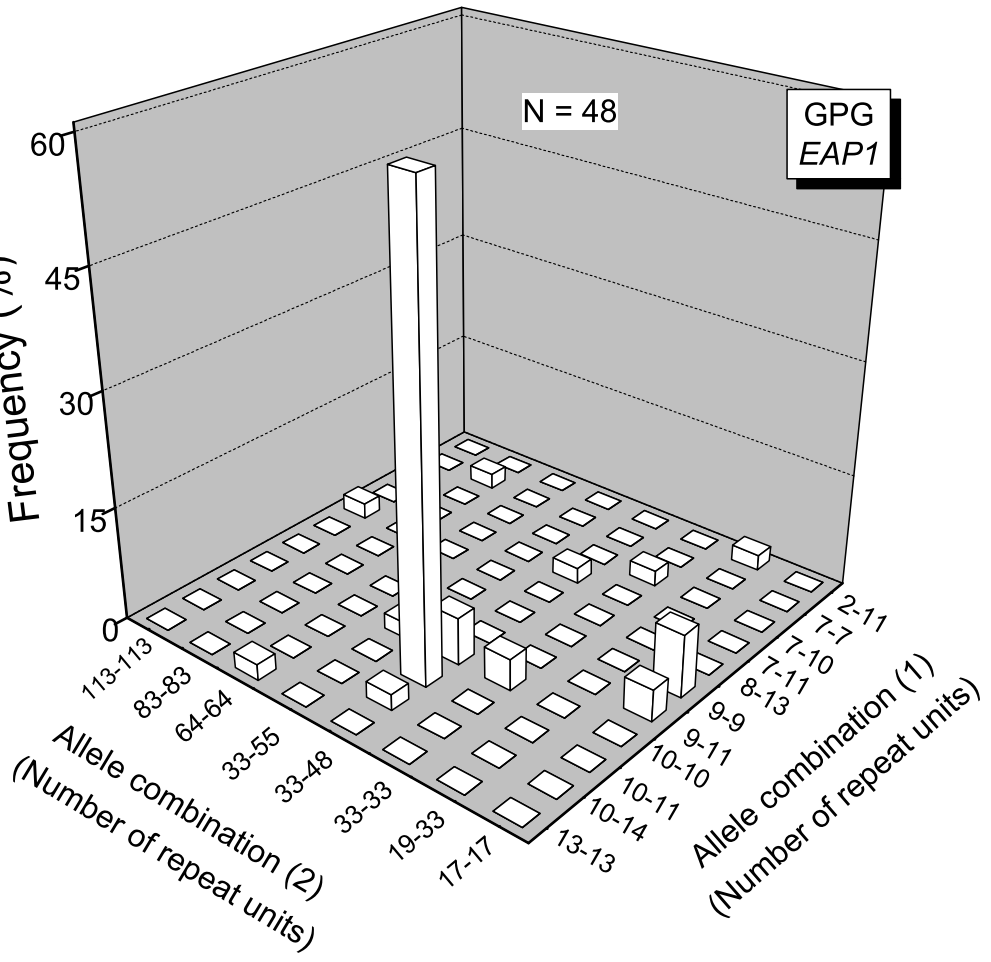

Figure 5.28: Distribution of genotypes of *EAP1* for GPG strains. N represents the number of strains characterized. There are 13 different genotypes reflected by the number of bars in the graph. Genotypes containing 33 and 48 repeat units of repeat region 1, and 10 and 11 repeat units of repeat region 2 predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains, compared to non-GPG strains shown in Figure 5.29 (the Chi-square test for contingency tables, $p < 0.001$).

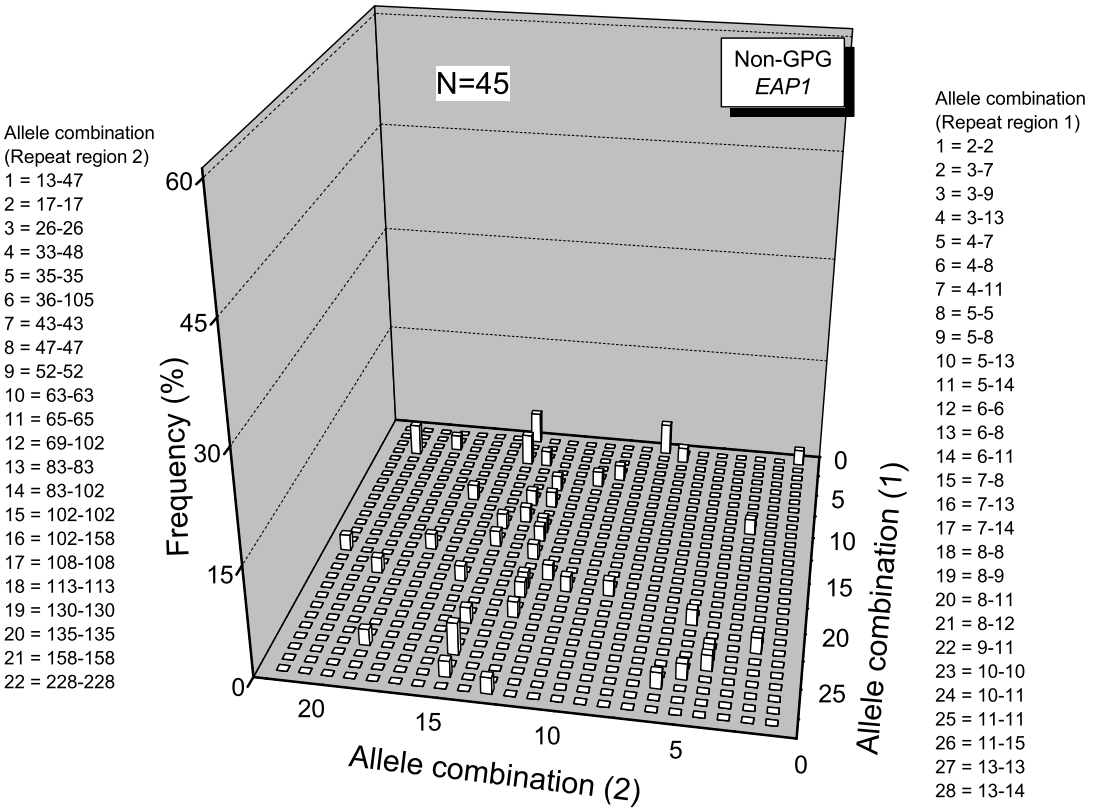

Figure 5.29: Distribution of genotypes of *EAP1* for non-GPG strains. N represents the number of strains characterized. Due to the limitation of space in the horizontal axes, allele combinations are written as numbers from 1 to 28 and 1 to 22 for repeat regions 1 and 2, respectively. The legends show the associated allele combinations. There are 42 different genotypes reflected by the number of bars in the graph. There is no obvious pattern of any predominant genotypes in non-GPG strains.

5.4.2 Genotypes of the *EAP1* Gene in Commensal Strains

Figure 5.30 shows the distribution of 11 genotypes of the *EAP1* gene identified in commensal strains. The figure shows that genotypes containing 33 and 48 repeat units of repeat region 1 and 10 and 11 repeat units of repeat region 2 predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$). The results of the Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal strains and infection strains shown in Figure 5.28 ($p > 0.05$).

5.5 Discussion

The aim of this particular study of the *EAP1* gene was to find out whether *EAP1* is a contingency gene, i.e. whether *EAP1* has a role in adaptation by changing the number of repeat units within the coding sequences. For this purpose, allelic distribution of *EAP1* in GPG and non-GPG strains, in commensal and infection strains, and in strains isolated from different sites of the humans body was examined and compared one with another.

5.5.1 Variability in the number of *EAP1* repeats

This study reveals that there is a variability in the number of *EAP1* repeats observed from different strains. From 96 infection strains, 15 and 25 different alleles of *EAP1* were detected for repeat regions 1 and 2, respectively, which formed 13 and 26 different allele combinations. Likewise, from 29 commensal strains, seven and six different alleles of *EAP1* were detected for repeat regions 1 and 2, respectively, which formed eight and seven different allele combinations. From total 125 strains, there were 15 different alleles detected, which form 38 different different allele combinations of *EAP1* for repeat region 1, and 26 different alleles detected, which form 29 different different allele combinations of *EAP1* for repeat region 2. The number of different *EAP1* alleles observed is more than that observed for *YWP1* and *HWP1* (eight and

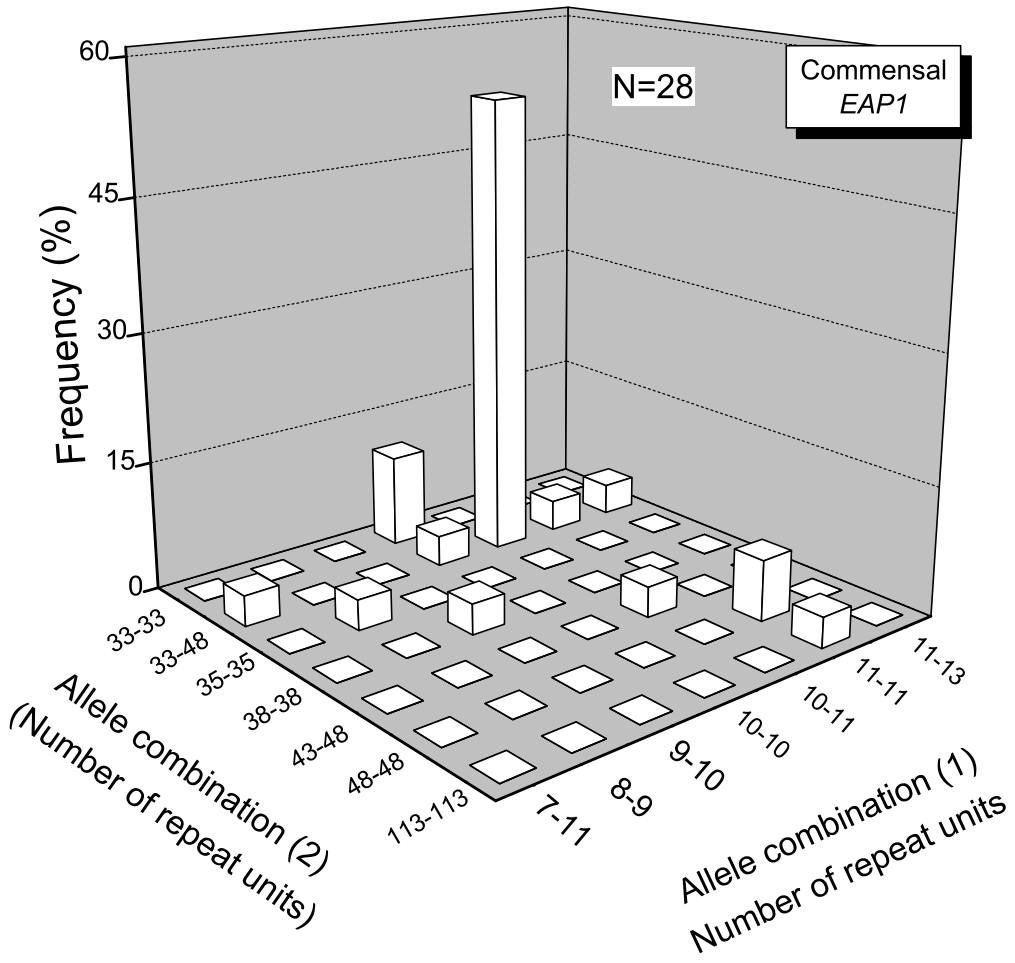

Figure 5.30: Distribution of genotypes of *EAP1* for commensal strains. N represents the number of strains characterized. There are 11 different genotypes reflected by the number of bars in the graph. Genotypes containing 33 and 48 repeat units of repeat region 1, and 10 and 11 repeat units of repeat region 2 predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these genotypes between commensal strains and infection strains of Figure 5.28 (the Chi-square test for contingency tables, $p > 0.05$).

five different alleles). The higher number of different *EAP1* alleles observed compared to *YWP1* and *HWP1* may indicate that *EAP1* has a relatively high mutation rate. The mutation rate of *EAP1* will be discussed in Chapter 7, while the indication that *EAP1* may have no role in adaptation will be discussed in the following section.

It has been described in Chapter 3 that the variability in the number of repeat units appears to relate to the the ability of the *C.albicans* cells to adhere to the host [27, 28, 29]. In relation to that, it has been suggested that EAP1p play a role in adhesion. For example, deleting *EAP1* results in 30% reduction of *C. albicans* adhesion ability in a humans cell line, and expression of *EAP1* in *S. cerevisiae* EAP1p increased the adherent ability of this fungus to HEK293 kidney epithelial cells [65]. In addition, it has been demonstrated that EAP1p has three domains, which relate to the adherent ability of the *C.albicans* cell: i) N-terminal ligand-binding domain, ii) N-terminal repeat domain (repeat region 2 of this study), and iii) C-terminal repeat domain (repeat region 1 of this study) [66]. The two Eap1p repeats appear to play a direct role in adhesion of *C. albicans*. N-terminal repeat domain (repeat region 2 of this study) and C-terminal repeat domain (repeat region 1 of this study) are required for adhesion to polystyrene while only N-terminal repeat domain is required to mediate adhesion to mammalian epithelial cells [66].

5.5.2 GPG and non-GPG strains have distinct predominant alleles and allele combinations of *EAP1*

The results of allelic characterization of the *EAP1* gene for repeat region 1 showed that GPG and non-GPG strains have distinct predominant alleles (see Figure 5.2 on page 105), but predominant allele combinations of *EAP1* were only observed in GPG strains (see Figure 5.4 on page 109). Alleles containing 10 and 11 repeat units predominated in GPG strains, while alleles containing 8 and 11 repeat units predominated in non-GPG strains. Allele combinations containing 10 and 11 repeat units predominated in GPG strains, but no obvious predominant allele combinations were observed in non-GPG strains. The results for allele combinations are in agreement

with the results for other genes of *C.albicans* such as *ALS7* [32], *PNG2* [51], and *SSR1* [52].

The results of allelic characterization of the *EAP1* gene for repeat region 2 showed that GPG and non-GPG strains have distinct predominant alleles and allele combinations (see Figure 5.15 on page 129 and Figure 5.17 on page 132). Allele combinations of *EAP1* containing 33 and 48 repeat units predominated in GPG strains, and overrepresented in GPG strains compared to non-GPG strains. In addition, allele combinations containing 83 and 83 repeat units, and 102 and 102 repeat units predominated in non-GPG strains, and overrepresented in non-GPG strains compared to GPG strains. These results showed that there are specific allele combinations of *EAP1* for repeat region 2 associated with each group of infection strains. However, the percentage of the predominant allele combinations in GPG strains is larger than that in non-GPG strains.

One of the differences in the distribution of allele combinations of *EAP1* for repeat regions 1 and 2 is that there are predominant allele combinations for repeat region 2 in non-GPG strains, but no obvious predominant allele combinations for repeat region 1 were observed in non-GPG strains. The percentage of the predominant allele combination of *EAP1* for repeat region 2 is larger in GPG strains compared to non-GPG strains. These results are similar to the results for the *YWP1* gene, which support the argument that : i) the distinct predominant allele combinations in GPG and non-GPG strains are related to the genetic backgrounds, and ii) the percentages of the predominant allele combinations were influenced by the genetic diversity among the strains in each group. The implications of the fact that some non-GPG strains have predominant allele combinations (*YWP1* and *EAP1* repeat region 2), while others show no obvious pattern (*EAP1* repeat region 1, *ALS7* [32], *PNG2* [51], and *SSR1* [52]) are still required more explanation. In Chapter 6, the possibility that the predominant allele combinations may relate to the gene interconnectedness will be discussed, and in Chapter 7, the possibility of the impact of the mutation rate on the predominant allele combinations will be discussed.

The allelic distributions of *EAP1* for both repeat regions 1 and 2 in GPG and non-GPG strains imply that *EAP1* is not involved in adaptation by changing the number of repeats units within coding sequences. Therefore, *EAP1* cannot be categorized as a contingency gene. The allelic distributions of *EAP1* for repeat regions 1 and 2 are different from the allelic distributions of *YWP1*, where GPG and non-GPG strains select the same predominant alleles beside distinct predominant alleles selected by non-GPG strains, and allelic distribution of *HWP1* where GPG and non-GPG strains select the same predominant alleles and allele combinations. If *EAP1* had a role in adaptation, GPG and non-GPG strains should select the same predominant alleles as shown by *YWP1* or the same allele combinations as shown by *HWP1*. In fact, for repeat region 1, there were no obvious predominant allele combinations of *EAP1* observed in non-GPG strains, while for repeat region 2 non-GPG strains select the predominant alleles and allele combinations different from those selected by GPG strains.

The allelic distributions of *EAP1* for both repeat regions in GPG and non-GPG strains show no indication of a selective pressure on the gene. It has been suggested that if there is a selective pressure on the gene, the influence of the genetic background is diminished [57]. Therefore, the *EAP1* allelic distributions are mostly influenced by the genetic background. This is similar to the results for *ALS7* [32], *PNG2* [51], and *SSR1* [52].

The fact that GPG and non-GPG strains select the same *YWP1* and *HWP1* predominant alleles, but different *EAP1* predominant alleles suggests that *EAP1* is not as important as *YWP1* and *HWP1* for *C. albicans* to function optimally in its host. This argument is in the sense that for *YWP1* and *HWP1*, there are particular number of repeats in the coding sequences required for *C. albicans* to function properly in its host, thus selected by both GPG and non-GPG strains, while for *EAP1*, there were no such requirements, i.e. *C. albicans* can still function with a variation in the number of repeats in the coding sequences.

However, how EAP1p mediates adhesion of *C. albicans* cell to host cell remains unclear. There are three possible interactions for that purpose: specific interaction

such as ligand-adhesin bonds, and unspecific interactions such as hydrophobic and electrostatic interactions [10, 90, 91, 92, 93, 94]. The attachment of *C. albicans* cell to the host cell through HWP1p involves ligand-adhesin bond. However, the cell surface component (ligand) that HWP1p binds is unidentified yet, and it might be through one of the three interactions mentioned above. So far, ligands for *C. albicans* adhesins have not been identified yet. However, for *C. glabrata*, Epa1p that has slight homology to EAP1p of *C. albicans* binds to N-acetyl lactosamine glycoconjugates of mammalian cell [95]. Whether N-acetyl lactosamine glycoconjugates also act as ligand in the interaction of *C. albicans* EAP1p with the host cell still needs more investigation. Whatever interaction between EAP1p and the host cell occurs, there is no need to choose specific allele combinations in order to adapt to the host.

The two repeat domains of EAP1p appear also to play no direct role in adhesion of *C. albicans* through their function in supporting the N terminal ligand-binding domain away into extracellular environment, where it can mediate yeast cell-cell adhesion [66]. The increased cell-cell adhesion has been found to increase invasive growth in *S. cerevisae* [96], and has been suggested also to cause invasive growth in *C. albicans* [66]. To do this job, EAP1p repeats do not show any requirements of adaptation.

5.5.3 The number of *EAP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic

The comparison of the distributions of allele combinations of *EAP1* for repeat region 1 in commensal and infection strains showed that no significant difference in the distribution of the predominant alleles and allele combinations of *EAP1* in commensal and infection strains (see Figure 5.6 on page 112 and Figure 5.8 on page 115). The results for repeat region 2 are similar to the results for repeat region 1 (see Figure 5.19 on page 136 and Figure 5.21 on page 138). In addition, the percentages of the predominant alleles and allele combinations are similar, and consequently the variation of allele combinations shown by the index of diversity is similar in both commensal and infection strains (see Figure 5.10 on page 117 and Figure 5.23 on page 141 for

repeat regions 1 and 2, respectively). The same predominant allele combinations of the repeat regions 1 and 2 of the *EAP1* gene of commensal and infection strains indicate no association between allele combinations with the immune status of the host. This suggests that the number of *EAP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic. This is in agreement with the results for the *YWP1* and *HWP1* genes described in the previous chapters. Like *YWP1* and *HWP1*, it seems that for *EAP1*, *C. albicans* select particular allele combinations when growing in both colonization of healthy individuals and in the invasion of immunocompromised individuals.

5.5.4 The number of *EAP1* repeats does not alter when *C. albicans* moves to particular sites of the humans body

This study showed no significant difference in the distribution of alleles and allele combinations of *EAP1* for any humans body sites sampled (see Figure 5.13 on page 124 and Figure 5.14 on page 125 for repeat region 1, and Figure 5.26 on page 147 and Figure 5.27 on page 148 for repeat region 2). The results suggest that The number of *EAP1* repeats does not alter when *C. albicans* moves to particular sites of the humans body. This is similar to the results for the *YWP1* and *HWP1* explained previously, which is in agreement with the results of studies on *ALS7* [32], *PNG2* [51], *SSR1* [52], and a study on bloodstream and non-bloodstream isolates [85].

This result shows that there is no influence of the different environmental conditions (provided by different sites of the humans body) on the *EAP1* allelic distributions. This is also supported by the evidence that there is a similarity in the *EAP1* allelic distributions in strains isolated from oral sites compared to all strains from different sites. For example, the distribution of allele combinations of *EAP1* for repeat region 1 in GPG and non-GPG strains isolated from oral sites (see Figure 5.5 on page 110) is similar to the distribution of allele combinations of *EAP1* for in GPG and non-GPG strains from all infection strains tested (see Figure 5.4 on page 109).

The same is true when comparing allele combinations of *EAP1* for repeat region 1 in commensal and infection strains (see Figures 5.9 and 5.8 on pages 116 and 115). The same results were shown by the allelic distributions of *EAP1* for repeat region 2.

5.5.5 The homozygous and heterozygous alleles of *EAP1*

The analysis of the combination of the two alleles of *EAP1* for repeat region 1 in an individual strain showed that *EAP1*(1) exhibits deficit of homozygotes in GPG strains, but the deficit of homozygotes is not significant (see Table 5.6 on page 121). This result is similar to the result for *ALS3* [33]. In contrast, the analysis showed that *EAP1*(1) exhibits a significant excesses of homozygotes in non-GPG strains, which is similar to the results for *ALS5* and *ALS6* [33]. In addition, *EAP1*(1) exhibits excesses of homozygotes in commensal strains, but the excesses of homozygotes is not significant.

The analysis of the combination of the two alleles in an individual strain showed that the two alleles of *EAP1*(1) in a GPG strain did not differ significantly from those expected by chance (see Table 5.7 on page 122). This result is similar to the result for *YWP1* in a non-GPG strain explained in chapter 3. In contrast, the analysis showed that the two alleles of *EAP1*(1) in a non-GPG strain are significantly more similar to each other than those expected by chance . This result is similar to the results for *ALS5*, *ALS6*, and *ALS7* [32, 33], and for *YWP1* in (GPG) infection and commensal strains explained in chapter 3. However, this result is in contrast to the result for *ALS3* which showed the observed difference is significantly larger than the expected difference by chance [27, 33]. In addition, the analysis also showed that the two alleles of *EAP1*(1) in a commensal strain did not differ significantly from those expected by chance, which is similar to the results for (GPG) infection strains.

The analysis of the combination of the two alleles of *EAP1* for repeat region 2 in an individual strain showed that *EAP1*(2) exhibits deficit of homozygotes in GPG strains, but the deficit of homozygotes is not significant (see Table 5.13 on page 143). This result is similar to the result for *ALS3* [33]. In contrast, the analysis showed that *EAP1*(2) exhibits a significant excesses of homozygotes in non-GPG strains, which is

similar to the results for *ALS5* and *ALS6* [33]. In addition, *EAP1*(2) exhibits deficit of homozygotes in commensal strains, but the deficit of homozygotes is not significant, similar to the result for (GPG) infection strains.

The analysis of the combination of the two alleles in an individual strain showed that the two alleles of *EAP1*(2) in a GPG strain and a non-GPG strain are significantly more similar to each other than those expected by chance (see Table 5.14 on page 146). This result is similar to the results for *ALS5*, *ALS6*, and *ALS7* [32, 33], and for *YWP1* in (GPG) infection and commensal strains explained in chapter 3. However, this result is in contrast to the result for *ALS3* which showed the observed difference is significantly larger than the expected difference by chance [27, 33]. In addition, the analysis also showed that the two alleles of *EAP1*(2) in a commensal strain did not differ significantly from those expected by chance, which is similar to the result for *YWP1* in a non-GPG strain explained in chapter 3.

Chapter 6

Interconnectedness Between Alleles of the *YWP1* and *HWP1* and *EAP1* Genes

This chapter describes the interconnectedness between alleles of *YWP1*, *HWP1*, and *EAP1*. The interconnectedness between alleles from different genes is defined as a non-random association between the alleles. A comparison of any interconnectedness in GPG and non-GPG strains may provide insights into the impact of the genetic background on the gene interconnectedness, while a comparison of any interconnectedness in infection and commensal strains may provide insights into the impact of the immune status of the host on the gene interconnectedness.

The chapter begins with a description of the allele combinations of pairs of genes in GPG, non-GPG, and commensal strains, followed by a description of its diversity. For convenience, allele combinations of pairs of genes are referred to as genotypes of the genes. The chapter continues with a description of the non-random association of alleles from different genes as an indicator of any gene interconnectedness, followed by the results of the calculation of the non-random association of alleles from different genes. The chapter ends with a discussion of the biological implications of the gene interconnectedness.

6.1 Genotypes of Pairs of Genes

In this section, the genotypes of pairs of genes in strains of different groups are identified, and compared one with another. A comparison of the genotypes of pairs of genes in GPG and non-GPG strains, two groups of strains with different genetic backgrounds, might reveal either the genetic background or the selective pressure impact the genotypes selected. A comparison of the genotypes of pairs of genes in commensal and infection strains might reveal the influence of the immune status of the host to the genotypes selected. The description of the genotypes can be considered as an extended description of allele combinations for the *YWP1*, *HWP1*, and *EAP1* genes described previously from 1 gene to pairs of genes.

For this purpose, the distribution of the genotypes is described in a three dimensional graph; two horizontal axes represent allele combinations of the two genes, and the vertical axis represents the frequency of the strains with the particular genotypes.

6.1.1 *YWP1-HWP1* Genotypes

The *YWP1-HWP1* genotypes in each strain of GPG, non-GPG, and (GPG) commensal strains were identified. The number of different *YWP1-HWP1* genotypes found in GPG strains was ten, while in non-GPG strains 26. In total, there were 32 different *YWP1-HWP1* genotypes found in infection strains. Figure 6.1 shows the distribution of *YWP1-HWP1* genotypes for GPG strains, while Figure 6.2 shows the distribution for non-GPG strains. Genotypes containing 5-5 repeat units of *YWP1*, and 4-5 repeat units of *HWP1* (5-5+4-5) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (42.8%), compared to non-GPG strains (2.1%) (the Chi-square test for contingency tables, $p < 0.001$). Genotypes containing 5-8 repeat units of *YWP1*, and 5-5 repeat units of *HWP1* (5-8+5-5) predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in non-GPG strains (15.0%), compared to GPG strains (0%) (the Chi-square test for contingency tables, $p < 0.001$). The results show that there are specific *YWP1-HWP1* genotypes associated with each

group of infection strains.

The number of different *YWP1-HWP1* genotypes found in commensal strains was seven, compared to ten in infection strains. Figure 6.3 shows the distribution of *YWP1-HWP1* genotypes for commensal strains. Genotypes containing 5-5 repeat units of *YWP1* and 4-5 repeat units of *HWP1* (5-5+4-5) predominated (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.05$); commensal and infection strains have the same predominant *YWP1-HWP1* genotypes. Although there is no significant difference in the distribution of genotypes containing 5-5+4-5 between commensal and infection strains, there is a noticeable change in the distribution of the genotypes when comparing the commensal and infection strain's graphs. The percentage of genotypes containing 5-5+4-5 of 63% in commensal strains decreases to 43% in infection strains, which is compensated by a change in the percentage of genotypes containing 5-5+5-5, from 8% in commensal strains to 24% in infection strains. However, the results of the Chi-square test for contingency tables showed that the distributions of these genotypes in the two groups of strains are not significantly different ($p > 0.05$).

6.1.2 *YWP1-EAP1* Genotypes

The *YWP1-EAP1* genotypes in each strain of GPG, non-GPG, and (GPG) commensal strains were identified. Since there are two repeat regions in *EAP1*, *YWP1-EAP1* genotypes for each repeat region of *EAP1* are described separately as *YWP1-EAP1*(1) and *YWP1-EAP1*(2) genotypes. Later, *YWP1-EAP1* genotypes which include both repeat regions of *EAP1* are described.

The number of different *YWP1-EAP1*(1) genotypes found in GPG strains was 12, while in non-GPG strains 36. In total, there were 45 different *YWP1-EAP1*(1) genotypes found in infection strains. Figure 6.4 shows the distribution of *YWP1-EAP1*(1) genotypes for GPG strains, while Figure 6.5 shows the distribution for non-GPG strains. Genotypes containing 5-5 repeat units of *YWP1* and 10-11 repeat

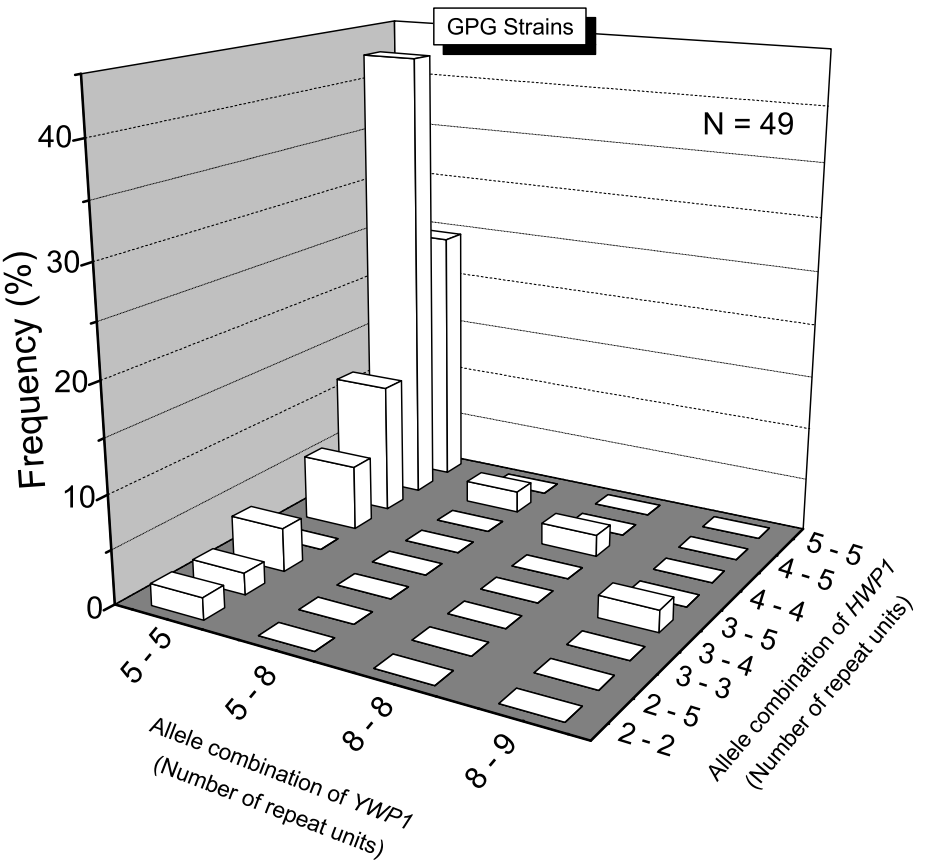

Figure 6.1: Distribution of *YWP1-HWP1* genotypes for GPG strains. N represents the number of strains characterized. There are ten different genotypes reflected by the number of bars in the graph. Genotypes 5-5+4-5 predominated (the Chi-square goodness of fit test, $p < 0.001$).

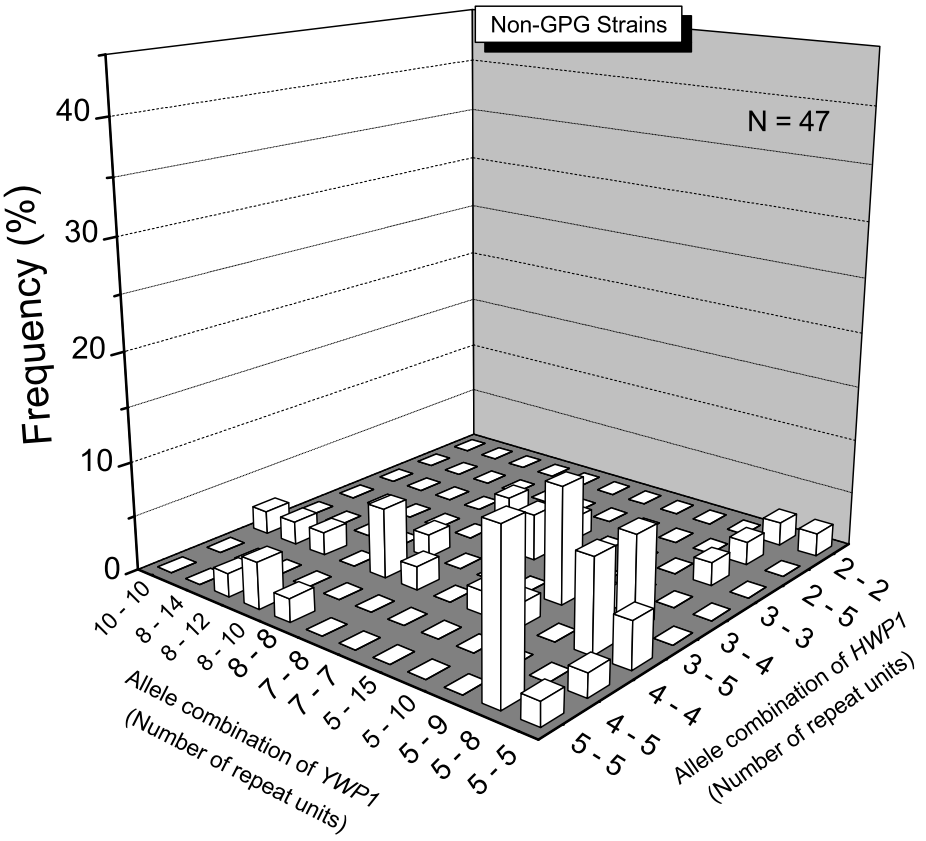

Figure 6.2: Distribution of *YWP1-HWP1* genotypes for non-GPG strains. N represents the number of strains characterized. There are 26 different genotypes reflected by the number of bars in the graph. Genotypes 5-8+5-5 predominated (the Chi-square goodness of fit test, $p < 0.001$).

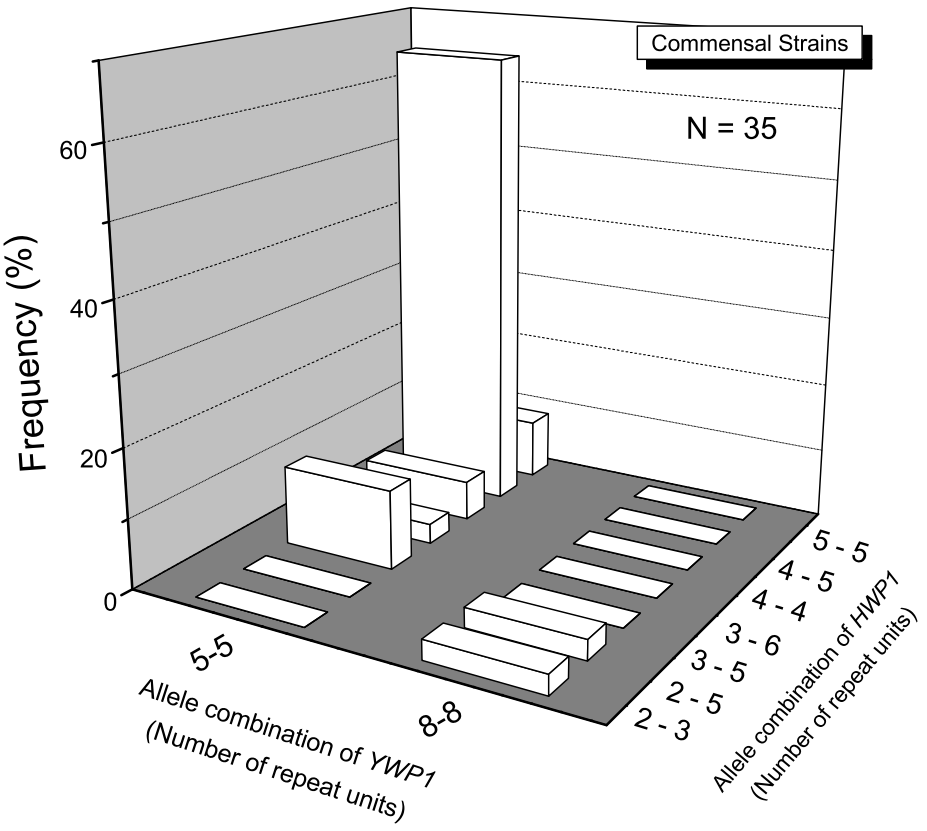

Figure 6.3: Distribution of *YWP1-HWP1* genotypes for commensal strains. N represents the number of strains characterized. There are seven different genotypes reflected by the number of bars in the graph. Genotypes 5-5+4-5 predominated (the Chi-square goodness of fit test, $p < 0.001$).

units of *EAP1*(1) (5-5+10-11) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (59.2%), compared to non-GPG strains (2.2%) (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *YWP1-EAP1*(1) genotypes observed in non-GPG strains. The results show that there is a predominant *YWP1-EAP1*(1) genotype associated with the GPG group.

The number of different *YWP1-EAP1*(1) genotypes found in commensal strains was ten, compared to 12 in infection strains. Figure 6.6 shows the distribution of *YWP1-EAP1*(1) genotypes for commensal strains. Genotypes containing 5-5 repeat units of *YWP1* and 10-11 repeat units of *EAP1*(1) (5-5+10-11) predominated (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$); commensal and infection strains have the same predominant *YWP1-EAP1*(1) genotypes.

The number of different *YWP1-EAP1*(2) genotypes found in GPG strains was eight, while in non-GPG strains 33. In total, there were 38 different *YWP1-EAP1*(2) genotypes found in infection strains. Figure 6.7 shows the distribution of *YWP1-EAP1*(2) genotypes for GPG strains, while Figure 6.8 shows the distribution for non-GPG strains. Genotypes containing 5-5 repeat units of *YWP1* and 33-48 repeat units of *EAP1*(2) (5-5+33-48) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (70.8%), compared to non-GPG strains (4.3%) (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *YWP1-EAP1*(2) genotypes observed in non-GPG strains. The results show that there is a predominant *YWP1-EAP1*(2) genotype associated with the GPG group strains.

The number of different *YWP1-EAP1*(2) genotypes found in commensal strains was seven, compared to eight in infection strains. Figure 6.9 shows the distribution of *YWP1-EAP1*(2) genotypes for commensal strains. Genotypes containing 5-5 repeat units of *YWP1* and 33-48 repeat units of *EAP1*(2) (5-5+33-48) predominated (the

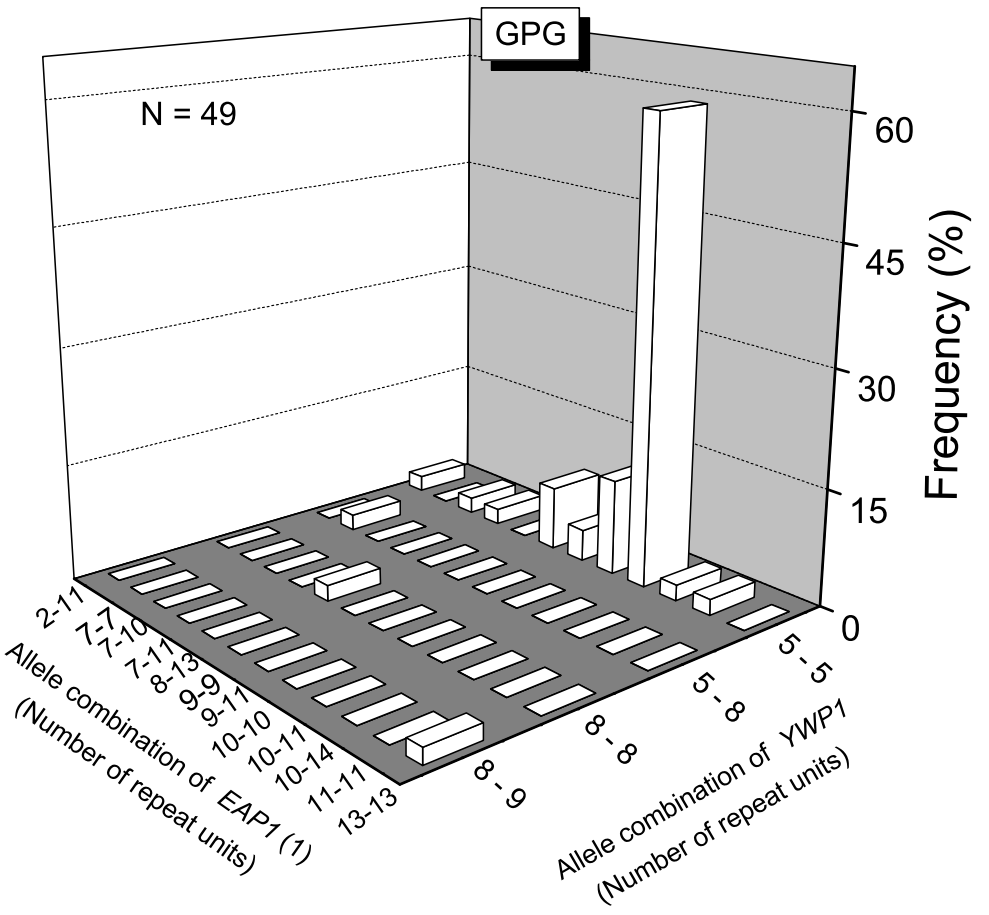

Figure 6.4: Distribution of *YWP1-EAP1*(1) genotypes for GPG strains. N represents the number of strains characterized. There are 12 different genotypes reflected by the number of bars in the graph. Genotypes 5-5+10-11 predominated (the Chi-square goodness of fit test, $p < 0.001$).

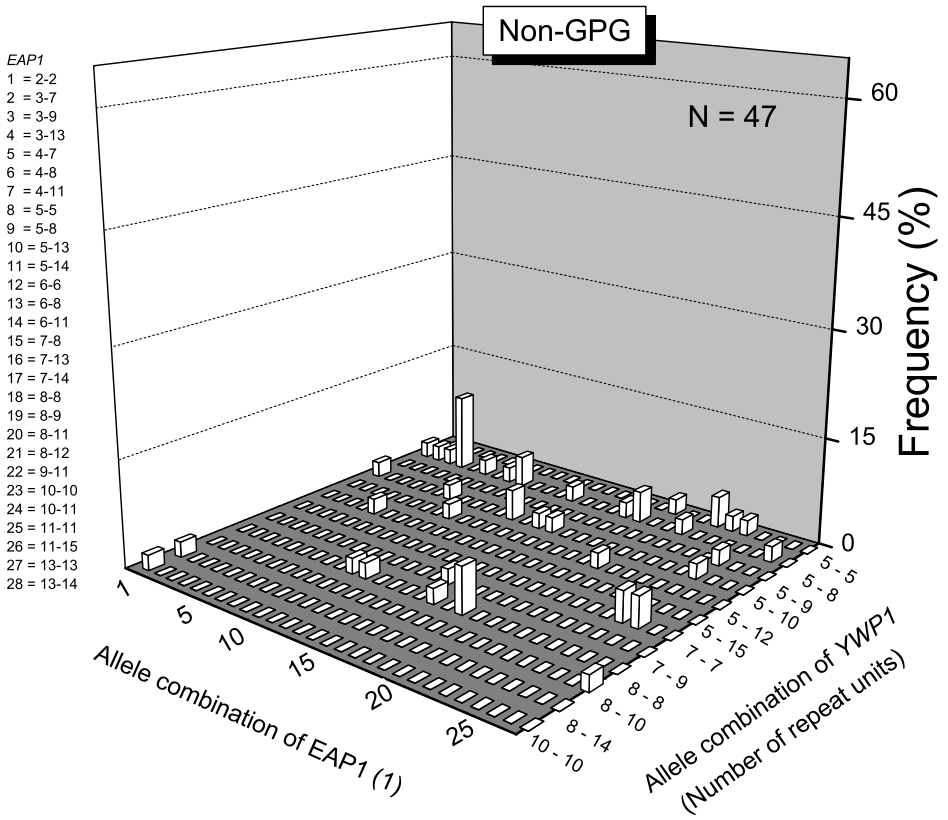

Figure 6.5: Distribution of *YWP1-EAP1*(1) genotypes for non-GPG strains. N represents the number of strains characterized. There are 36 different genotypes reflected by the number of bars in the graph.

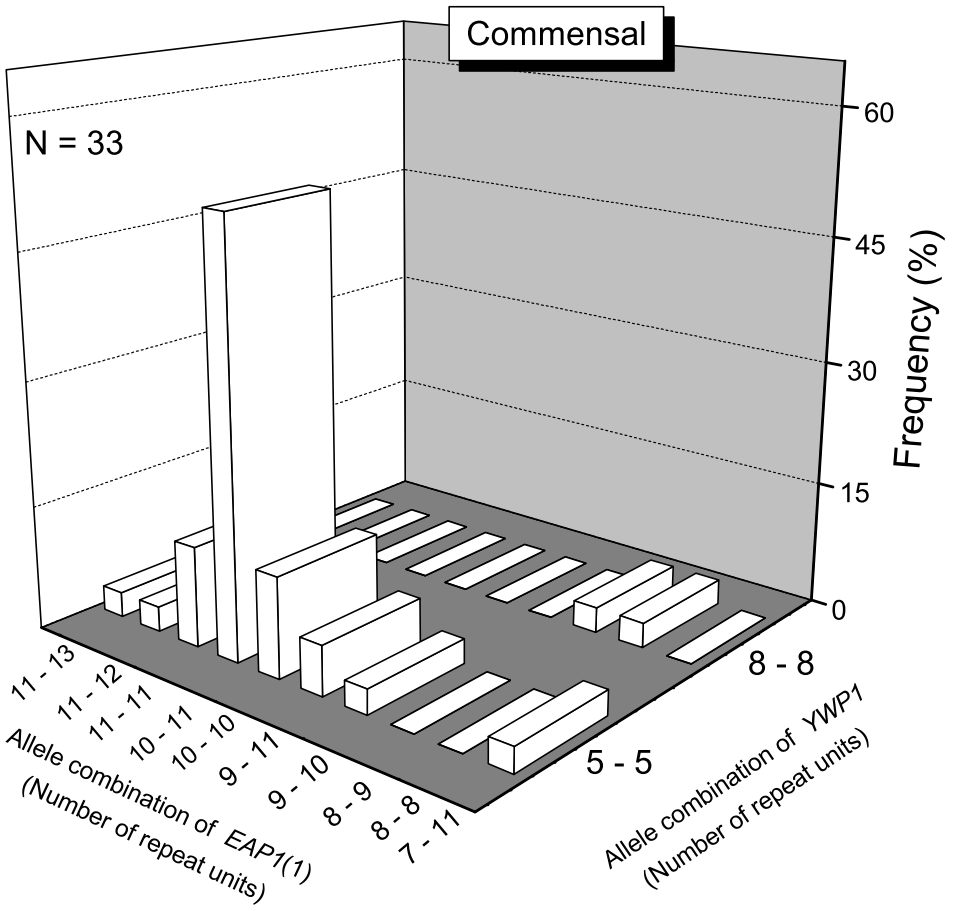

Figure 6.6: Distribution of *YWP1-EAP1*(1) genotypes for commensal strains. N represents the number of strains characterized. There are ten different genotypes found reflected by the number of bars in the graph. Genotypes 5-5+10-11 predominated (the Chi-square goodness of fit test, $p < 0.001$).

Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$); commensal and infection strains have the same predominant *YWP1-EAP1*(2) genotypes.

The number of different *YWP1-EAP1*, which include both repeat regions of *EAP1*, found in GPG strains was 13, while in non-GPG strains 46. In total, there were 56 different *YWP1-EAP1* genotypes found in infection strains. Figure 6.10.a shows the distribution of *YWP1-EAP1* genotypes for GPG strains, while Figure 6.10.b shows the distribution for non-GPG strains. Genotypes containing 5-5 repeat units of *YWP1*, 10 and 11 repeat units of *EAP1*(1), and 33-48 repeat units of *EAP1*(2) (5-5+10-11+33-48) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (60.4%), compared to non-GPG strains (2.2%) (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *YWP1-EAP1* genotypes observed in non-GPG strains. The results show that there is a predominant *YWP1-EAP1* genotype associated with the GPG group. The figure also shows that most of the non-GPG strains have distinct genotypes, i.e. only two strains have the same genotypes 5-8+3-13+102-102, and other two strains have genotypes 5-8+3-13+158-158. In addition, most of the genotypes in non-GPG strains are different from genotypes in GPG strains, i.e. only 3 strains of non-GPG strains having the same genotypes as those in GPG strains indicated by three arrows in the figure.

The number of different *YWP1-EAP1*, which include both repeat regions of *EAP1*, found in commensal strains was 11, compared to 13 in infection strains. Figure 6.10.c shows the distribution of *YWP1-EAP1* genotypes for commensal strains. Genotypes containing 5-5 repeat units of *YWP1*, 10 and 11 repeat units of *EAP1*(1), and 33-48 repeat units of *EAP1*(2) (5-5+10-11+33-48) predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results shown by infection strains (Figure 6.10.a). The Chi-square test for contingency tables showed no

significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$); commensal and infection strains have the same predominant *YWP1-EAP1* genotypes.

6.1.3 *HWP1-EAP1* Genotypes

The *HWP1-EAP1* genotypes in each strain of GPG, non-GPG, and (GPG) commensal strains were identified. Since there are two repeat regions in *EAP1*, *HWP1-EAP1* genotypes for each repeat region of *EAP1* are described separately as *HWP1-EAP1*(1) and *HWP1-EAP1*(2) genotypes. Later, *HWP1-EAP1* genotypes which include both repeat regions of *EAP1* are described.

The number of different *HWP1-EAP1*(1) genotypes found in GPG strains was 21, while in non-GPG strains 40. In total, there were 58 different *HWP1-EAP1*(1) genotypes found in infection strains. Figure 6.11 shows the distribution of *HWP1-EAP1*(1) genotypes for GPG strains, while Figure 6.12 shows the distribution for non-GPG strains. Genotypes containing 4-5 repeat units of *HWP1* and 10-11 repeat units of *EAP1*(1) (4-5+10-11) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (33.3%), compared to non-GPG strains (2.1%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of these genotypes between GPG and non-GPG strains ($p < 0.001$). There are no obvious predominant *HWP1-EAP1*(1) genotypes observed in non-GPG strains. The results show that there is a predominant *HWP1-EAP1*(1) genotype associated with the GPG group.

The number of different *HWP1-EAP1*(1) genotypes found in commensal strains was 17, compared to 21 in infection strains. Figure 6.13 shows the distribution of *HWP1-EAP1*(1) genotypes for commensal strains. Genotypes containing 4-5 repeat units of *HWP1* and 10-11 repeat units of *EAP1*(1) (4-5+10-11) predominated (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$); commensal and infection strains have the same predominant *HWP1-EAP1*(1)

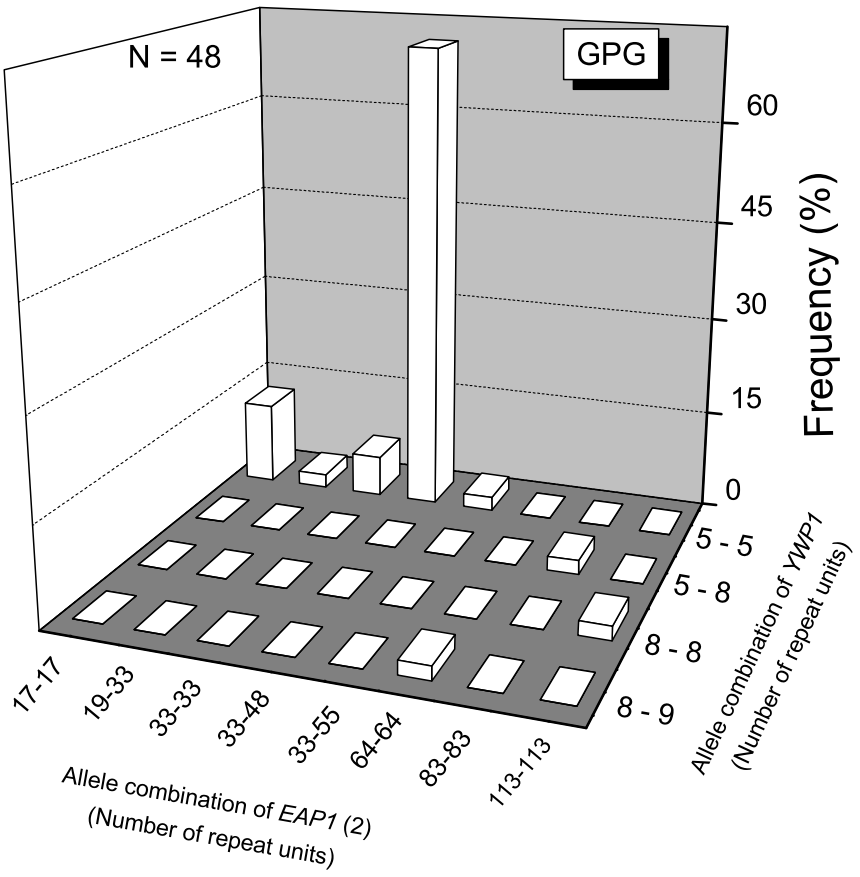

Figure 6.7: Distribution of *YWP1-EAP1*(2) genotypes for GPG strains. N represents the number of strains characterized. There are 12 different genotypes reflected by the number of bars in the graph. Genotypes 5-5 + 33-48 predominated (the Chi-square goodness of fit test, $p < 0.001$).

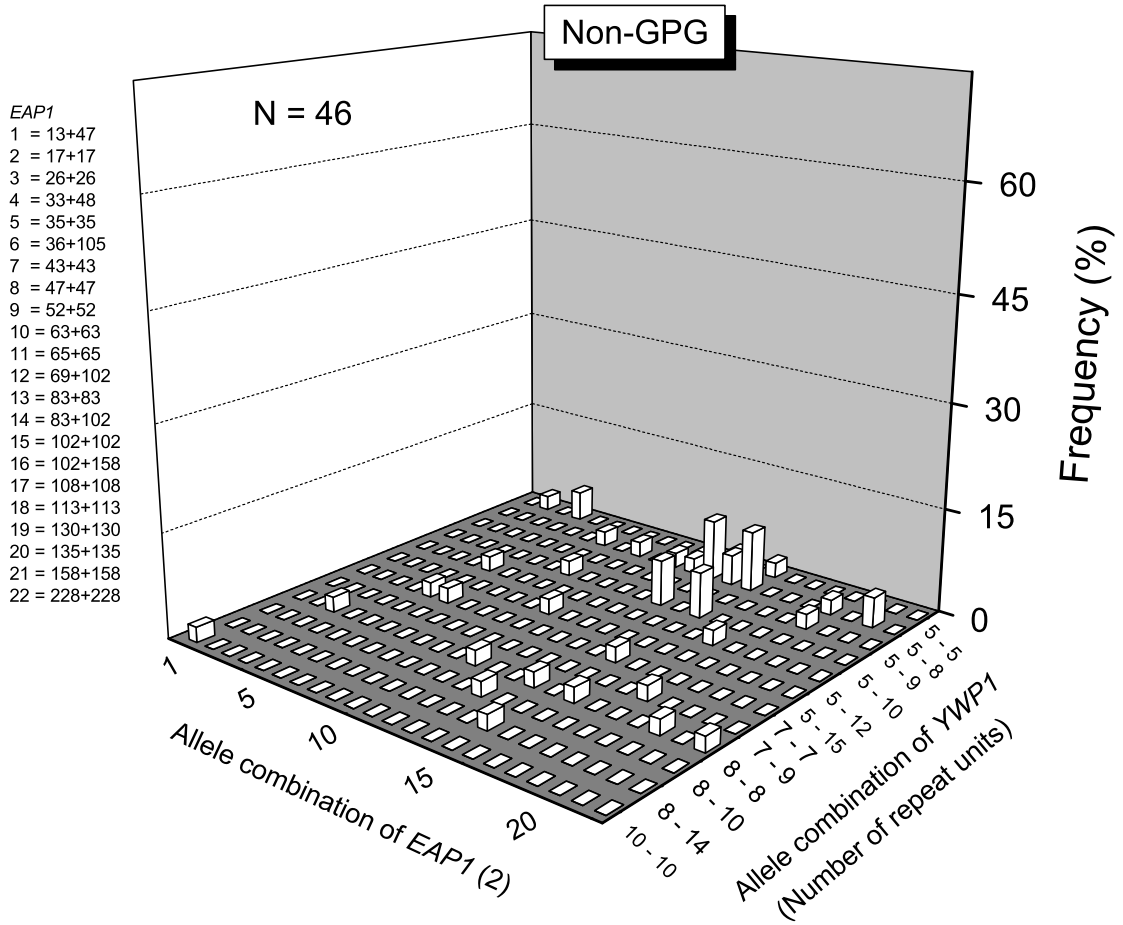

Figure 6.8: Distribution of *YWP1-EAP1*(2) genotypes for non-GPG strains. N represents the number of strains characterized. There are 33 different genotypes reflected by the number of bars in the graph. There is no obvious pattern of any predominant genotypes observed.

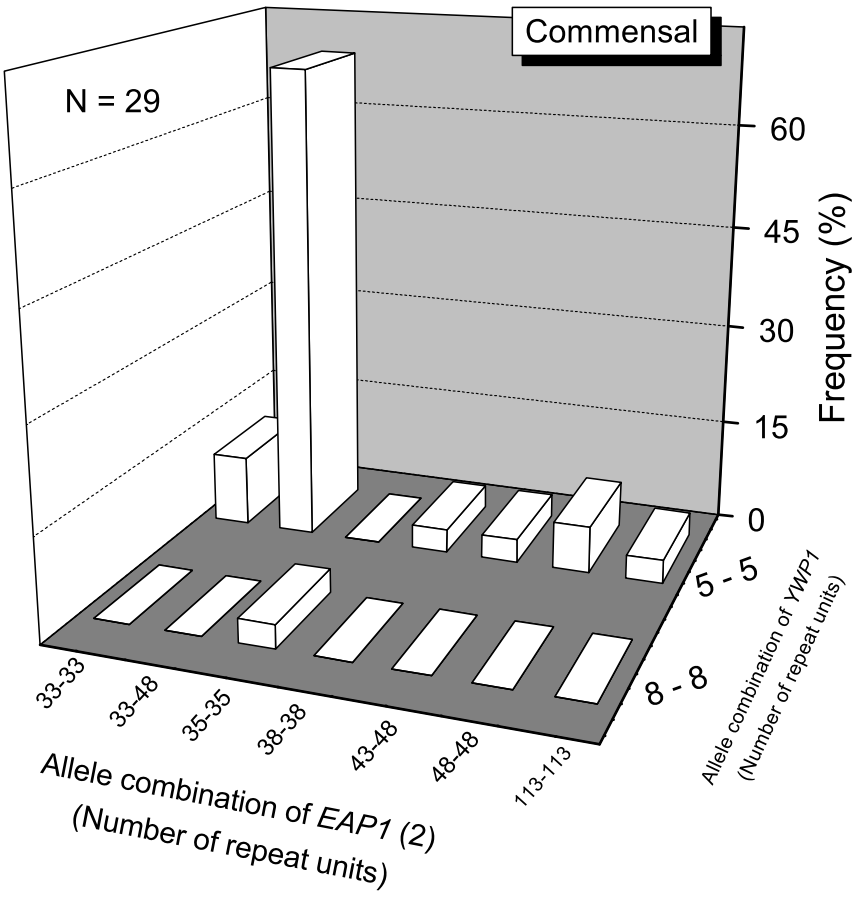

Figure 6.9: Distribution of *YWP1-EAP1*(2) genotypes for commensal strains. N represents the number of strains characterized. There are seven different genotypes reflected by the number of bars in the graph. Genotypes containing 5-5 + 33-48 predominated (the Chi-square goodness of fit test, $p < 0.001$).

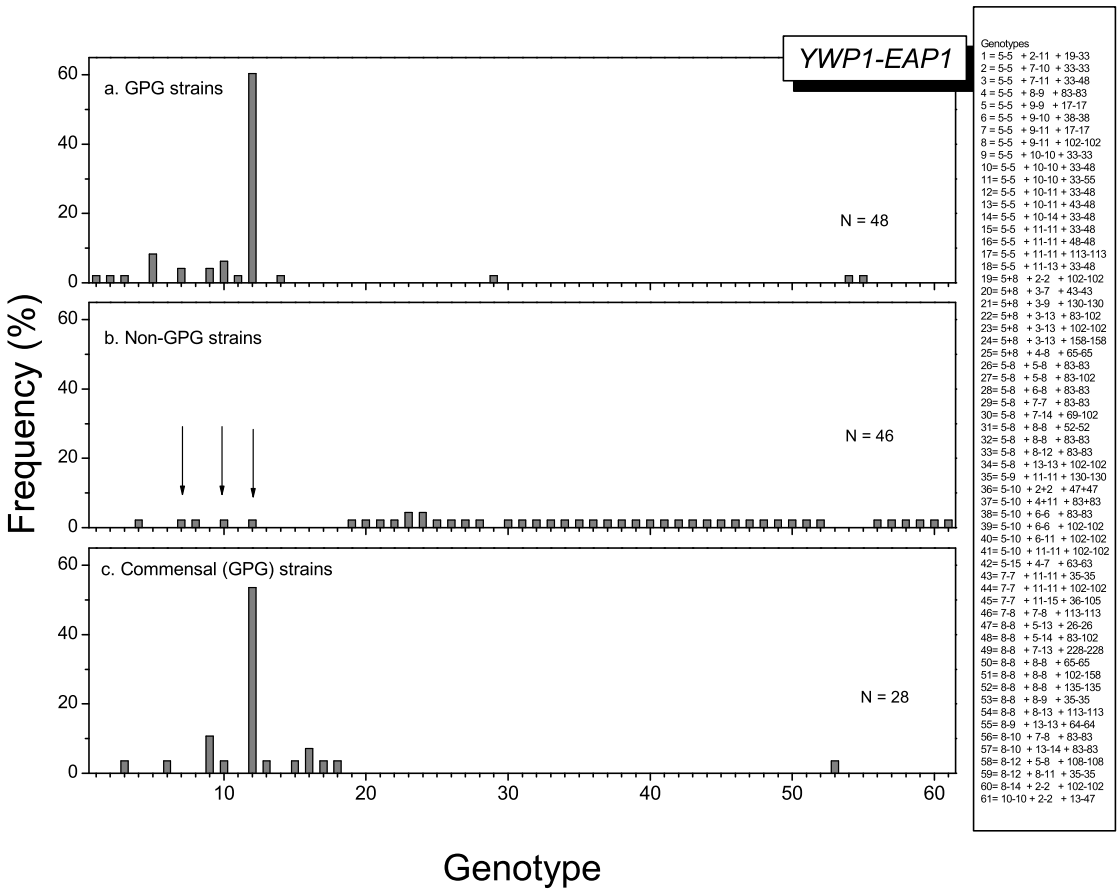

Figure 6.10: Distribution of *YWP1-EAP1* genotypes for GPG, non-GPG, and commensal strains. In this analysis, both repeat regions of *EAP1* are included. N represents the number of strains characterized. Genotypes containing 5-5+10-11+33-48 predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (60.4%), compared to non-GPG strains (2.2%) (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *YWP1-EAP1* genotypes observed in non-GPG strains. Genotypes containing 5-5+10-11+33-48 predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), and no significant difference in the distribution of these genotypes between commensal and infection strains (the Chi-square test for contingency tables, $p > 0.1$). The three arrows indicate the genotypes in non-GPG strains which match the genotypes in GPG strains; other genotypes in non-GPG strains are different from genotypes in GPG strains.

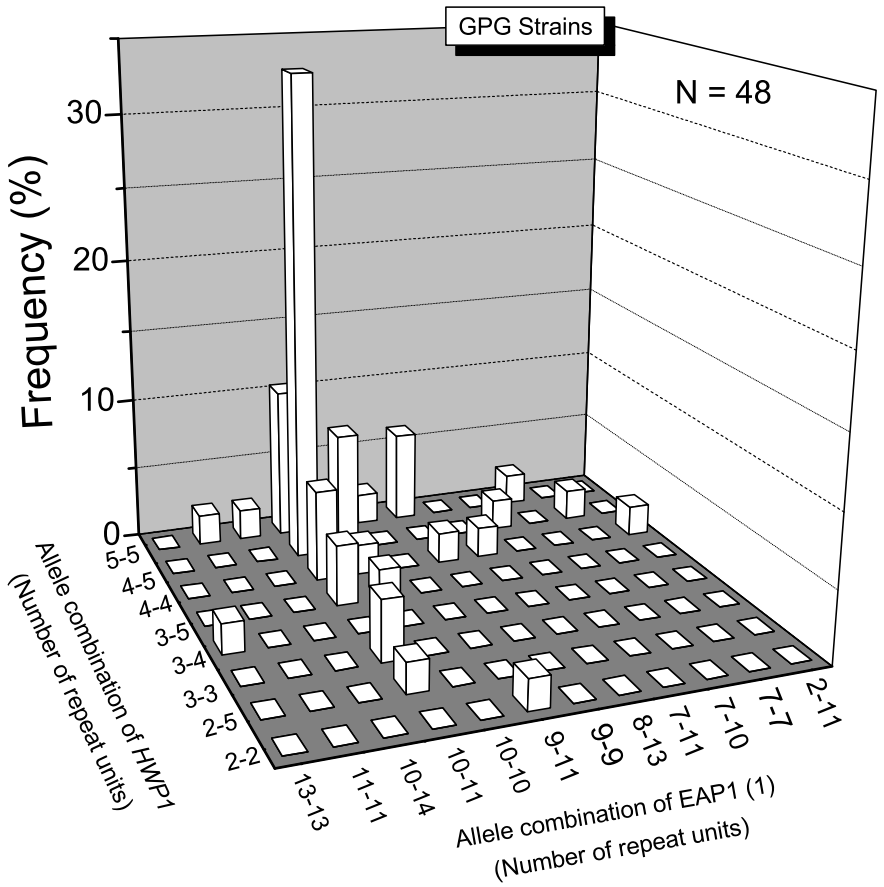

Figure 6.11: Distribution of *HWP1-EAP1*(1) genotypes for GPG strains. N represents the number of strains characterized. There are 21 different genotypes reflected by the number of bars in the graph. Genotypes 4-5+10-11 predominated (the Chi-square goodness of fit test, $p < 0.001$)

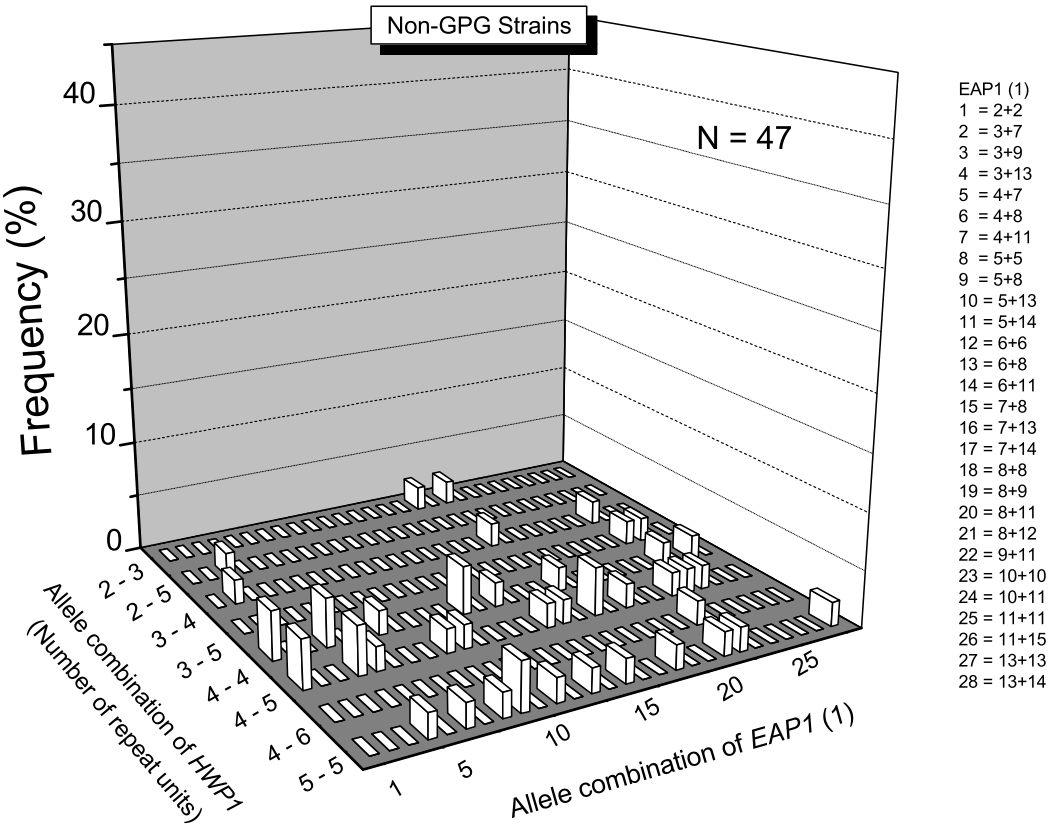

Figure 6.12: Distribution of *HWP1-EAP1*(1) genotypes for non-GPG strains. N represents the number of strains characterized. There are 40 different genotypes reflected by the number of bars in the graph.

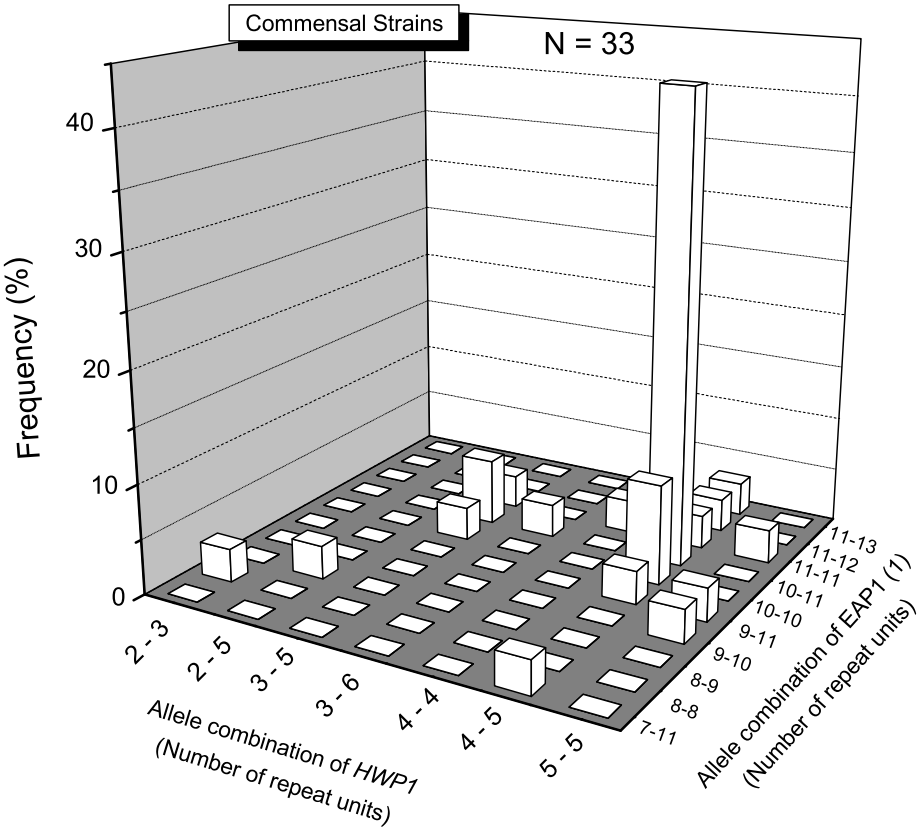

Figure 6.13: Distribution of *HWP1-EAP1*(1) genotypes for commensal strains. N represents the number of strains characterized. There are 17 different genotypes reflected by the number of bars in the graph. Genotypes 4-5+10-11 predominated (the Chi-square goodness of fit test, $p < 0.001$)

genotypes.

The number of different *HWP1-EAP1*(2) genotypes found in GPG strains was 17, while in non-GPG strains 32. In total, there were 47 different *HWP1-EAP1*(2) genotypes found in infection strains. Figure 6.14 shows the distribution of *HWP1-EAP1*(2) genotypes for GPG strains, while Figure 6.15 shows the distribution for non-GPG strains. Genotypes containing 4-5 repeat units of *HWP1* and 33-48 repeat units of *EAP1*(2) (4-5+33-48) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (41.7%),

compared to non-GPG strains (4.3%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of these genotypes between GPG and non-GPG strains ($p < 0.001$). Likewise, genotypes containing 4-4+102-102 and 5-5+83-83 predominated in non-GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in non-GPG strains (26.1%), compared to GPG strains (0%), i.e. the Chi-square test for contingency tables showed a significant difference in the distribution of these genotypes between GPG and non-GPG strains ($p < 0.001$). The results show that there are predominant *HWP1-EAP1*(2) genotypes associated with each group of infection strains.

The number of different *HWP1-EAP1*(2) genotypes found in commensal strains was 10, compared to 17 in infection strains. Figure 6.16 shows the distribution of *HWP1-EAP1*(2) genotypes commensal strains. Genotypes containing 4-5 repeat units of *HWP1* and 33-48 repeat units of *EAP1*(2) (4-5+33-48) predominated (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$); commensal and infection strains have the same predominant *HWP1-EAP1*(2) genotypes.

The number of different *HWP1-EAP1*, which include both repeat regions of *EAP1*, found in GPG strains was 21, while in non-GPG strains 45. In total, there were 63 different *HWP1-EAP1* genotypes found in infection strains. Figure 6.17.a shows the distribution of *HWP1-EAP1* genotypes for GPG strains, while Figure 6.17.b shows the distribution for non-GPG strains. Genotypes containing 5-6 repeat units of *HWP1*, 10-11 repeat units of *EAP1*(1), and 33-48 repeat units of *EAP1*(2) (5-6+10-11+33-48) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (33.3%), compared to non-GPG strains (2.2%) (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *HWP1-EAP1* genotypes observed in non-GPG strains. The results show that there is a predominant *HWP1-EAP1* genotype associated with

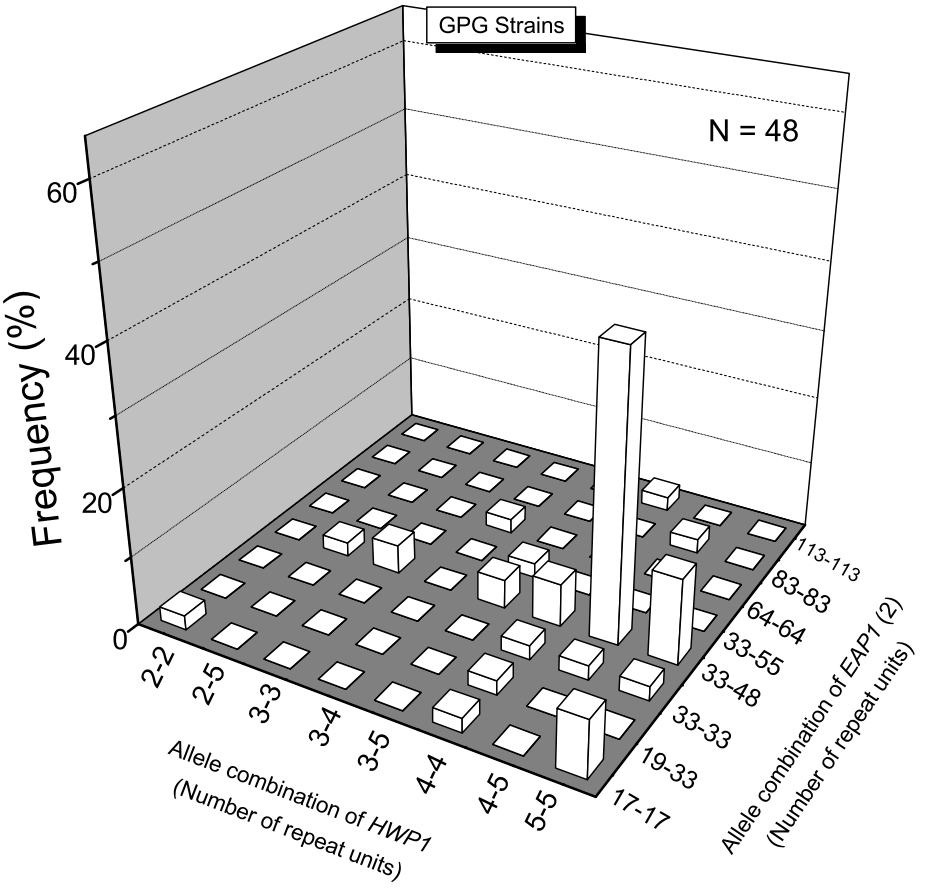

Figure 6.14: Distribution of *HWP1-EAP1*(2) genotypes for GPG strains. N represents the number of strains characterized. There are 17 different genotypes reflected by the number of bars in the graph. Genotypes 4-5+33-48 predominated (the Chi-square goodness of fit test, $p < 0.001$).

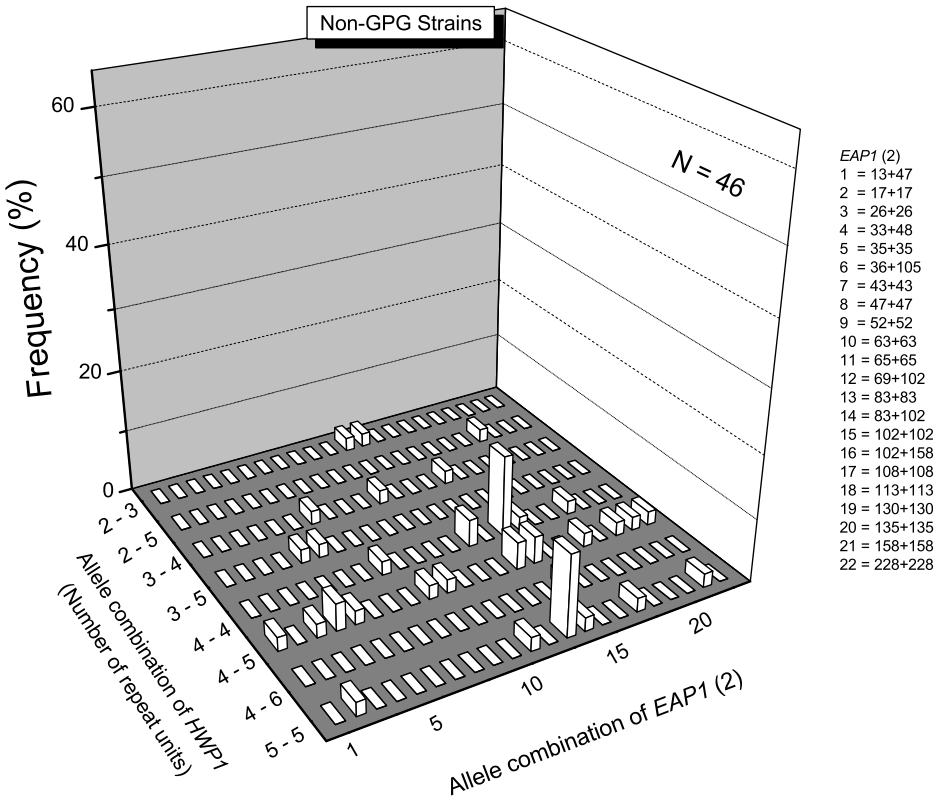

Figure 6.15: Distribution of *HWP1-EAP1*(2) genotypes for non-GPG strains. N represents the number of strains characterized. There are 32 different genotypes reflected by the number of bars in the graph. Genotypes 4-4+102-102 and 5-5+83-83 predominated (the Chi-square goodness of fit test, $p < 0.001$).

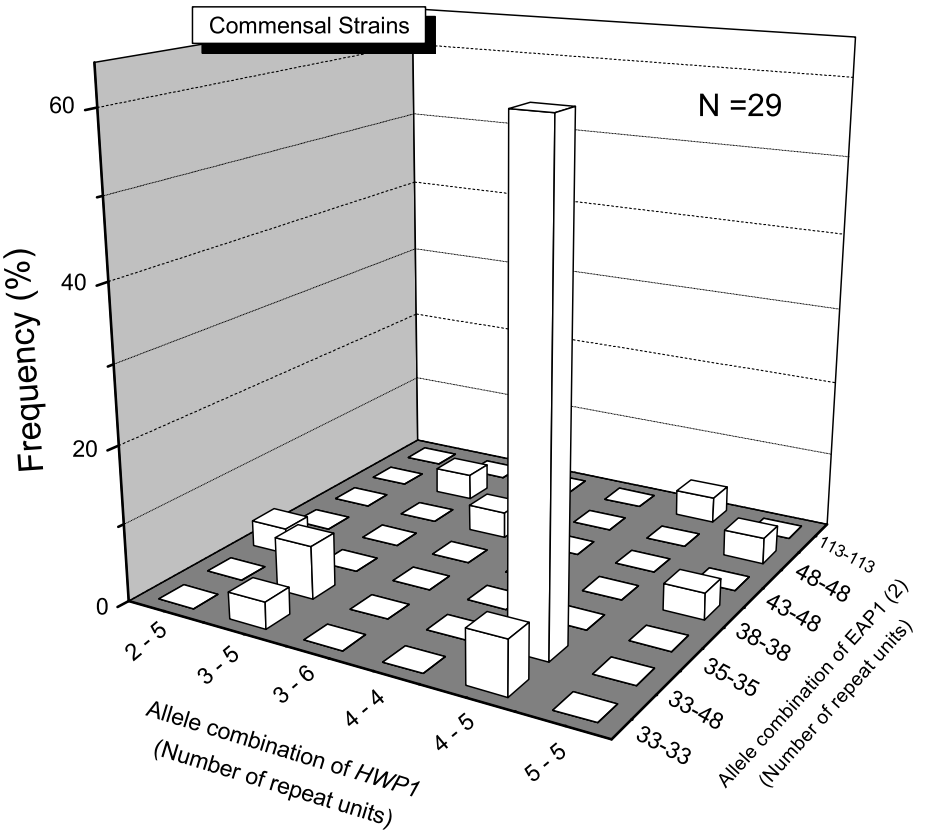

Figure 6.16: Distribution of *HWP1-EAP1*(2) genotypes for commensal strains. N represents the number of strains characterized. There are 10 different genotypes reflected by the number of bars in the graph. Genotypes 4-5+33-48 predominated (the Chi-square goodness of fit test, $p < 0.001$)

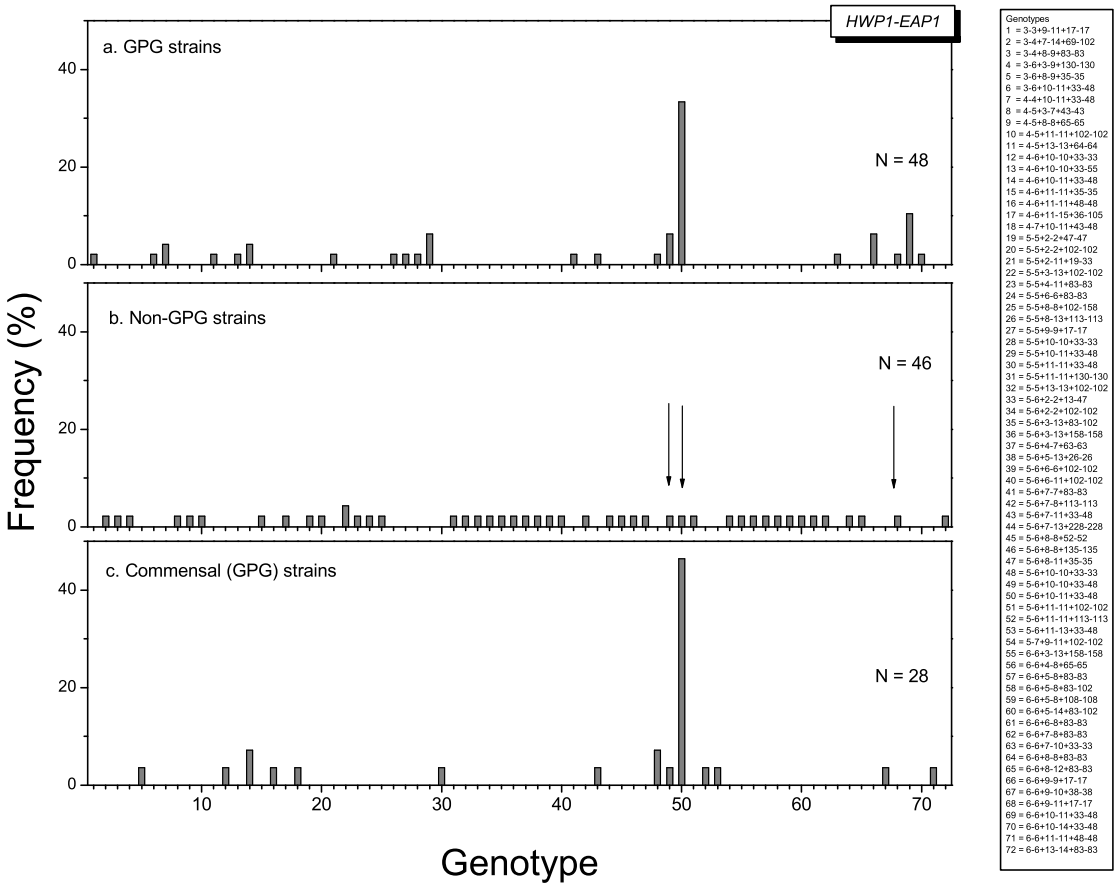

Figure 6.17: Distribution of *HWP1-EAP1* genotypes for GPG, non-GPG, and commensal strains. N represents the number of strains characterized. Genotypes containing 5-5+10-11+33-48 predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains (33.3%), compared to non-GPG strains (2.2%) (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *HWP1-EAP1* genotypes observed in non-GPG strains. Genotypes containing 5-6+10-11+33-48 predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains (Figure.a). The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$). The three arrows indicate the genotypes in non-GPG strains which match the genotypes in GPG strains; other genotypes in non-GPG strains are different from genotypes in GPG strains.

the GPG group. The figure also shows that most of the non-GPG strains have distinct genotypes, i.e. only two strains have the same genotypes 5-5+3-13+102-102. In addition, most of the genotypes in non-GPG strains are different from genotypes in GPG strains, i.e. only three strains of non-GPG strains having the same genotypes as those of GPG strains indicated by three arrows in the figure.

The number of different *HWP1-EAP1*, which include both repeat regions of *EAP1*, found in commensal strains was 14, compared to 21 in infection strains. Figure 6.17.c shows the distribution of *HWP1-EAP1* genotypes for commensal strains. Genotypes containing 5-6 repeat units of *YWP1*, 10-11 repeat units of *EAP1*(1), and 33-48 repeat units of *EAP1*(2) (5-6+10-11+33-48) predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains (Figure 6.17.a). The Chi-square test for contingency tables showed no significant difference in the distribution of these genotypes between commensal and infection strains ($p > 0.1$); commensal and infection strains have the same predominant *HWP1-EAP1* genotypes.

6.2 *YWP1-HWP1-EAP1* Genotypes

In this section, the description of genotypes of two genes in the previous section is extended to description of the genotypes of three genes, i.e. *YWP1-HWP1-EAP1* genotypes.

The *YWP1-HWP1-EAP1* genotypes in each strain of GPG, non-GPG, and commensal strains were identified. Since there are two repeat regions in *EAP1*, *YWP1-HWP1-EAP1* genotypes for each repeat region of *EAP1* are described separately as *YWP1-HWP1-EAP1*(1) and *YWP1-HWP1-EAP1*(2) genotypes.

The number of different *YWP1-HWP1-EAP1*(1) genotypes found in GPG strains was 21, while in non-GPG strains 44. In total, there were 62 different *YWP1-HWP1-EAP1*(1) genotypes found in infection strains. In addition, there were 17 different *YWP1-HWP1-EAP1*(1) genotypes found in (GPG) commensal strains. Figure 6.18 shows the distribution of *YWP1-HWP1-EAP1*(1) genotypes for GPG and non-GPG

strains, while Figure 6.19 shows the distribution for (GPG) commensal strains, where the results of (GPG) infection strains were added for comparison.

Genotypes containing 5-5 repeat units of *YWP1*, 4-5 repeat units of *HWP1*, and 10-11 repeat units of *EAP1*(1) (5-5+4-5+10-11) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$), and overrepresented in GPG strains, compared to non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *YWP1-HWP1-EAP1*(1) genotypes observed in non-GPG strains. The figure also shows that most of the non-GPG strains have distinct genotypes, i.e. only six strains have the same genotypes. In addition, most of the genotypes in non-GPG strains are different from genotypes in GPG strains, i.e. only three strains of non-GPG strains having the same genotypes as those of GPG strains indicated by three arrows in the figure.

Genotypes containing 5-5+4-5+10-11 predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. There is no significant difference in the distribution of these genotypes between commensal and infection strains (the Chi-square test for contingency tables, $p < 0.001$).

The number of different *YWP1-HWP1-EAP1*(2) genotypes found in GPG strains was 17, while in non-GPG strains 35. In total, there were 49 different *YWP1-HWP1-EAP1*(2) genotypes fund in infection strains. In addition, there were 11 different *YWP1-HWP1-EAP1*(2) genotypes found in (GPG) commensal strains. Figure 6.20 shows the distribution of *YWP1-HWP1-EAP1*(2) genotypes for GPG and non-GPG strains, while Figure 6.21 shows the distribution for commensal strains, where the results of (GPG) infection strains were added for comparison.

Genotypes containing 5-5 repeat units of *YWP1*, 4-5 repeat units of *HWP1*, and 33-48 repeat units of *EAP1*(2) (5-5+4-5+33-48) predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). These genotypes were overrepresented in GPG strains, compared to non-GPG strains (the Chi-square test for contingency tables, $p < 0.001$). There are no obvious predominant *YWP1-HWP1-EAP1*(2) genotypes observed in non-GPG strains. The figure also shows only two different genotypes

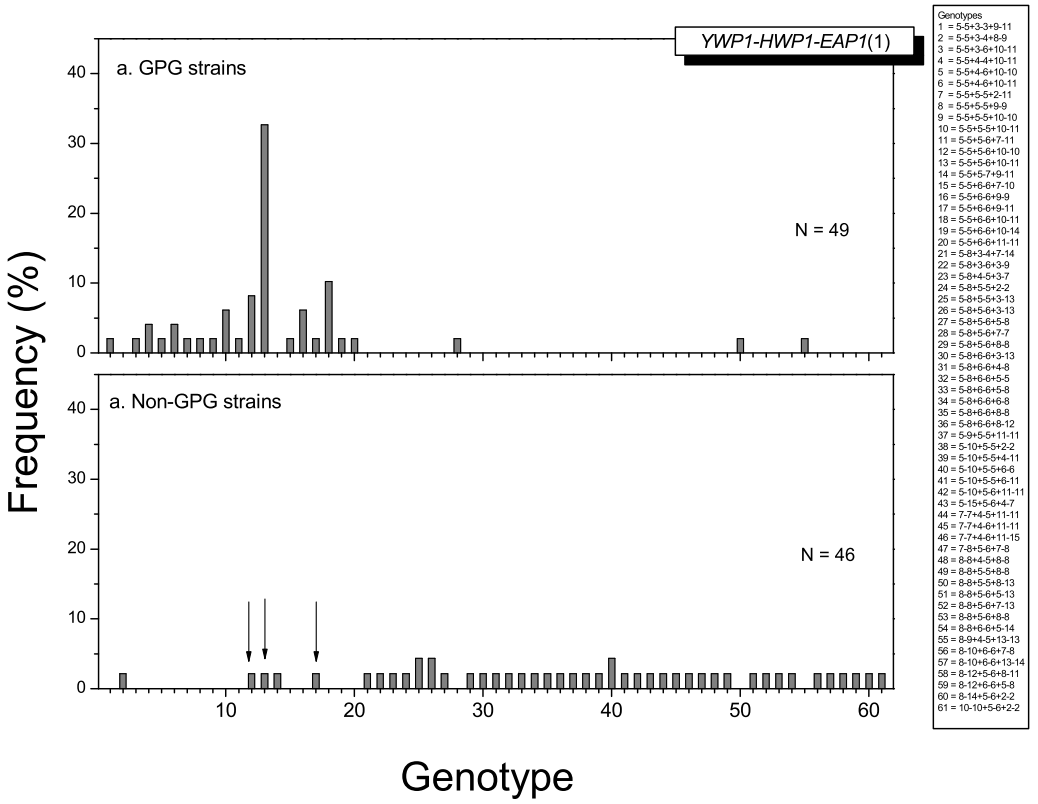

Figure 6.18: Distribution of *YWP1-HWP1-EAP1*(1) genotypes for: a. GPG strains, and b. non-GPG strains. There are 21 and 44 different genotypes in GPG and non-GPG strains, respectively. In total, there are 62 different genotypes for infection strains. N represents the number of strains characterized. Genotypes containing 5-5+4-5+10-11 predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There are no obvious predominant *YWP1-HWP1-EAP1*(1) genotypes observed in non-GPG strains.

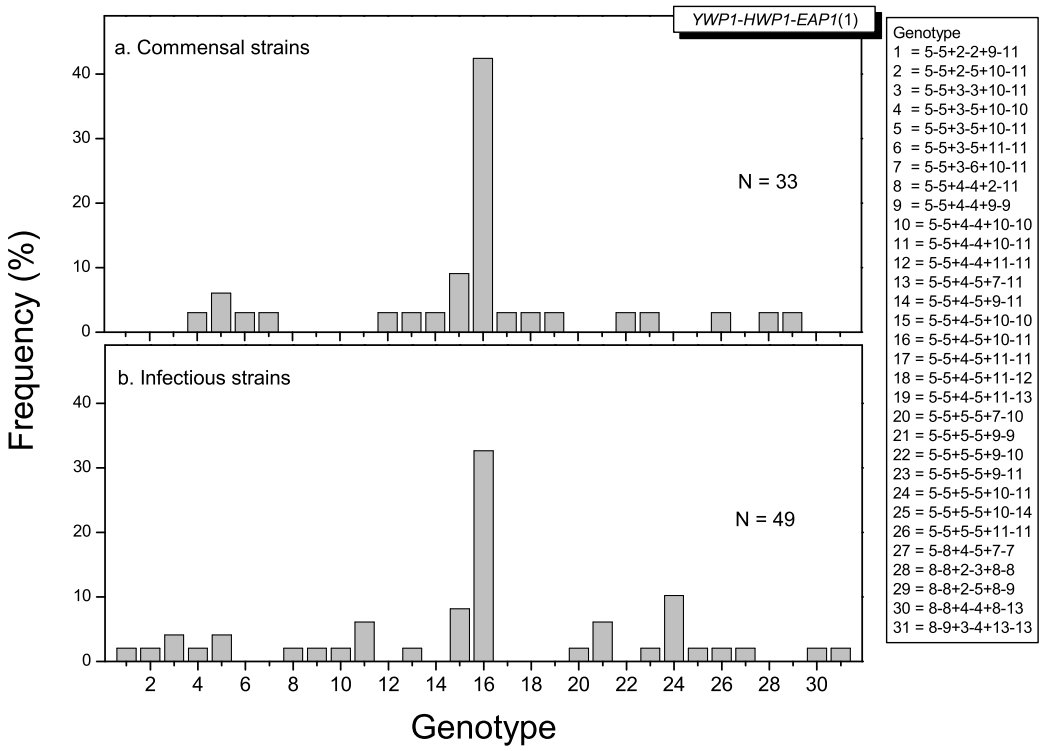

Figure 6.19: Distribution of *YWP1-HWP1-EAP1*(1) genotypes for: a. commensal strains, and b. infection strains. There are 17 different genotypes in commensal strains, compared to 21 in infection strains. Genotypes containing 5-5+4-5+10-11 predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. There is no significant difference in the distribution of these genotypes between commensal and infection strains (the Chi-square test for contingency tables, $p < 0.001$).

in non-GPG strains match the genotypes in GPG strains indicated by two arrows in the figure.

Genotypes containing 5-5 repeat units of *YWP1*, 4-5 repeat units of *HWP1*, and 33-48 repeat units of *EAP1*(2) (5-5+4-5+33-48) predominated in commensal strains (the Chi-square goodness of fit test, $p < 0.001$), i.e. similar to the results for infection strains. There is no significant difference in distribution of these genotypes between commensal and infection strains (the Chi-square test for contingency tables, $p < 0.001$).

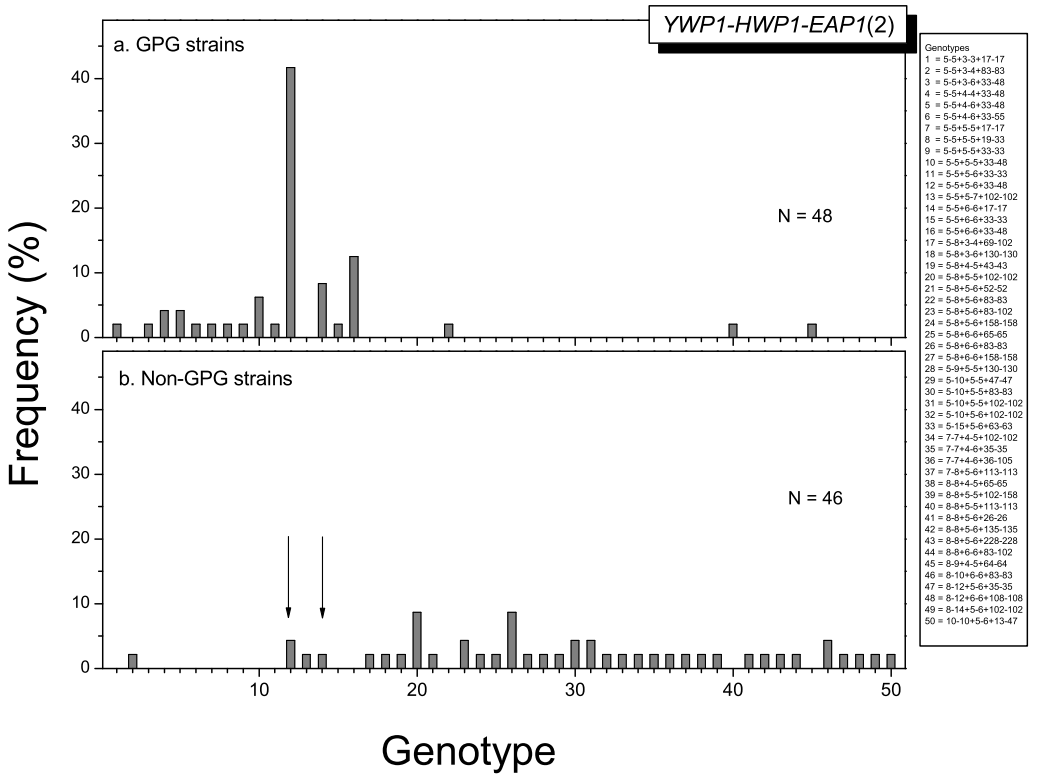

Figure 6.20: Distribution of *YWP1-HWP1-EAP1*(2) genotypes for: a. GPG strains, and b. non-GPG strains. There are 17 and 35 different genotypes found in GPG and non-GPG strains, respectively. In total, there are 49 different genotypes for infection strains. N represents the number of strains characterized. Genotypes containing 5-5+4-5+10-11 predominated in GPG strains (the Chi-square goodness of fit test, $p < 0.001$). There are no obvious predominant *YWP1-HWP1-EAP1*(2) genotypes observed in non-GPG strains.

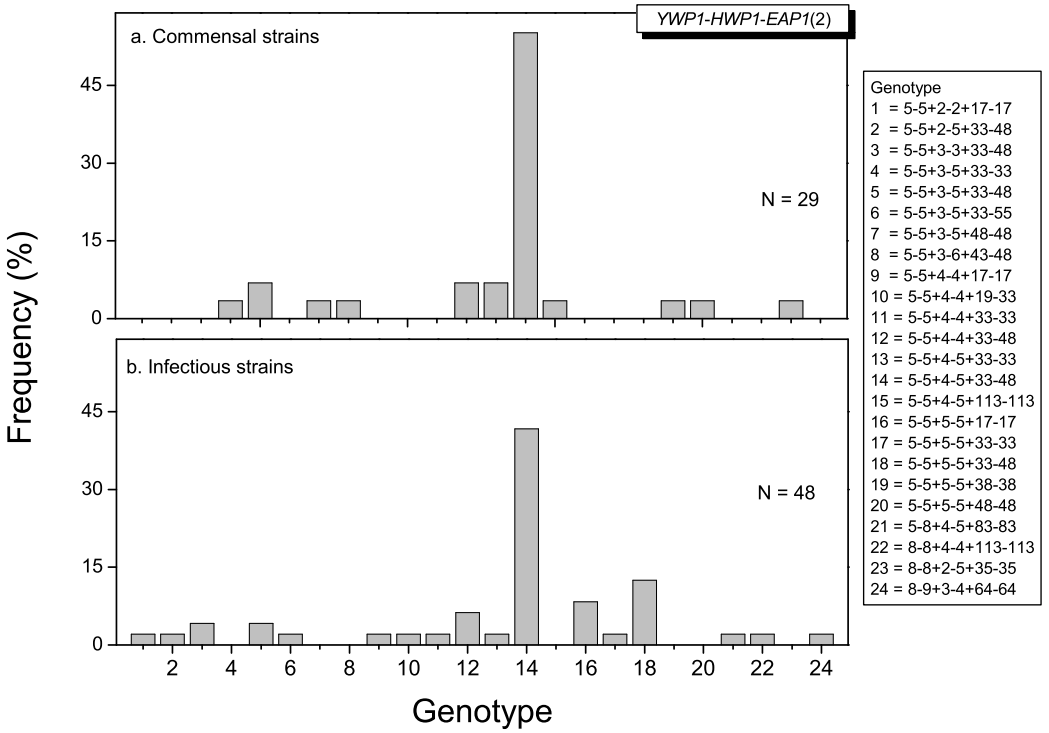

Figure 6.21: Distribution of *YWP1-HWP1-EAP1*(2) genotypes for: a. commensal strains, and b. infection strains. There are 11 different genotypes found in commensal strains, compared to 17 in infection strains. Genotypes containing 5-5+4-5+33-48 predominated in both commensal and infection strains. (the Chi-square goodness of fit test, $p < 0.001$). There is no significant difference in the distribution of these genotypes between commensal and infection strains (the Chi-square test for contingency tables, $p < 0.001$).

6.3 Diversity of Genotypes of Pairs of Genes

The diversity of the genotypes of pairs of genes is quantified by the index of diversity λ_G which is defined as,

$$\lambda_G = \sum_{j=1}^{m} \left(1 - \frac{1}{N(N-1)} \sum_{i=1}^{k} n_i(n_i - 1) \right), \tag{6.1}$$

where m is the number of genes, k is the number of different genotypes, N is the number of strains, and n_i is the number of strains having ith genotype of jth gene. Notice that Equation 6.1 is generalized form of the index of diversity Equation 3.1 (see page 48) from one gene to m genes.

The index of diversity λ_G of the genotypes of *YWP1-HWP1*, *YWP1-EAP1*, and *HWP1-EAP1*, was calculated for GPG, non-GPG, and commensal strains. The results of the calculation are summarized in Table 6.1, and presented graphically in Figure 6.22.

Figure 6.22.a shows that the values of the index of diversity of genotypes of all pairs of genes for non-GPG strains are systematically higher than those for GPG strains. In addition, the values of the index of the diversity of genotypes for GPG strains are dependent on the pair of genes; pairs that include *HWP1* have higher index of diversity values than other pairs. Figure 6.22.b shows that the variation of the index of diversity, in respect to the pairs of genes, is similar in both commensal and infection strains, with an exception of the *YWP1-HWP1* pair.

6.4 Non-Random Association Between Alleles of Different Genes

In this section, the non-random association between alleles of different genes is analysed. Any association can be quantified by a linkage disequilibrium analysis. This section begins with a description of linkage disequilibrium, followed by the determination of the non-random association between alleles of different genes.

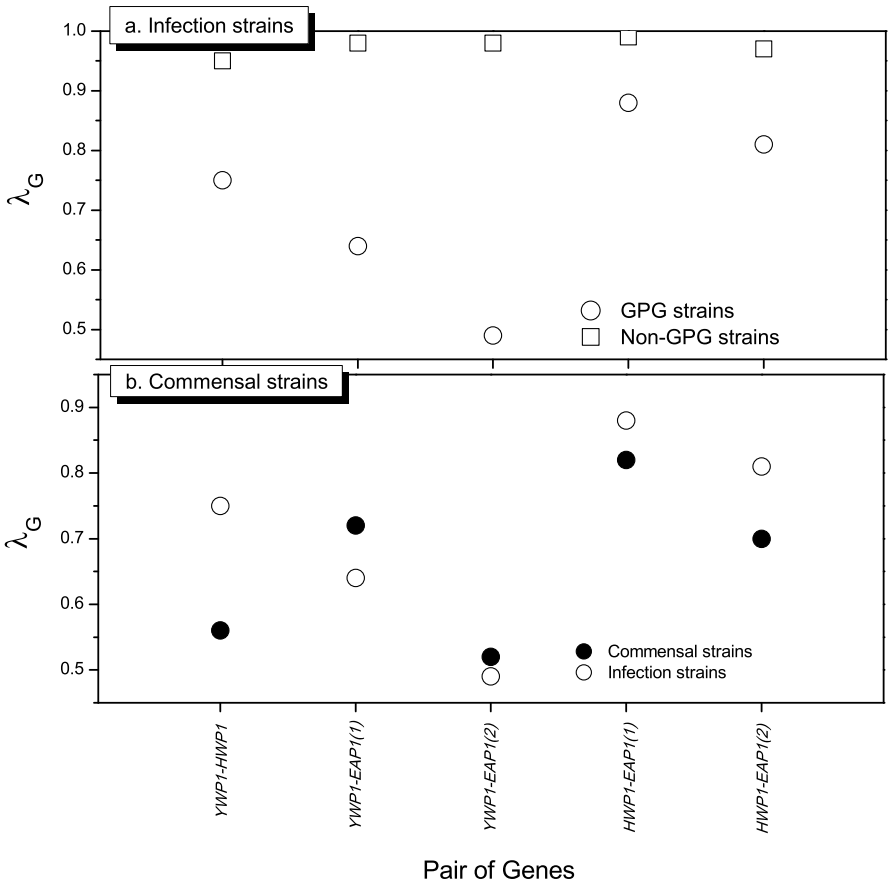

Figure 6.22: The index of diversity of genotypes of pairs of genes in a. GPG and non-GPG strains, and b. (GPG) commensal strains. The data of (GPG) infection strains is added to b for comparison.

Table 6.1: The index of diversity λ_G of genotypes of pairs of genes in GPG, non-GPG, and (GPG) commensal strains.

Pair of genes	Index of Diversity		
	GPG strains	Non-GPG strains	Commensal strains
YWP1-HWP1	0.75	0.95	0.56
YWP1-EAP1(1)	0.64	0.98	0.72
YWP1-EAP1(2)	0.49	0.98	0.52
HWP1-EAP1(1)	0.88	0.99	0.82
HWP1-EAP1(2)	0.81	0.97	0.70
YWP1-HWP1-EAP1(1)	0.88	0.99	0.82
YWP1-HWP1-EAP1(2)	0.81	0.98	0.70

6.4.1 Linkage Disequilibrium : A Measure of Non-Random Association Between Alleles of Different Genes

Linkage disequilibrium analysis gives a measure of any non-random association between alleles of different genes. Determination of the linkage disequilibrium between alleles of different genes in these studies followed a similar procedure used to determine the linkage disequilibrium of multilocus alleles, where an index of association (I_A) is used as a measure of the linkage [97, 98, 99]. Determination of I_A is based on the evaluation of a pairwise distance defined as the number of loci at which two individual strains differ one from another. For a single locus j analyzed in n haploid strains, the probability that two strains selected at random have different alleles at that locus can be written as,

$$h_j = 1 - \frac{1}{N(N-1)} \sum_{i=1}^{k} n_i(n_i - 1), \tag{6.2}$$

where n_i is the number of strains which have ith allele, k is the number of different alleles, and N is number of strains. Notice that Equation 6.2 is identical to the Simpson index of diversity Equation 3.1 (see page 48). For N haploid strains, there are $N(N-1)/2$ possible pairs of alleles, and each pair can have a distance of 0 (for the same allele) or 1 (for different alleles). For diploid strains such as *C. albicans*, each pair can have a distance of 0, 1 or 2. Equation 6.2 can also be used to calculate the mean distance d between all $N(N-1)/2$ possible pairs of strains, and the variance

of the distances can be written as

$$var_j = h_j(1 - h_j). \tag{6.3}$$

Similarly, generalizing for m loci, if D is defined as the distance between two strains over all loci, then the mean distance between two strains over all loci can be written as

$$\bar{D} = \sum_j h_j, \tag{6.4}$$

and the variance of the distances is

$$V_D = \sum_j var_j + 2\sum_j \sum_k cov_{j,k}, \tag{6.5}$$

where $cov_{j,k}$ is the covariance between the distance at locus j and the distance at locus k. All these covariances are expected to be zero if there is no association between loci. Therefore I_A is defined as

$$I_A = \frac{V_D}{\sum var_j} - 1, \tag{6.6}$$

which is zero if there is no association between loci.

The I_A between alleles of *YWP1*, *HWP1* and *EAP1* was determined using multilocus software designed by Agapow and Burt [100]. The software calculates I_A using Equation 6.6, where,

$$V_D = \frac{\sum D^2 - \frac{(\sum D)^2}{n_p}}{n_p}, \tag{6.7}$$

and,

$$var_j = \frac{\sum d^2 - \frac{(\sum d)^2}{n_p}}{n_p}. \tag{6.8}$$

In Equations 6.7 and 6.8, $n_p = N(N-1)/2$ is the number of possible pairs between N strains, D is the total distance between strains over all loci, and d is the distance at locus j.

The I_A measures to what extent individual strains which are the same at one locus are more likely than random to be the same at other loci. The significance of the I_A can be analysed by comparing the observed value to one expected by chance

(random data) [101]. In the calculation of I_A between alleles from *YWP1*, *HWP1* and *EAP1*, a random data set of 1000 replicates was used to test the significance. The significance of the I_A shown by value of $p < 0.05$ indicates that there is a non-random association between alleles from different genes. The non-significant value of $p > 0.05$ indicates that there may be an association between the alleles, but it could be due to a random process (random association), and for convenience, this is referred to as no (significant) association between the alleles.

The software also determine the value of r, as an alternative to I_A, which is dependent on the number of loci in the analysis [100]. However, since these studies mostly focus on the association between two genes, only I_A is calculated, i.e. for two genes, the values of I_A and r are similar. Later in this chapter, the value of r is shown together with I_A when analysing the association among three genes. With this, when comparing the association of three genes to the association of two gene, the value of r for three genes should be comparable to the value of I_A for two genes. The significance of the value of I_A and r is identical.

6.4.2 Linkage Disequilibrium Between Alleles of Different Genes

In this section, the quantification of the linkage disequilibrium between alleles of different genes is described. This includes the linkage disequilibrium between alleles of *YWP1-HWP1*, *YWP1-EAP1*, and *HWP1-EAP1*, and the linkage disequilibrium between alleles of *YWP1-HWP1-EAP1*. The linkage disequilibrium is quantified by the index of association I_A defined by Equation 6.6 (see page 195).

Index of Association of Alleles of *YWP1-HWP1*

The values of index of association I_A of alleles of *YWP1-HWP1* were determined to be 0.1 ($p = 0.13$) for GPG strains, 0.002 ($p = 0.49$) for non-GPG strains, and 0.41 ($p=0.008$) for (GPG) commensal strains. These results showed that there is no association between alleles of *YWP1-HWP1* for GPG and non-GPG strains. However, there is a non-random association between the alleles for (GPG) commensal strains.

Index of Association of Alleles of *YWP1-EAP1*

Since there are two repeat regions in *EAP1*, I_A between alleles from *YWP1-EAP1* of each repeat region was calculated separately as *YWP1-EAP1*(1) and *YWP1-EAP1*(2), for repeat regions 1 and 2 of *EAP1*, respectively.

The values of I_A of alleles of *YWP1-EAP1*(1) were determined to be 0.44 ($p < 0.001$) for GPG strains, 0.17 ($p < 0.001$) for non-GPG strains, and 0.55 ($p = 0.004$) for (GPG) commensal strains. The results showed that there is a non-random association between alleles of *YWP1-EAP1*(1) of all groups of strains, where the association in GPG strains is stronger than non-GPG strains, and the association in (GPG) commensal strains is stronger than (GPG) infection strains.

The values of I_A of alleles of *YWP1-EAP1*(2) were determined to be 0.42 ($p < 0.001$) for GPG strains, 0.11 ($p = 0.04$) for non-GPG strains, and 0.39 ($p = 0.01$) for commensal strains. The results showed that there is a non-random association between alleles of *YWP1-EAP1*(2) of all groups of strains, where the association in GPG strains is stronger than non-GPG strains, and the association in (GPG) infection strains is similar to (GPG) commensal strains.

Index of Association of Alleles of *HWP1-EAP1*

Since there are two repeat regions in *EAP1*, I_A between alleles of *HWP1-EAP1* of each repeat region was calculated separately as *HWP1-EAP1*(1) and *HWP1* and *EAP1*(2), for repeat regions 1 and 2 of *EAP1*, respectively.

The values of I_A of alleles of *HWP1-EAP1*(1) were determined to be 0.12 ($p = 0.10$) for GPG strains, 0.05 ($p = 0.08$) for non-GPG strains, and 0.31 ($p = 0.01$) for (GPG) commensal strains. The results indicate that there is no association between alleles of *HWP1-EAP1*(1) for GPG and non-GPG of infection strains. However, there is a non-random association between the alleles for (GPG) commensal strains.

The values of I_A of alleles of *HWP1-EAP1*(2) were determined to be 0.22 ($p = 0.01$) for GPG strains, 0.06 ($p = 0.10$) for non-GPG strains, and 0.34 ($p=0.02$) for (GPG) commensal strains. The results showed that there is a non-random association between alleles of *HWP1-EAP1*(2) of (GPG) infection and (GPG) commensal strains,

where the association in commensal strains is stronger than infection strains. There is no association between the alleles observed in non-GPG strains.

Index of Association of Alleles of Repeat Regions 1 and 2 of *EAP1*

Since there are two repeat regions in *EAP1*, I_A of *EAP1* alleles for these two repeat regions was determined to observe how these the alleles associate with each other. The values of I_A for repeat regions 1 and 2 of *EAP1* were determined to be 0.85 ($p < 0.001$) for GPG strains, 0.05 ($p = 0.10$) for non-GPG strains, and 0.61 ($p < 0.001$) for (GPG) commensal strains. The results showed that there is a non-random association between *EAP1* alleles for repeat regions 1 and 2 for (GPG) infection and (GPG) commensal strains, where the association in infection strains is stronger than commensal strains. There is no association between the alleles observed in non-GPG strains.

Index of Association of Alleles of *YWP1-HWP1-EAP1*

For the linkage disequilibrium of alleles of *YWP1-HWP1-EAP1*, beside I_A, r is also determined. It has been shown previously that the value of I_A is influenced by the number of loci involved in the calculation. Therefore, the I_A of three genes cannot be compared to that of two genes. However the value of r is independent of number of genes, and for 2 genes, the value of r and I_A is similar. The value of p for both r and I_A is the same.

Since there are two repeat regions in *EAP1*, I_A of alleles of *YWP1-HWP1-EAP1* of each repeat region is described separately as *YWP1-HWP1-EAP1*(1) and *YWP1-HWP1-EAP1*(2), for repeat regions 1 and 2 of *EAP1*, respectively.

The I_A values of alleles of *YWP1-HWP1-EAP1*(1) were determined to be 0.44 (r=0.22) ($p < 0.001$) for GPG strains, 0.11 (r=0.06) ($p = 0.04$) for non-GPG strains, and 0.86 ($r = 0.43$) ($p < 0.001$) for (GPG) commensal strains. The results showed that there is a non-random association between alleles of *YWP1-HWP1-EAP1*(1), where the association in GPG strains is stronger than non-GPG strains, and the association in (GPG) commensal strains is stronger than (GPG) infection strains.

The I_A values of alleles of *YWP1-HWP1-EAP1*(2) were determined to be 0.52 (r=0.27) ($p < 0.001$) for GPG strains, 0.05 (r=0.03) ($p = 0.01$) for non-GPG strains, and 0.63 (r=0.33) ($p = 0.002$) for (GPG) commensal strains. The results showed that there is a non-random association between alleles of *YWP1-HWP1-EAP1*(1), where the association in GPG strains is stronger than non-GPG strains, and the association in (GPG) commensal strains is stronger than (GPG) infection strains.

6.4.3 Summary of Linkage Disequilibrium Between Alleles of Different of Genes

The results of calculation of I_A and its significance for all pairs of genes and all groups of strains are summarized in Table 6.2, and presented graphically in Figure 6.23. The results for the three genes are summarized in Table 6.3.

Table 6.2: The index of association between alleles of different genes for GPG, non-GPG, and commensal strains. The significance value p is shown in parenthesis next to I_A.

Pair of genes	Index of Association I_A (p value)		
	GPG strains	Non-GPG strains	Commensal strains
YWP1-HWP1	0.1 (0.13)	0.002 (0.49)	0.41 (0.008)
YWP1-EAP1(1)	0.44 (<0.001)	0.17 (<0.001)	0.55 (0.004)
YWP1-EAP1(2)	0.42 (<0.001)	0.11 (0.04)	0.39 (0.01)
HWP1-EAP1(1)	0.12 (0.1)	0.05 (0.08)	0.31 (0.01)
HWP1-EAP1(2)	0.22 (0.01)	0.06 (0.12)	0.34 (0.02)
EAP1(1)-*EAP1*(2)	0.85 (<0.001)	0.05 (0.08)	0.61(<0.001)

Table 6.3: The index of association between alleles of different genes for GPG, non-GPG, and commensal strains. The significance value p is shown in parenthesis next to I_A.

Pair of genes	I_A, r (p value)		
	GPG strains	Non-GPG strains	Commensal strains
YWP1-HWP1-EAP1(1)	0.44, 0.23 (<0.001)	0.11, 0.06 (0.04)	0.86, 0.43 (<0.001)
YWP1-HWP1-EAP1(2)	0.52, 0.27 (<0.001)	0.06, 0.03 (0.20)	0.63, 0.33 (=0.002)

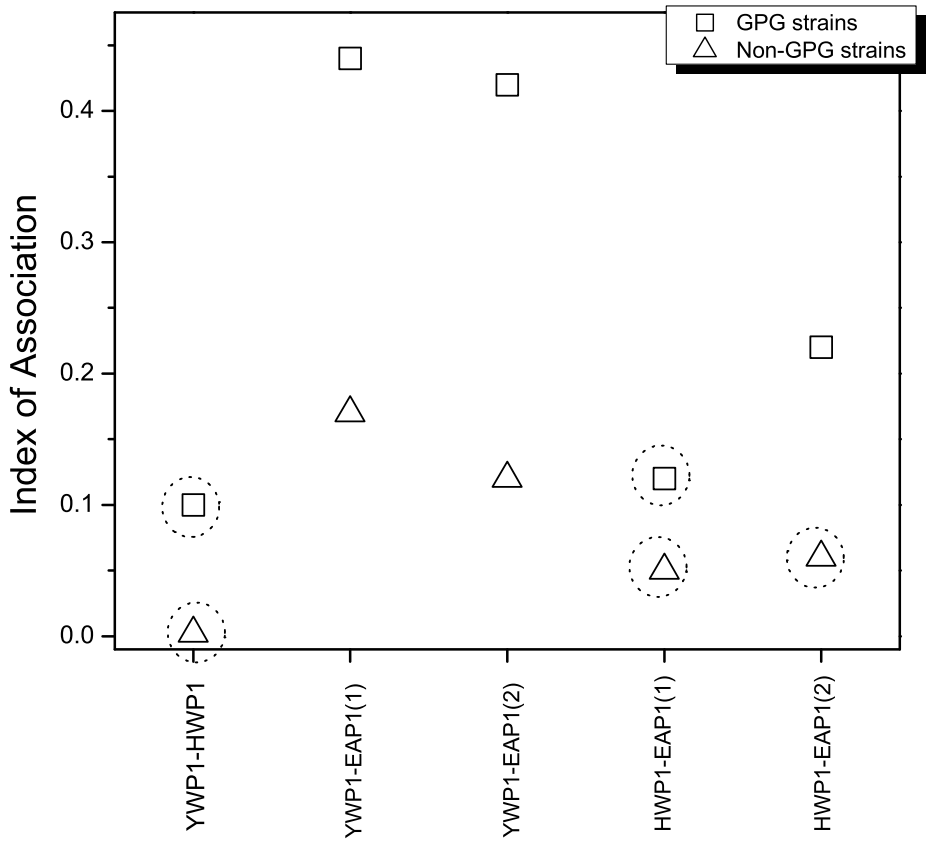

Figure 6.23: Index of association (I_A) between alleles of different genes for GPG and non-GPG strains. The circular dashed-lines indicate that I_A is not significant for the particular data, and is referred to as no association.

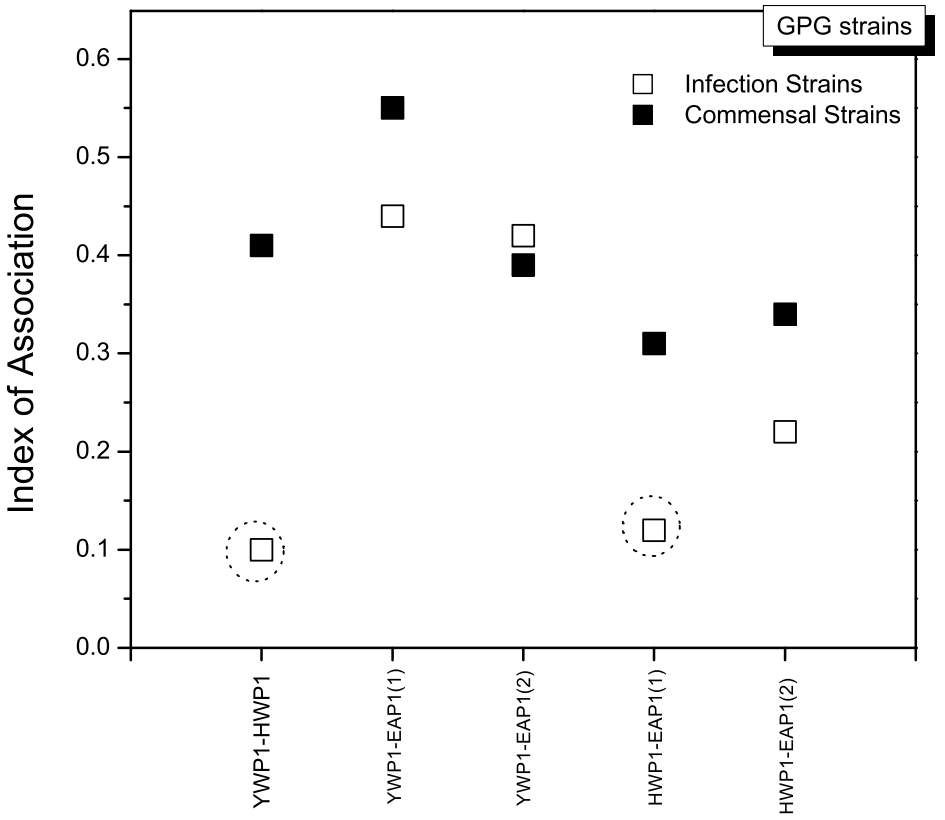

Figure 6.24: Index of association between alleles of different genes for commensal strains. The data for (GPG) infection strains was added for a comparison. The circular dashed-lines indicate that I_A is not significant for the particular data, and is referred to as no association.

6.5 Discussion

The aim of this chapter is to describe the gene interconnectedness between alleles of *YWP1*, *HWP1*, and *EAP1* for GPG, non-GPG, and commensal strains.

6.5.1 Distribution of genotypes of the pairs of genes

The distributions of genotypes of pairs of genes indicate that there were predominant genotypes observed in GPG strains for all pairs of genes (see Figures 6.1, 6.4, 6.7, 6.11, and 6.14, on pages 164, 168, 173, 177, and 181). However, the predominant genotypes in non-GPG strains were only observed for *YWP1-HWP1* and *HWP1-EAP1*(2) pairs (see Figures 6.2 and 6.15 on pages 165 and 182). Similarly, the distributions of genotypes of three genes, i.e. *YWP1-HWP1-EAP1* genotypes, also indicate that there were predominant genotypes observed in GPG strains, but no obvious predominant genotypes were observed in non-GPG strains (see Figures 6.18 and 6.20 on pages 187 and 190). The distributions of *YWP1-HWP1-EAP1* genotypes may be used to distinguish GPG strains from non-GPG strains, since most genotypes in non-GPG strains are different from genotypes in GPG strains. For example, only three *YWP1-HWP1-EAP1*(1) genotypes in non-GPG strains (indicated by three arrows in Figure 6.18 on page 187) match *YWP1-HWP1-EAP1*(1) genotypes in GPG strains, and only two *YWP1-HWP1-EAP1*(2) genotypes in non-GPG strains (indicated by two arrows in Figure 6.20 on page 190) match *YWP1-HWP1-EAP1*(2) genotypes in GPG strains. This is also case for the distribution of *YWP1-EAP1* and *HWP1-EAP1* genotypes, where both repeat regions in *EAP1* are included in the analysis (see Figures 6.10 on page 176 and 6.17 on page 184 for *YWP1-EAP1* and *HWP1-EAP1* genotypes, respectively).

The predominant genotypes of pairs of genes were formed by the predominant allele combinations of individual genes, and the percentage of the predominant genotypes is less than or equal to the percentage of the predominant allele combinations of individual genes. For example, *YWP1-EAP1*(1) genotypes containing 5-5+10-11 which predominated in GPG strains (59.2%) were formed by *YWP1* allele combinations

containing 5-5 and *EAP1*(1) allele combinations containing 10-11, where both allele combinations predominated in GPG strains (94% and 59.2% for *YWP1* and *EAP1*(1), respectively), i.e. the percentage of the predominant *YWP1-EAP1*(1) genotypes is equal to the percentage of the predominant *EAP1*(1) allele combinations. This implies that all GPG strains, which have predominant *EAP1*(1) allele combinations also have predominant *YWP1* allele combinations. Similar results were also observed for *YWP1-EAP1*(2) genotypes. The percentages of the predominant genotypes of pair of genes and the predominant allele combinations of individual genes for GPG strains are summarized in Figure 6.25.a. The figure also shows the cases, where the percentage of the predominant genotypes of pairs of genes is less than the percentage of individual genes.

Similar to the results for GPG strains, Figure 6.25.b shows the percentage of the predominant genotypes of pairs of genes and the predominant allele combinations of individual genes for commensal strains. The figure shows that the percentage of the predominant *YWP1-HWP1* genotypes is equal to the percentage of the predominant *HWP1* allele combinations, which implies that all commensal strains, which have predominant *HWP1* allele combinations, also have predominant *YWP1* allele combinations. The percentage of genotypes of other pairs of genes is less than the percentage of allele combinations of individual genes. Figure 6.25.c shows the percentage of the predominant genotypes of pairs of genes in infection and commensal strains. The figure shows that the percentage of the predominant *YWP1-HWP1* and *HWP1-EAP1* genotypes in commensal strains is larger than that in infection strains, while the percentage of *YWP1-EAP1* genotypes in commensal strains is smaller than that in infection strains. Figures 6.26.a and b show the the percentages of the predominant *YWP1-HWP1-EAP1* genotypes and the predominant allele combinations of individual genes for infection and commensal strains, while Figure 6.26.c shows a comparison of the percentages of the predominant genotypes between commensal and infection strains. The later figure shows that the percentage for commensal strains is larger than that for infection strains. This may indicate that the interaction between these three genes may be more important when *C albicans* is in a commensal state

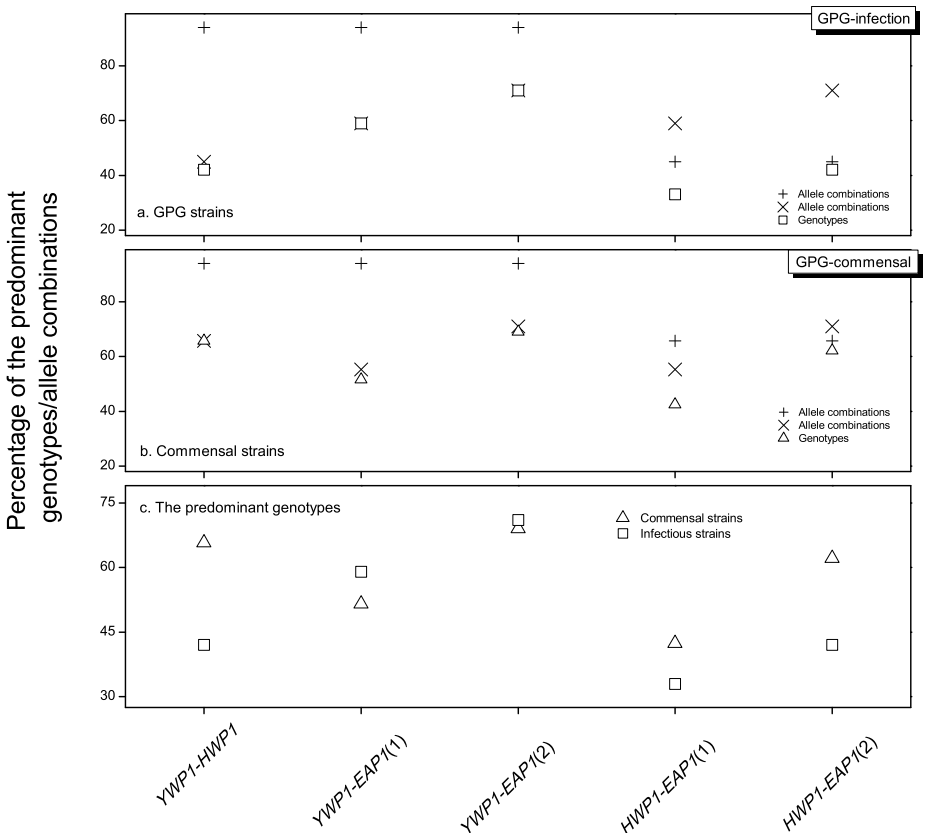

Figure 6.25: The percentages of the predominant genotypes of pairs of genes and the predominant allele combinations of individual genes for a. GPG strains, b. Commensal strains, and c. The percentage of the predominant genotypes of pairs of genes in (GPG) infection and commensal strains.

compared to when it is in a pathogenic state.

6.5.2 Genetic background influences the gene interconnectedness

The results of the I_A calculation for alleles of two from the three genes in GPG and non-GPG strains (see Figure 6.23 on page 200) showed that the interconnectedness between alleles of *YWP1-EAP1* is stronger than other pairs of genes. Considering both GPG and non-GPG strains, the non-random association exists only between alleles of *YWP1-EAP1*. For GPG strains, there is also a non-random association

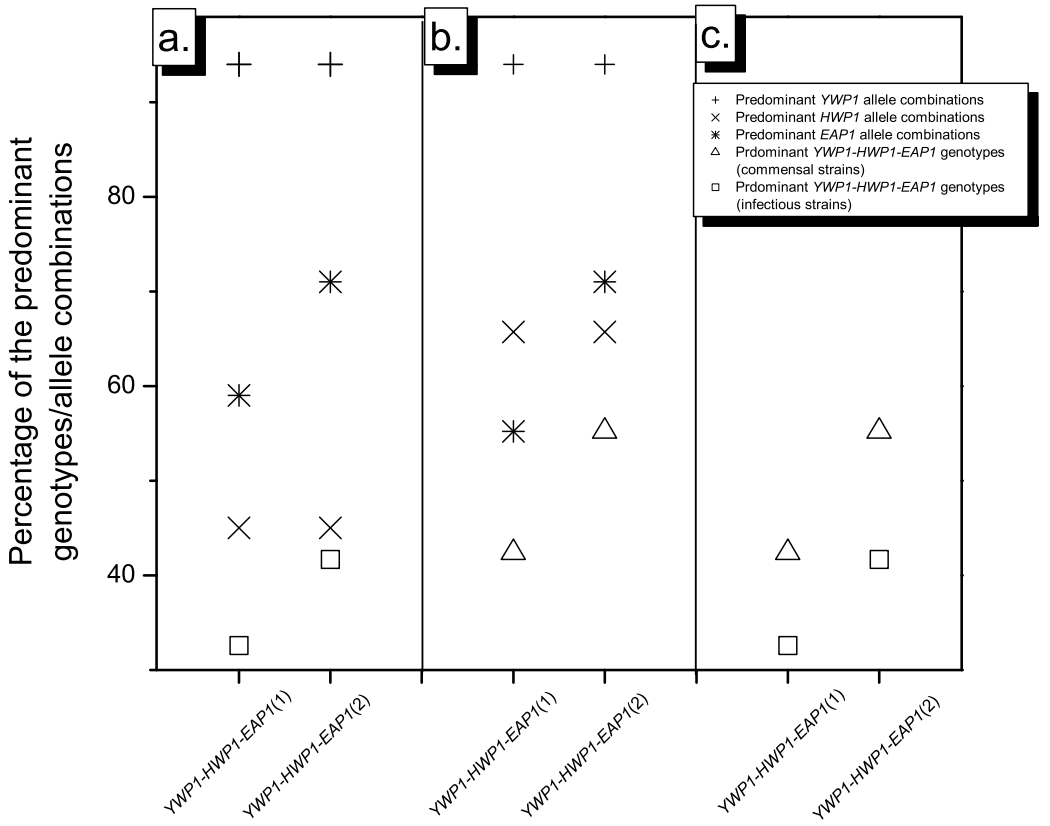

Figure 6.26: The percentages of the predominant genotypes of three genes and the predominant allele combinations of individual genes for a. GPG strains, b. Commensal strains, and c. The percentage of the predominant genotypes of three genes in (GPG) infection and commensal strains.

between alleles of *HWP1-EAP1*(2), but the value of I_A is lower than that of *YWP1-EAP1*. A similar pattern was observed for commensal strains, with an exception for *YWP1-EAP1*(2), where the I_A value is comparable to the value of the other pairs of genes (see Figure 6.24 on page 201); this exception will be discussed in the next section (6.5.3).

If *C. albicans* reproduces sexually, a strong interconnectedness between alleles of *YWP1-EAP1* may be related to the fact that these two genes are located on the same chromosome (chromosome 2); they are in a physical link one to another. Since the two genes are located on the same chromosome, the interconnectedness between the two genes may be caused only by their physical link not by selection. Non-random association of alleles from unlinked loci on organism that reproduces sexually has often been interpreted as evidence for selection [102]. However, it has been suggested that *C. albicans* reproduces clonally rather than sexually [103, 104, 105, 106]. Therefore there should be no effect on whether the genes are located on the same or different chromosomes.

The results presented in chapters 3 and 5 suggest that particular genetic backgrounds may have an impact on the selection of predominant allele combinations of *YWP1* and *EAP1*. For example, the results of the *YWP1* gene showed that GPG strains (branch A) have 5-5 as a predominant allele combination (see Figure 3.7 on page 41), while branch B non-GPG strains have 8-8 and branch C non-GPG strains have 5-8 as the predominant allele combinations (see Figure 3.18 on page 64). The results of this chapter show that the interconnectedness between alleles of these two genes is the strongest compared to other pairs of genes in GPG and non-GPG strains (see Figure 6.24 on page 201). This may indicate that the impact of the genetic background on the selection of the predominant allele combinations of the *YWP1* and *EAP1* genes may contribute to the strong interaction between these two genes. A genetic background has been suggested to have an impact on the selection of alleles, if fitness is influenced strongly by gene interaction, i.e. fitness of an allele at one gene is related to the interaction of this allele with alleles of other genes [107]. The predominant allele combination observed in non-GPG strains for *YWP1* and *EAP1*,

which is due to the influence of the genetic background, may indicate a possible interconnectedness between this gene and other genes.

Although the predominant allele combinations of *HWP1* were due to selection (chapter 4), while predominant allele combinations of *EAP1* (chapter 5) were influenced by a genetic background, the results of this chapter show that there was a non-random association of alleles of *HWP1-EAP1*(2) for GPG strains, and *HWP1-EAP1* for (GPG) commensal strains. This indicates that the genetic background does influence gene interconnectedness, although one of the linked genes is due to selection. However, the interconnectedness is not as strong as it would be if the two linked genes were influenced by the genetic background. It has been shown that for an asexual population, if there is a selective pressure on genes, the influence of the genetic background is diminished, and this weakens the gene interconnectedness [57].

A comparison of I_A values calculated for alleles of different genes in GPG and non-GPG strains also showed that the gene interconnectedness in GPG strains is systematically stronger than the gene interconnectedness in non-GPG strains. This result is expected, since the non-random association between genes measures a 'distance' defined as the number of loci at which two individual strains differ, similar to what the classification of GPG and non-GPG is based on, i.e. a genetic distance. Therefore, a similar genetic background relates to a strong gene interconnectedness, while a diverse genetic background will favor a weak interconnectedness between genes; GPG strains with a mean genetic distance between any two isolates of 0.19, showed stronger gene interconnectedness than non-GPG strains with a mean genetic distance between any two isolates of 0.37.

6.5.3 The gene interconnectedness distinguishes commensal strains from infection Strains

The I_A values calculated for alleles of two from the three genes in commensal and infection strains (see Figure 6.24 on page 201) show that the association in commensal strains is systematically stronger than the association in infection strains, except for

the association between alleles of *YWP1-EAP1*(2), where the I_A value is similar in commensal and infection strains. A similar result was also found for alleles of the three genes; the association between alleles of *YWP1-HWP1-EAP1* in commensal strains is stronger than the association in infection strains (see Table 6.3 on page 199).

The clear result that the non-random association between alleles of different genes is systematically stronger in commensal strains than in infection strains suggests that the interconnectedness between alleles of different genes distinguishes commensal strains from infection strains. From the results of the allele characterization of the *YWP1*, *HWP1*, and *EAP1* genes described in chapters 3 to 5, commensal and infection strains were found to have the same predominant allele combinations. A search to find distinct predominant allele combinations in commensal and infection strains, which would be expected if the gene repeat variability was used for adaptation purposes, failed. This leads to the conclusion that a change of the repeats of *YWP1*, *HWP1*, and *EAP1* is not required when *C. albicans* state changes from commensal to pathogenic. Similar results were found in the genotypes distributions of commensal and infection strains for pairs of genes and for the three genes described in this chapter, i.e. there was no significant difference in the distribution of the predominant genotypes in commensal and infection strains. Although the results from previous chapters for the allele combination of a single gene, and this chapter for genotypes of pairs of genes showed that there is no change in the number of repeats required when *C. albicans* state changes from commensal to pathogenic, one of the key findings in this chapter showed that there is a change in the degree of the interconnectedness between alleles of different genes.

The facts that the gene interconnectedness in commensal strains is systematically stronger than the gene interconnectedness in infection strains for any pair of genes (and among the three genes), except *YWP1* and *EAP1*(2), suggest that the interconnectedness between *YWP1*, *HWP1*, and *EAP1* is more important in a commensal state than it is in a pathogenic state. For example, when *C. albicans* changes from a commensal to a pathogenic state, the strong non-random association between *YWP1* and *HWP1* becomes insignificant. This implies that the interconnectedness between

YWP1 and *HWP1* is not important in pathogenicity of *C. albicans*. This may be explained by the known functions of the encoded protein. It has been found that HWP1p is involved in hyphal adhesion only on host epithelial cells, not on host endothelial cells [62, 108, 109], and YWP1p is involved in yeast dispersal in the host [60]. Since in a commensal state the *C. albicans* only interacts with host epithelial cells, the interconnectedness between these two genes may be important resulting in a significant association. In this case, YWP1p facilitates the yeast cells moving from one location to another, and HWP1p mediates in adhesion of the hyphae (produced by the yeast cells) to the host cells. In an infection state, the fungus interacts with both host epithelial and endothelial cells. Adhesion of the fungus to host endothelial cells is not mediated by HWP1p but rather by different proteins such as ALS3p [110]. This may explain why there is no significant association between *HWP1* and *YWP1* in infection strains. In an infection state, *HWP1* or *YWP1* may associate with other genes or each gene may function independently. A similar I_A value for the alleles of *YWP1-EAP1*(2) in both commensal and infection strains indicates that the interconnectedness between *YWP1* and *EAP1*(2) is as important in a commensal state as it is in a pathogenic state. Based on these results, it is reasonable to expect that there are some *C. albicans* genes which are interconnected more strongly in infection strains than in commensal strains. Although there are no such pairs of genes observed in this study, the interconnectedness between alleles of *EAP1* for repeat regions 1 and 2 is consistent with the expectation of the existence of such genes; the interconnectedness between these alleles is stronger in infection strains than in commensal strains (see Table 6.2 on page 199). These results imply that pairs of genes may be selected to interact with one another when the immune status of the host changes.

Chapter 7

An Attempt to Detect New Alleles

This chapter describes an attempt to find new alleles of the *YWP1*, *HWP1*, and *EAP1* genes. Any detection of new alleles would allow the mutation rate of each gene to be determined. For the purpose to generate new alleles, *in vitro* serial transfer for 300 generations was used. This experiment, therefore, was aimed to answer questions of whether *YWP1*, *HWP1*, and *EAP1*, are hypermutable genes. In addition to this experiment, the possible new alleles generated was also investigated in strains isolated from an AIDS patient with recurrent candidiasis.

7.1 Alleles of *YWP1*, *HWP1*, and *EAP1* in Strain RIHO30 from Serial Transfer

The methods to serial transfer are described in the materials and methods chapter (see page 27). Mutation rate is defined as a ratio between the number of mutated cells and the number of colonies tested multiplied by the generations. The mutation rate M_{rate} can be expressed mathematically as,

$$M_{rate} = \frac{\#\ of\ mutated\ cells}{\#\ of\ collonies\ tested \times generations}\ . \tag{7.1}$$

This experiment used 60 colonies from 300 generation samples for strain RIHO30. Primers YWP1BF and YWP1BR, HWP1F and HWP1R, and EAP1AF and EAP1AR (see Table 2.5 on page 24) were used to amplify the repeat regions of *YWP1* and

HWP1, and repeat region 1 of *EAP1*, respectively. Alleles of *YWP1*, *HWP1*, and *EAP1*(1) in strain RIHO30 at zero and 300 generations were characterized by length using genotyping.

Table 7.1: Results of the allele characterization of *YWP1*, *HWP1*, and *EAP1*(1) in strain RIHO30 at zero and 300 generations.

Generation	Allele combination, bp (# of repeats)		
	YWP1	*HWP1*	*EAP1*(1)
0	275-293 (5-8)	335-375 (2-3)	729-729 (7-7)
300	275-293 (5-8)	335-375 (2-3)	729-729 (7-7)

The results at zero generations showed that allele combinations of *YWP1*, *HWP1*, and *EAP1*(1) were (in bp) 275-293, 335-375, and 729-729, respectively. The same allele combinations for each gene were observed at 300 generations. The results of the characterization are summarized in Table 7.1. These results show that alleles are unchanged after 300 generations, thus the mutation rate cannot be determined for those genes.

7.2 Alleles of the *YWP1*, *HWP1*, and *EAP1* Genes in Strains from an AIDS Patient with Recurrent Candidiasis

The aim of this particular study is to determine whether there is a change of alleles of the *YWP1*, *HWP1*, and *EAP1* genes in strains isolated from an AIDS patient with recurrent candidiasis. For this purpose, six *C. albicans* isolates from an AIDS patient: 3207, 3208, 3209, 3210, 3211, and 3212, were characterized. There was an interval of 4 years between the isolation of isolates 3207 and 3212, and an interval of 28 months between isolates 3209 and 3210 [70]. Typing of the six sequential isolates used in this experiment (isolates 3207-3212) showed that they are from the same progenitor. The isolates were donated by Dr P.T. Magee's laboratory, University of Minnesota, USA.

Primers YWP1BF and YWP1BR, HWP1F and HWP1R, EAP1AF and EAP1AR, and EAP1BF and EAP1BR (see Table 2.5 on page 24) were used to amplify the repeat regions of *YWP1*, *HWP1*, and repeat regions 1 and 2 of *EAP1*, respectively. Alleles of the *YWP1*, *HWP1*, and *EAP1* genes from six isolates were characterized. For *YWP1*, allele combination containing 275 and 293 bp length (2 and 5 repeat units) were observed in all 6 isolates. For *HWP1*, allele combination containing 405 and 405 bp length (4 and 4 repeat units) were observed in all 6 isolates. For *EAP1*(1), allele combination containing 1123 and 1650 bp length (13 and 19 repeat units) were observed in all 6 isolates. For *EAP1*(2), allele combination containing 2800 and 2800 bp length (147 and 147 repeat units) were observed in all 6 isolates. The results of the characterization of the *YWP1*, *HWP1*, and *EAP1* genes for the six isolates are presented graphically in Figure 7.1. The results show that there is no change in *YWP1*, *HWP1*, and *EAP1* alleles over a four-year period.

7.3 Discussion

7.3.1 Prediction of mutation rate of *YWP1*, *HWP1*, and *EAP1*

Since alleles of *YWP1*, *HWP1*, and *EAP1* in strain RIHO30 were found to be unchanged after 300 generations, the mutation rate cannot be determined for those genes. The same '300 generation sample' used for this study, had already been used to detect new alleles of *SSR1*, and from 60 colonies tested, two new alleles were observed; the mutation rate of *SSR1* was determined to be 1.1×10^{-4} per cell division [52]. Since the same '300 generation sample' was used in both studies, it can be inferred that the mutation rate of *YWP1*, *HWP1*, and *EAP1* is smaller than that of *SSR1*. Indeed, it is less than 5.5×10^{-5} per cell division, the rate expected if only one new allele was observed in 60 colonies. The mutation rate of *PNG2* was determined with the same method to be 2.8×10^{-5} per cell division [51].

Although the mutation rate of *YWP1*, *HWP1*, and *EAP1* could not be determined

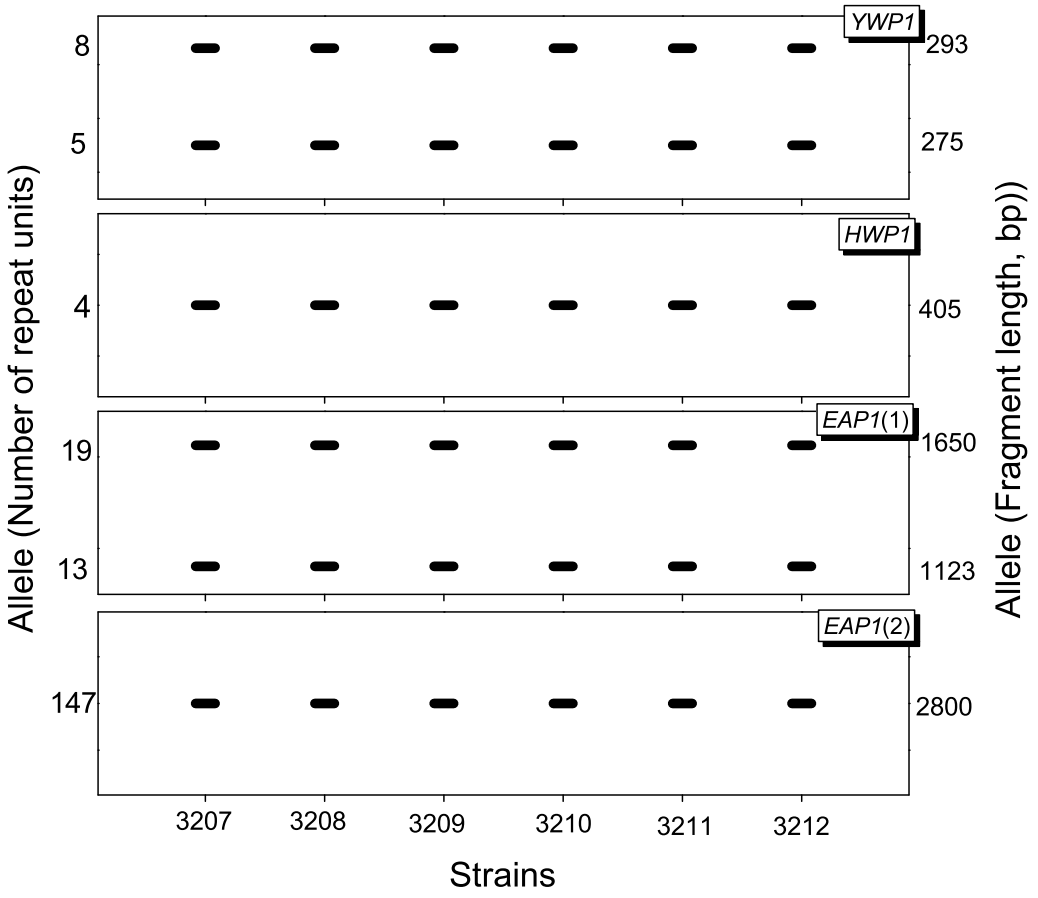

Figure 7.1: Results of the allele characterization of *YWP1*, *HWP1*, and *EAP1* for isolates from an AIDS patient with recurrent candidiasis. Interval time between strains 3207 and 3212 is 4 years.

in this *in vitro* experiment, it can be predicted using the VARScore algorithm [42]. The VARScore algorithm calculates a VARScore, and based on the experimental results, a relation between the VARScore and the mutation rate was established. The VARScore of a repeat region depends on the number of repeats, the length of the repeat unit, and the purity of the repeats. The VARScore for *YWP1*, *HWP1*, and *EAP1* was determined using SERV software available at www.hulsweb1.cgr.harvard.edu/TandemRepeat/, from which the mutation rate for these three genes was predicted. The values of the VARScore and the predicted mutation rate for *YWP1*, *HWP1*, and *EAP1* are presented graphically in Figure 7.2. For a comparison, the figure also included the predicted and observed mutation rates for *PNG2* [51] and *SSR1* [52]. The figure shows that the predicted mutation rates of the three genes are below 5.5×10^{-5} per cell division, which is a minimum rate that can be resolved by the measurement used. According to Legendre et al. [42], the VARScore algorithm can be applied to *C. albicans* genes. This could explain why new alleles of *YWP1*, *HWP1*, and *EAP1* could not be detected in this experiment. The VARSCore predicted mutation rates for *YWP1*, *HWP1*, and *EAP1* appear reasonable considering allele diversity of those three genes. *EAP1* with high allele diversity is predicted to have relatively higher mutation rate than *YWP1* and *HWP1* with less allele diversity. However, the reasonable VARScore predicted mutation rate for these three genes was found not to be the case for *PNG2* and *SSR1*, as the observed mutation rates for *PNG2* and *SSR1* are approximately 10 and 100 times, respectively, larger than the predicted values by the VARScore algorithm.

Based on the predicted VARScore mutation rates for *YWP1*, *HWP1*, and *EAP1*, and using the standard of hypermutable genes ranging from 1×10^{-5} to 1×10^{-2} per cell division [55], only *EAP1* can be categorized as a hypermutable gene.

A failure to detect new alleles of the *C.albicans* genes in *in vitro* experiments has also been reported in previous studies [47, 33, 111, 112, 113]. The genes include: *ALS1* [111], *EF3* [47], *ERK1* [113], *ALS3*, *ALS5*, and *ALS6* [33], *ALS3* and *ALS51* [112]. The results for *YWP1*, *HWP1*, and *EAP1* indicate that these genes can be added to the list of genes with relatively low mutation rate (no new alleles were observed

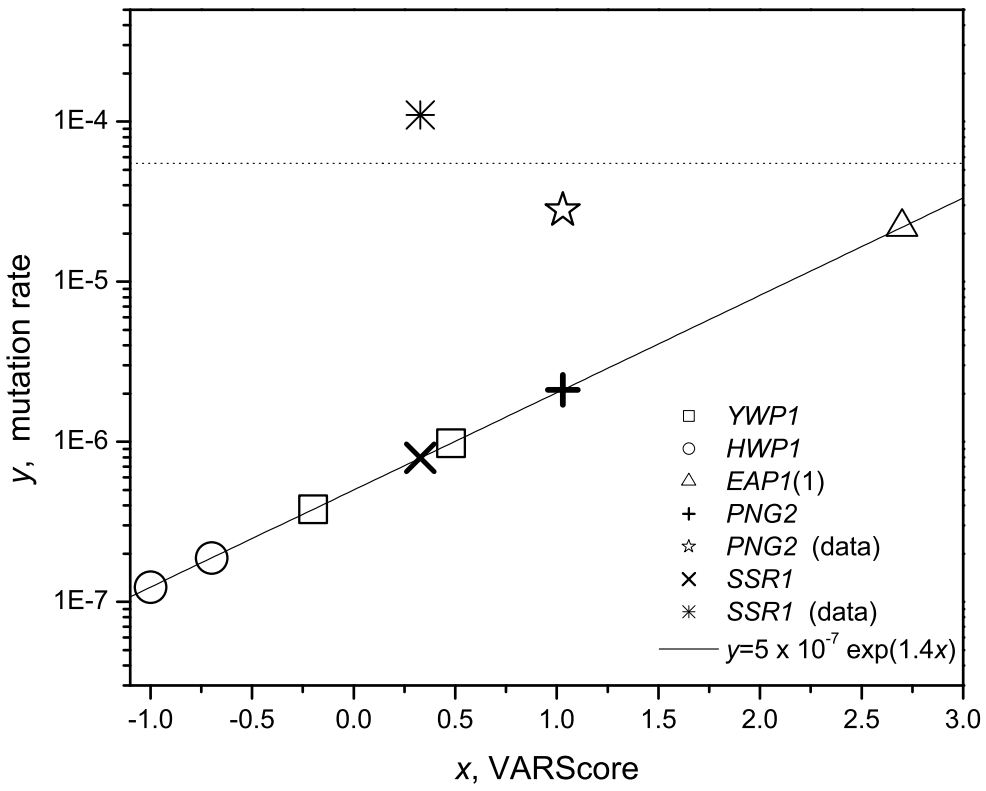

Figure 7.2: Prediction of the mutation rate for *YWP1*, *HWP1*, and *EAP1* in strain RIHO30 based on the VARScore calculation as described by Legendre et al. [42]. The mutation rate predictions for *PNG2* and *SSR1* and the measured mutation rates [51, 52] were included for comparison. The line representing $y = 5 \times 10^{-7} \exp(1.4x)$ shows the relation between the mutation rate y and the VARScore x. The horizontal dotted line denotes a minimum rate the methods used in these studies can resolve. Two *YWP1* and *HWP1* in RIHO30 are shown indicating the heterozygosity of the alleles.

in these genes). Although the mutation rates of these genes must be relatively low, there was a variability in the number of repeat units observed for each gene. There were 24 different alleles observed for *ALS1* [31], eight different alleles for *CEF3* [47], 16 and 12 different alleles for *CEF3* and *ERK1*, respectively [31], 14 different alleles for *ALS3* [27], nine different alleles for *ALS5* [33], and seven different alleles for *ALS6* [33]. In addition, the numbers of different alleles observed for *YWP1*, *HWP1*, and *EAP1*(1) were 8, 5, and 14, respectively.

7.3.2 A contingency gene has a low mutation rate

The *HWP1* gene in RIHO30 has been shown to have a low mutation rate compared to the experimental measured rates for *PNG2* and *SSR1* [51, 52]. In addition, the VARScore algorithm predicted that *HWP1* has lower mutation rate compared to *PNG2* and *SSR1* (see Figure 7.2 on page 215). Since the mutation rate tends to increase when the number of repeats increases, and there are 3 repeats of *HWP1* in RIHO30, there is a possibility that the mutation rate of *HWP1* will be larger when the number of repeats increases. However, although the predicted mutation rate for 6 repeats (the maximum repeats observed) is larger than that for 3 repeats, it is still less than 1×10^{-6} per cell division, lower than both the predicted and measured rates for *PNG2* and *SSR1*.

The experimental results show that strains with different genetic backgrounds select the same predominant alleles of *HWP1*. This indicates that *HWP1* has a role in adaptation through the variation of its repeats, i.e. *HWP1* functions as a contingency gene. In contrast, the experimental results show that strains with different genetic backgrounds select distinct predominant alleles of *PNG2* and *SSR1*, i.e. *PNG2* and *SSR1* do not function as contingency genes. Since *HWP1* has been suggested to have a low mutation rate, while *PNG2* and *SSR1* have been observed to have high mutation rates, i.e. *PNG2* and *SSR1* are hypermutable genes, it can be inferred that a hypermutable gene does not necessarily function as a contingency gene, and a contingency gene is not necessarily hypermutable. This is reasonable, since a high mutation rate facilitates adaptation when the cells are exposed to a condition to

which they are poorly adapted by producing diverse alleles. If the cells are well adapted, most of the diverse alleles resulting from a high mutation rate are likely to be deleterious due to their low fitness, i.e. once the cells became adapted, the advantages of high mutation rates are lost. This suggests that the well adapted cells tend to have low mutation rates [77]. This study shows that *HWP1* of *C. albicans* acts as a contingency gene, and has a low mutation rate. It is a common belief that a high mutation rate benefits to the functionality of a contingency gene [55]. However, the results of *HWP1*, *PNG2*, and *SSR1* suggest that this may not be the case for *C. albicans* genes, with an exception for *SSR1* when considering an adaptation when *C. albicans* changes from commensal to a pathogenic state; it has been claimed that *SSR1*, a hypermutable gene, acts as a contingency gene when *C. albicans* changes from commensal to a pathogenic state [52].

The VARScore algorithm predicts that the mutation rate of *EAP1*(1) is higher than that of *YWP1*, and the mutation rate of *YWP1* is slightly higher than that of *HWP1* (see Figure 7.2 on page 215 for RIHO30). In addition to this, the prediction for the mutation rate of these three genes based on the range of alleles observed (from the minimum to the maximum number of repeats) is consistent. The VARScore predicted mutation rates range from 3.4×10^{-7} to 2.3×10^{-6} per cell division for *YWP1*, 1.2×10^{-7} to 6.1×10^{-7} per cell division for *HWP1*, and 3.1×10^{-6} to 2.7×10^{-4} per cell division for *EAP1*(1); the number of repeats varies from 5 to 14, 2 to 6, and 2 to 15, for *YWP1*, *HWP1*, and *EAP1*(1), respectively (see Figure 7.3 on page 218). This is in agreement with a hypothesis that a well adapted gene tends to have low mutation rates [77], which is the case for *HWP1*. Using this thought, it can be inferred, that *YWP1* with mutation rate predicted to be relatively low (slightly higher than that of *HWP1*), potentially is a contingency gene, and may be at a state closer to the well adapted gene as *HWP1*. This is supported by the results described in Chapter 3, where *YWP1* allele combinations of 5-5 repeat units predominated in GPG strains, while 5-8 predominated in non-GPG strains, i.e. the two different genetic background strains select *YWP1* alleles of 5 repeat units. In the final state, alleles 8 repeat units of non-GPG strains may be eliminated leaving the predominant allele combinations of

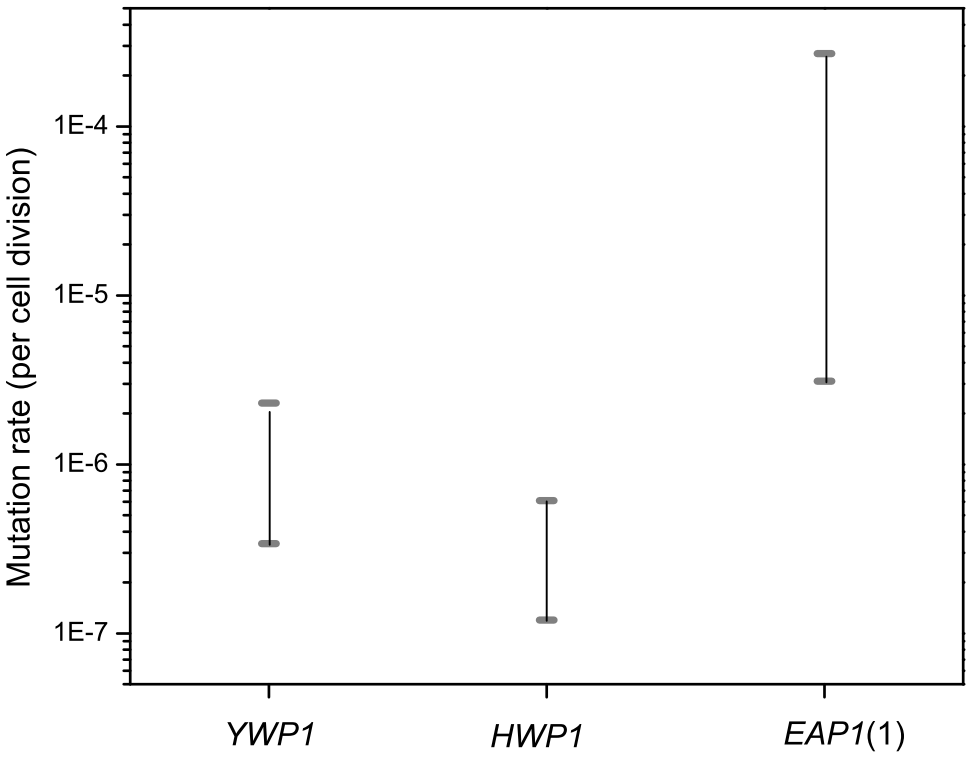

Figure 7.3: The range of the VARScore predicted mutation rates for *YWP1*, *HWP1*, and *EAP1* in strain RIHO30 based on the range of the number of repeats observed. The number of repeats observed varies from 5 to 14, 2 to 6, and 2 to 15, for *YWP1*, *HWP1*, and *EAP1*(1), respectively.

GPG and non-GPG strains the same. This scenario assumed that alleles of non-GPG strains adapt to follow those of GPG strains; it has been indicated that GPG strains are more virulent than non-GPG strains in young candidemia patients [80], GPG strains are 10 to 100 times more likely to cause disease than non-GPG strains [56], and most GPG strains are identical to clade 1 [78, 79], which has been proposed to be better adapted as humans epithelial colonizer and invader [79] compared to other clades in *C. albicans*.

7.3.3 No change in *YWP1*, *HWP1*, and *EAP1* alleles occurred during recurrent candidiasis over a four-year period

Alleles of *YWP1*, *HWP1*, and *EAP1* in six isolates from an AIDS patient with recurrent candidiasis were unchanged over a four-year period. There are three possible scenarios which can occur for these results : 1) each recurrent candidiasis is caused by the same strain, 2) although the same strain is responsible for each recurrent candidiasis, there is minor variation in genotype, i.e the strain undergoes microevolution, or 3) each recurrent candidiasis is caused by a different strain, i.e. there is strain replacement [114]. However, since the six isolates used in this experiment (strains 3207-3212) have been found to come from the same progenitor [70], scenario 3 can be ruled out. It has been suggested that a person commonly carries only one *C. albicans* strain [115]. If the strain undergoes microevolution that alters these genes (scenario 2), there would be variants selected for the three genes after drug treatment in each successive infection, resulting in different alleles of the genes for each infection. The result of this chapter seems to fit to scenario 1, where there was no microevolution and the same strain was responsible in each recurrent candidiasis. The results indicate that *YWP1*, *HWP1*, and *EAP1* may not respond to antifungal therapeutics.

The results are consistent with the results of other studies on *CAI* microsatellite, a non-coding locus of *C. albicans*; from 8 patients with recurrent candidiasis, 6 patient showed that the same strain was responsible for each recurrent candidiasis [116].

However, the same studies on *CAI* microsatellite also showed one case, where the same strain was responsible for the recurrent candidiasis, but the strain underwent microevolution (scenario 2), and one case where a different strain was responsible for each recurrent candidiasis (scenario 3).

Chapter 8

Summary and Conclusion

8.1 Summary

Candida albicans is a predominant causative agent of fungal infections in humans. It is an opportunistic pathogen that resides on the mucosal surfaces of the gastrointestinal tract, genitourinary tract, and oral sites of humans as harmless commensals. However, it can cause infection in humans when there is attenuation of normal host immune system i.e. in immunocompromised persons because of AIDS and organ transplantation. In immunocompromised persons, it can cause infection in almost all parts of the body, which indicates its ability to adapt well to different environmental conditions. The ability of *C. albicans* to alter its cell wall proteome in different conditions has been suggested to play important roles in the virulence of this fungus, and enable this fungus to adapt well to different host conditions [11]. The ability to have variation in the fungal cell surface in different conditions can help the fungus both to colonize different sites of the humans body and to evade the host immune system. Changes of the characteristics of individual cell surface protein through the variation in the number of repeats within the coding sequences is one of the processes that can result in cell surface variation. Variation in the repeats also can affect the exposure of the functional domain into extracellular space where it can carry its function. Approximately 115 genes in *C. albicans* have been predicted to encode cell surface proteins [19], and 50 % of the genes have repeat sequences [11].

Studies of *C. albicans* populations using DNA fingerprinting with the Ca3 probe indicate the existence of a strain group called general-purpose genotypes (GPG) [56]. This group was found to cause disease 10 to 100 times more often than a non-GPG strain group. GPG strains have been suggested to be better adapted in colonizing and causing disease on humans compared to non-GPG strains, and are related to clade 1 of 5 clades classified by MLST typing [78, 79]. Clade 1 was found to be more often associated with both epithelial colonization and infection than other clades [79]. Study on the genome of GPG and non-GPG strains using AFLP indicates that polymorphism on repeat-ORF has been found to be over-represented in GPG strains and is suggested to be important for the success of GPG strains [78].

The aim of this study was to identify genes which encode cell surface proteins, and are responsible for the adaptation process of *C. albicans* in humans by changing the number of repeats within the coding sequences. These genes are frequently called contingency genes, and have been identified in bacteria and fungi that cause disease in humans [44, 45]. In this study, a contingency gene is defined as a gene which has a role in adaptation by changing the number of repeat units within the coding sequences. *YWP1*, *HWP1*, and *EAP1* were chosen as potential contingency genes, since these genes encode cell surface proteins [13, 17, 25, 58] and contain repeats [59]. To find out whether *YWP1*, *HWP1*, and *EAP1* are contingency genes, their allelic distributions in two groups of strains with different genetic background, GPG and non-GPG strains, were examined. In addition, the allelic distributions in commensal and infection strains, and in strains isolated from different sites of the humans body were also examined.

8.1.1 Variability in the number repeats in *YWP1*, *HWP1*, and *EAP1*

This study reveals that there is a variability in the number of repeats in *YWP1*, *HWP1*, and *EAP1* observed from different strains. The number of different alleles and allele combinations of *YWP1*, *HWP1*, and *EAP1* is summarized in Table 8.1.

The number of different alleles and allele combinations of *HWP1* in all strains tested is the least, and the number of different alleles and allele combinations of *EAP1* is the highest.

Table 8.1: Summary of the number of different alleles and allele combinations of *YWP1*, *HWP1*, and *EAP1*.

	Group	*YWP1*	*HWP1*	*EAP1*(1)	*EAP1*(2)
Number of different alleles	GPG strains	3	4	8	8
	Non-GPG strains	8	5	14	22
	Commensal strains	3	5	7	6
	All strains	**8**	**5**	**15**	**26**
Number of different allele combinations	GPG strains	4	8	12	8
	Non-GPG strains	12	8	28	22
	Commensal strains	2	7	8	7
	All strains	**13**	**11**	**38**	**29**

8.1.2 Allelic distributions of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains

In this study, allelic distributions of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains were examined to determine whether the genes have a role in adaptation by changing the number of repeats in the coding sequences, i.e. whether *YWP1*, *HWP1*, and *EAP1* are contingency genes.

Allelic distributions of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains indicate that there are predominant alleles and allele combinations in both groups, with the exception that no predominant *EAP1* allele combinations were observed in non-GPG strains. The percentage of the predominant allele combinations in GPG strains is larger than that in non-GPG strains. Table 8.2 contains the summary of the predominant alleles and allele combinations of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains and the percentage of strains which have the predominant allele combinations.

Allelic distributions of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains are unique for each gene in terms of whether there is an impact of either the genetic

Table 8.2: The predominant alleles and allele combinations of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains. The percentages of the predominant alleles and allele combinations are shown in the parenthesis next to the predominant alleles and allele combinations.

Genes	Predominant alleles / allele combinations (%)			
	alleles		allele combinations	
	GPG	Non-GPG	GPG	Non-GPG
YWP1	5(95)	5(38), 8(38)	5-5 (94)	5-8 (38)
HWP1	4(38), 5(51)	4(40), 5(48)	4-5 (45)	4-5 (34)
EAP1(1)	10 (44), 11(36)	8(21), 11(15)	10-11 (59)	-
EAP1(2)	33(44), 48(35)	83(23), 102(27)	33-48 (71)	83-83(20),102-102(22)

background or the selective pressure on the genes. First, the predominant alleles and allele combinations of *HWP1* in GPG and non-GPG strains are the same, which show that there is a strong selective pressure on *HWP1* resulting in the selection of the same alleles by different genetic background strains. Second, the predominant allele combinations of *YWP1* in GPG and non-GPG strains are different, which shows an impact of the genetic background. However, GPG and non-GPG strains share the same predominant alleles, alleles containing 5 repeat units, which show that there is also a selective pressure on the gene. Therefore, both the genetic background and selective pressure influence the allelic distributions of *YWP1*. Third, there are predominant allele combinations of *EAP1* for repeat region 1 observed in GPG strains, but no obvious predominant allele combinations were observed in non-GPG strains. For repeat region 2, the predominant allele combinations of *EAP1* in GPG and non-GPG strains are different. Therefore, there is no indication of selective pressure on *EAP1*, and allelic distributions of *EAP1* are influenced by the genetic background.

Based on the genetic background of the strains tested and the allelic distributions of the genes observed, *YWP1* and *HWP1* can be categorized as contingency genes, while *EAP1* cannot be categorized as a contingency gene.

8.1.3 Allelic distributions of *YWP1*, *HWP1*, and *EAP1* in commensal and infection strains

In this study, allelic distributions of *YWP1*, *HWP1*, and *EAP1* in commensal and infection strains were examined to determine whether the genes act as contingency genes when *C. albicans* state changes from commensal to pathogenic.

Comparison of the allelic distributions of *YWP1*, *HWP1*, and *EAP1* in commensal and infection strains showed that commensal and infection strains have the same predominant alleles and allele combinations. This leads to the conclusion that the number of *YWP1*, *HWP1*, and *EAP1* repeats does not alter when *C. albicans* state changes from commensal to pathogenic. The predominant alleles and allele combinations of *YWP1*, *HWP1*, and *EAP1* in commensal and infection strains and the percentage of the strains having the predominant alleles and allele combinations are summarized in Table 8.3.

Similar results were observed in the genotype distributions of pairs of genes and all three genes as a single set in commensal and infection strains. There was no significant difference in the distribution of the predominant genotypes of *YWP1-HWP1*, *YWP1-EAP1*, *HWP1-EAP1*, and *YWP1-HWP1-EAP1* in commensal and infection strains.

Table 8.3: The predominant alleles and allele combinations of *YWP1*, *HWP1*, and *EAP1* in commensal and infection strains. The percentages of the predominant alleles and allele combinations are shown in the parenthesis next to the predominant alleles and allele combinations.

Genes	Predominant alleles / allele combinations (%)			
	alleles		allele combinations	
	Commensal	Infection	Commensal	Infection
YWP1	5(94)	5(95)	5-5 (94)	5-5 (94)
HWP1	4(39), 5(41)	4(40), 5(48)	4-5 (66)	4-5 (45)
EAP1(1)	10(43), 11(47)	10 (44), 11(36)	10-11 (55)	10-11 (59)
EAP1(2)	33(45), 48(44)	33(44), 48(35)	33-48 (71)	33-48 (71)

8.1.4 Allelic distributions of *YWP1*, *HWP1*, and *EAP1* in strains isolated from different sites of the humans body

In this study, allelic distributions of *YWP1*, *HWP1*, and *EAP1* in strains isolated from different sites of the humans body were examined to determine whether the genes act as contingency genes when *C. albicans* moves to particular sites of the humans body. The results of this study show no significant difference in the distribution of alleles and allele combinations of *YWP1* and *EAP1* in strains isolated from different sites of the humans body. Although there is no significant difference in the distribution of *HWP1* alleles observed in strains isolated from different sites of the humans body, there is a significant difference in the *HWP1* allele combinations. This leads to the conclusion that *YWP1* and *EAP1* do not act as contingency genes, but *HWP1* may act as a contingency gene, when *C. albicans* moves to particular sites of the humans body.

8.1.5 Are *YWP1*, *HWP1*, and *EAP1* contingency genes ?

Based on the allelic distributions of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains, or in commensal and infection strains, or in strains isolated from different sites of the humans body, it can be inferred that *YWP1* and *HWP1* can be categorized as contingency genes, while *EAP1* cannot be categorized as a contingency gene (see Table 8.4 for the summary). The result shows that although there is a variability in the number of repeats in *YWP1*, *HWP1*, and *EAP1* observed from different strains, the variability was used for adaptation purposes only in *YWP1* and *HWP1*.

Table 8.4: The summary of whether there is any indication that *YWP1*, *HWP1*, and *EAP1* are involved in adaptation, thus contingency genes.

Genes	Adaptation (contingency gene)		
	Genetic background	Immune status	Particular sites
YWP1	YES	NO	NO
HWP1	YES	NO	YES
EAP1	NO	NO	NO

8.1.6 The gene interconnectedness distinguishes commensal from infection strains

Although commensal and infection strains cannot be distinguished by the allelic distribution of *YWP1*, *HWP1*, and *EAP1*, the gene interconnectedness does distinguish these two groups. The gene interconnectedness for commensal strains is systematically stronger than the gene interconnectedness for infection strains for any pair of genes and for all three genes as a single set, with the exception of the *YWP1-EAP1*(2) pair. These results suggest that the interconnectedness between *YWP1*, *HWP1*, and *EAP1* is more important in a commensal state than it is in a pathogenic state. These results imply that there may be pairs of genes selected to interact one with another when the immune status of the host changes.

Beside the gene interconnectedness, the percentage of the predominant *YWP1-HWP1-EAP1* genotypes in commensal strains is larger compared to infection strains. This also indicates that the interaction between these three genes may be more important when *C albicans* is in a commensal state than when it is in a pathogenic state.

8.1.7 Are *YWP1*, *HWP1*, and *EAP1* hypermutable genes ?

Bacterial genes with mutation rates in the range 10^{-5} to 10^{-2} per cell division have been categorized as hypermutable gene [55], and the same range of mutation rates was applied to the *C. albicans* genes characterized in this study. Since no mutated alleles of *YWP1*, *HWP1*, and *EAP1* were observed in RIHO30 strain used in these studies, the gene mutation rate could not be determined. Given the methodology used, it can be inferred that the mutation rate of the three genes is less than 5.5×10^{-5} per cell division. In addition, using the VARScore algorithm, the mutation rate of the *YWP1* and *HWP1* is predicted to be lower than 1×10^{-6} per cell division, and of the *EAP1* to be 2.2×10^{-5} per cell division. Both the inference from the observation and the prediction from the VARScore algorithm suggest that *YWP1* and *HWP1* may not be categorized as hypermutable genes, while *EAP1* may be categorized as a

hypermutable gene. The fact that the mutated alleles of *EAP1* could not be observed may be due to the nature of the experiment designed. The prediction that *EAP1* is a hypermutable gene, while *YWP1* and *HWP1* are not hypermutable genes is consistent with the fact that the numbers of different alleles of *EAP1* is the most compared to *YWP1*, *HWP1*.

8.2 Conclusions and Future Directions

8.2.1 Conclusions

1. Based on the allelic distributions of *YWP1*, *HWP1*, and *EAP1* in GPG and non-GPG strains, *YWP1* and *HWP1* can be categorized as contingency genes, while *EAP1* cannot be categorized as a contingency gene. For *YWP1*, although GPG and non-GPG strains select distinct predominant allele combinations, they share common predominant alleles, which are optimal alleles for both groups. For *HWP1*, GPG and non-GPG strains select the same predominant alleles and allele combinations, which indicates that these predominant alleles are important alleles for both groups. For *EAP1*, allelic distributions in GPG strains and non-GPG strains are significantly different.

2. Based on the allelic distributions of *YWP1*, *HWP1*, and *EAP1* in commensal and infection strains, it can be inferred that *YWP1*, *HWP1*, and *EAP1* do not act as contingency genes when *C. albicans* state changes from commensal to pathogenic. Although the allelic distributions of the genes cannot distinguish commensal from infection strains, the gene interconnectedness does distinguish these two groups, i.e. the *YWP1-HWP1-EAP1* interconnectedness is stronger in commensal strains than it is in infection strains.

3. Based on the allelic distributions of *YWP1*, *HWP1*, and *EAP1* in strains isolated from different sites of he humans body, it can be inferred that *YWP1* and *EAP1* do not act as contingency genes, but *HWP1* may act as a contingency gene, when *C. albicans* moves to particular sites of the humans body.

4. The results of this study cannot determine whether *YWP1*, *HWP1*, and *EAP1*

are hypermutable genes, since there were no new alleles observed. However, from the prediction of the VARScore algorithm, *EAP1* may be categorized as a hypermutable gene, while *YWP1* and *HWP1* may not be categorized as hypermutable genes. Based on the VARScore predicted mutation rates, it can be inferred that the contingency genes tend to have low mutation rates, while hypermutable genes may not necessarily act as contingency genes.

8.2.2 Future Directions

Using 454 sequencing to determine the mutation rate of *YWP1*, *HWP1* and *EAP1*

In these studies, the mutation rate of *YWP1*, *HWP1*, and *EAP1* could not be determined, since no new alleles were observed. This may be due to the fact that the mutation rate of these three genes is less than the minimum rate that the experimental design can detect. This is consistent with the low mutation rates predicted by the VARScore algorithm for these three genes. One method for measuring the mutation rates of these genes would be to using 454 sequencing [117]. 454 sequencing can generate approximately a million reads per sample, which obviously can cover all DNA templates present in a sample. In 454 sequencing the average read length is 450 to 500 bp, which is longer than the repeat regions of *YWP1* and *HWP1* for all strains tested in these studies. Since the *EAP1* repeat regions for all strains is longer than 450 bp, the 454 sequencing is not feasible. However, for *EAP1*, the predicted mutation rate is larger than those for *YWP1* and *HWP1*. Therefore it may be possible to observe *EAP1* mutations, simply by increasing either the number of colonies and/or the number of generations above the 60 colonies and the 300 generations used in these studies.

Allelic characterization of *YWP1*, *HWP1*, and *EAP1* in strains from five different clades of *C. albicans*

The strains of *C. albicans* have been assigned by MLST and Ca3 probes into five large clades [79]. Strains from each of the five clades have distinct genetic backgrounds, but similar genetic diversity. The GPG strains used in these studies have been associated with clade 1 [78]. GPG strains have a different genetic background to non-GPG strains, and the genetic diversity of non-GPG strains is much greater than GPG strains [56]. Therefore, characterization of *YWP1*, *HWP1*, and *EAP1* in strains of these five different clades is of interest, and may reveal more information about the significance of the repeat regions. For example, allelic characterization of *HWP1* in strains of these five different clades could confirm the results of these studies that *HWP1* is a contingency gene. In addition, allelic characterization of *YWP1* in strains of these five different clades could reveal whether different clade strains share alleles containing 5 repeat units, allowing a more distinctive categorization of *YWP1* as a contingency gene. Likewise, allelic characterization of *EAP1* in strains of these five different clades could reveal whether different clade strains select different predominant alleles and alleles combinations, which show an impact of the genetic background on the allelic distribution of *EAP1*.

Relation Between *HWP1* Allele Variability and β 1.2 Mannosides

$\beta-1.2$ Mannosides (β-Mans) are molecules found in the *Candida albicans* cell wall that are associated with virulence due to their adhesin and immunomodulatory properties [118]. Different numbers of β-Mans have been found in different strains, cell forms, and growth conditions [119, 120, 121, 122]. The number varies from 1 to 14 [123]. β-Mans have been found to stimulate the production of a specific antibody that protects mice from candidiasis [124]. In humans with candidiasis, this antibody has also been found to be present [125], to mediate the attachment of *C. albicans* cells to the host macrophage cells [126], and to stimulate macrophage cells to produce cytokines, that function in antimicrobial defense system [127]. The production of cytokines by macrophage cells is not optimal if the number of β-Mans is less than seven [127].

HWP1p is a cell wall protein that has been proven to be β-mannosylated [128]. Since *HWP1* appears to be a gene that is involved in adaptation of *C. albicans* in its host as shown in these studies, this gene most likely plays an important role in *C. albicans* virulence. Therefore, it is of interest to find out whether there is a correlation between the allele variability of *HWP1* found in these studies, and the number of β-Mans attached to HWP1p. Since the predominant alleles of *HWP1* can be considered to be better adapted alleles, these alleles may associated with particular numbers of β-Mans. If this is the case, then the number of β-Mans may be less than 7 which then leads to a significantly decreased production of cytokines by the host macrophage cells. If this hypothesis is proven, it would be further evidence for the importance of the *HWP1* protein in colonization and invasion of *C. albicans*.

In vivo Experiments to Observe Gene Interconnectedness in Commensal and Infection Strains

These studies show that gene interconnectedness among *YWP1*, *HWP1*, and *EAP1* is systematically higher in commensal strains than it is in infection strains. Since the gene interconnectedness was determined by analyzing allele combinations of the three genes in commensal and infection strains, an *in vivo* experiment to observe this interconnectedness would be of interest.

A possible experimental plan would involve commensal and pathogenic states of *C. albicans* in rats. Since there was a significant change in the interconnectedness between *YWP1* and *HWP1* in these states, it is reasonable to expect that a significant change would be seen in *in vivo* experiments. Cells of a strain of interest would be grown in series of rats until they reached a certain number of generations, that could reasonably be expected to give rise to mutations in both genes. The commensal and pathogenic *C. albicans* cells in rats would then be isolated, and DNA analysed using 454 sequencing [117] to identify the mutated alleles of *YWP1* and *HWP1*. Allele combinations of *YWP1* and *HWP1* for strains from commensal and pathogenic *C. albicans* would then be analysed to calculate the non-random association between the two genes. The results can be compared to confirm the results from the calculations

done in these studies.

Bibliography

[1] F.C. Odds. Candida and candidosis.1988. 2nd ed. london. balliere tindall.

[2] W.G. Powderly, K. Robinson, E.J.Keath.1993. Molecular pidemiology of recurrent oral candidiasis in human immunodeficiency virus-positive patients:evidence of two paterns of reccurence. J. Infect.D 168;463-466.

[3] A. Voss, R.J. Hollis, M.A. Pfaller, R.P.Wenzel,B.N.Doebbeling.1994.Investigation of the sequence of colonization and candidemia in nonneutropenic patients. J. Clin. Microbiol. 32:975-980.

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

[5] R.D.Cannon and Chaffin, W.L. 1999. Oral colonization by *Candida albicans*. Crit. rev. oral Biol. med. 10: 359-383.

[6] J.D.Buck, P.M. Bubucis, T.J. Combs. 1977. Occurance of human-associated yeast in bivalve shellfish fromLong Island Sound. Appl.Environ.Microbial. 33;370-378.

[7] R.S. Togbi, L.P. Samaranayake, and T.W. MacFarlane. 1988. *In vitro* susceptibility of Candida species to lysozyme. Oral Microbiol. Immunol. 3:35-39.

[8] B.S. Schonwetter, E.D. Stolzenberg, and M.A. Zasloff. 1995. Epithelial antibiotics induced at sites of inflammation. Science 267:1645-1648.

[9] H. Nikawa, L.P. Samaranayake, and J. Tenovuo J. 1993. The fungicidal effect of human lactoferrin on *Candida albicans* and *Candida krusei*. Arch. Oral Biol. 38:1057-1063.

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

[11] K Nather and C. A. Munro. 2008. Generating cell surface diversity in *Candida albicans* and other fungal pathogens. FEMS Microbiol. Lett. 285: 137-145.

[12] J.C. Kapteyn, L.L. Hoyer, J.E. Hecht, W.H. Muller, A. Andel, A.J. Verkleij, M. Makarow, H. Van Den Ende, and F.M. Klis. 2000. The cell wall architecture of Candida albicans wild-type cells and cell wall-defective mutants. Mol. Microbiol. 35:601-611.

[13] P.W. De Groot, K.J. Hellingwerf, and F.M. Klis. 2003. Genome-wide identification of fungal GPI proteins. Yeast 20:781-796.

[14] G.J. Smits, J. C. Kapteyn, H. van den Ende, and F.M. Klis. 1999. Cell wall dynamics in yeast. Curr. Opin. Microbiol. 2:348-352.

[15] W.L. Chaffin. *Candida albicans* cell wall proteins. 2008. Microbiol. Mol. Biol. Rev. 72: 495544.

[16] Q.Y. Yin, P.W.J. de Groot, C.G. de Koster, and F.M.Klis. 2007. Mass spectrometry-based proteomics of fungal wall glycoproteins. Trends in Microbiology 1: 20-26.

[17] F. Li, M.J. Svarovsky, A.J. Karlsson, J.T. Wagner, K. Marcillo, P. Oshel, D. Andes, and S.P. Palecek. 2007. EAP1p, an adhesin that mediates *Candida albicans* biofilm formation *in vitro* and *in vivo*. Eukarayot. Cell 6:931-939.

[18] Q.T. Phan, C.L. Myers, F. Yue, D.C. Sheppard, M.R. Yeaman, W.H. Welch, A.S. Ibrahim, J.E. Edwards, and S.G. Filler. 2007. *Als3* is a *Candida albicans* invasions that binds to cadherins and induces endocytosis by host cells. PLoS Biology 5: 1-15.

[19] M.L. Richard and A. Plaine. Comprehensive analysis of glycosylphosphatidylinositolanchoredproteins in *Candida albicans*. 2007. Eukaryot. Cell 6: 119-133.

[20] M.B. Fierman and B.P. Cormack. The omega-site sequence of glycosylphosphatidylinositol-anchored proteins in *saccharomyces cerevisiae* can determine distribution between the membrane and the cell wall. 2003. Mol. Microbiol. 50:883-8966.

[21] P.W. De Groot, A.D. de Boer, J. Cunningham, H.L. Dekker, L. de Jong , K.J. Hellingwerf, C. de Koster, and F.M. Klis. 2004. Proteomicanalysis of *Candida albicans* cell walls reveals covalently boundcarbohydrate-active enzymes and adhesins. Eukaryot. Cell 3:955-965.

[22] F.M. Klis, A. Booorsma, and P.W.J. de Groot. 2006. Cell wall construction in *Saccharomyces cerevisiae*. Yeast 23:185-202.

[23] E.R. Setiadi, T. Doedt, F. Cottier, C. Noffs, and J. Ernst. 2006. Transcriptional response of *Candida albicans* to hypoxia: linkage of oxygen sensing and Efg1p-regulatory networks. J. Mol. Biol. 361:399-411.

[24] R.O. Ebanks, K. Chisholm, S. McKinnon, M. Whiteway, and D.M. Pinto. 2006. Proteomic analysis of *Candida albicans* yeast andhyphal cell wall and associated proteins. Proteomics 6:2147-2156.

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

[26] E. Levdansky, H. Sharon, N. Osherov. 2008. Coding fungal tandem repeats as generators of fungal diversity. Fungal Biol. Rev. 22:85-96.

[27] S.H. Oh, G. Cheng, J.A. Nuessen, R. Jajko, K.M. Yeater, X. Zhao, C. Pujol, D.R. Soll, L.L. Hoyer. 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 151:673681.

[28] L. Loza, Y. Fu, A.S. Ibrahim, D.C. Sheppard, S.G. Filler, J.E. Edwards Jr. 2004. Functional analysis of the *Candida albicans ALS1* gene product. Yeast 2:473482.

[29] J.M. Rauceo, R. De Armond, H. Otoo, P.C. Kahn, S.A. Klotz, N.K. Gaur, and P.N. Lipke. 2006. Threonine-rich repeats increase fibronectin binding in the *Candida albicans* adhesin Als5p. Eucaryot. Cell 5:16641673.

[30] L. Hoyer, C. Green, S. Oh, X. Zhao. 2008. Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family- a sticky pursuit. Med. Mycol. 46:1-15.

[31] T.J. Lott, B.P. Holloway, D.A. Logan, R. Fundyga, J. Arnold. 1999. Towards understanding the evolution of the human commensal yeast *Candida albicans*. Microbiology 145:11371143.

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

[33] X. Zhao, S.H. Oh, R. Jajko, D.J. Diekema, M.A. Pfaller, C. Pujol, D.R. Soll, and L.L. Hoyer. 2007. Analysis of *ALS5* and *ALS6* allelic variability in a geographically diverse collection of *Candida albicans* isolates. Fungal Genet. Biol. 44:12981309.

[34] P. Jordan, L. Snyder, and N. Saunders. 2003. Diversity in coding tandem repeats in related *Neisseria spp.* BMC Microbiology 3:23-37.

[35] E. Levdansky,J. Romano,Y. Shadkchan,H. Sharon, J.Kevin,Verstrepen, G. R. Fink, N. Osherov. Coding Tandem Repeats Generate Diversity in *Aspergillus fumigatus* Genes. 2007. Eukaryot. Cell 6:1380-1391.

[36] G.F. Richard and F. Paques. 2000. Mini- and microsatellite expansions: The recombination connection. EMBO Reports 11:122-126.

[37] T.A. Kunkel. 1993. Slippery DNA and diseases. Nature 365:207208.

[38] C.E. Pearson, K. Nichol Edamura, J.D. Cleary. 2005. Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. 6:729742.

[39] H. Murakami, N. Sugaya, M. Sato, A. Imaizumi, S. aburatami, and K. Horimoto. 2004. Detection of Inter-spread repeat sequence in genomic DNA sequence. Genome Inform. 15: 170-179.

[40] G. Benson. 1999. Tandem repeats finder: A program to analyze DNA sequences. 1999. Nucleic Acids Res. 27:573-80.

[41] D.R. Kurtz and C. Schleiermacher. REPuter: Fast computation of maximal repeats in complete genomes. 1999. Bioinformatics 15:426-427.

[42] M. Legendre, N. Pochet, T. Pak, and K.J. Verstrepen. Sequence-based estimation of minisatellite and microsatellite repeat variability. 2007. Genome Res. 17:1787-1796.

[43] D. Field, M.O. Magnasco, E.R. Moxon, D. Metzgar, M.M. Tanaka, C. Wills, and D.S. Thaler. 1999. Contingency loci, mutator alleles, and their interactions. Mol. Strat. Biol. Evol. 870:378-382.

[44] Moxon, E.R., Rainey, P.B., Novak, M.A., Lenski, R.E.1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. 1:24-33.

[45] E.R.Moxon and D.S.Thaler.1997. Microbial genetics. The tinkerer's evolving toolbox. Nature 387:659.

[46] D. Metzgar and C. Wills. 2000. Evolutionary changes in mutation rates and spectra and their influence on the adaptation of pathogens. Microbes Infect. 2:1513-1522.

[47] S. Bretagne, J-M. Costa,C. Besmond, R. Carsique, and R. Calderone. 1997. Microsatellite polymorphism in the promoter sequence of the Elongation Factor 3 gene of *Candida albicans* as the basis for a typing system. J. Clin. Microbiol. 35:17771780.

[48] P. Sampaio, E. Nogueira, A.S. Loureiro, Y. Delgado-Silva, A. Correia, and C. Pais. 2009. Increased number of glutamine repeats in the C-terminal of *Candida albicans* Rlm1p enhances the resistance to stress agents. Anton. Van. Lee. 96:395404.

[49] L. Hoyer. 2001. The ALS gene family of *Candida albicans*. Trends Microbiol. 9:176180.

[50] D.M. MacCallum, L. Castillo, K. Nather , C.A. Munro, A.J.P. Brown, N.A.R. Gow, and F. C. Odds. 2009 Property differences among the four major *Candida albicans* strain clades. Eukaryot. Cell 8: 373387.

[51] N. Zhang, R.D. Cannon, B.R. Holland, M.L. Patchett, and J. Schmid. 2010. Impact of genetic background on allele selection in a highly mutable *Candida albicans* gene, *PNG2*. PLos One 5:e9614.

[52] Zhuo Zhou. Evidence that *SSR1* can act as hypermutable contingency gene in *Candida albucans*. 2010. MSc Thesis, Institute of Molecular BioSciences, Massey University.

[53] D.W. Hood, M.E. Deadman, and M.P. Jennings. 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc. Natl. Acad. Sci. USA 93:11121-11125.

[54] J.N. Weiser, J.M. Love, and E.R. Moxon. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell 59:657-665.

[55] R. Moxon, C. Bayliss, D. Hood. 2006. Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. Annu. Rev. Genet. 40:307333.

[56] J. Schmid, S. Herd, P.R. Hunter, R.D. Cannon, M.S. Yasin, S. Samad, M. Carr, D. Parr, W. McKinney, M. Schousboe, B. Harris, R. Ikram, M. Harris, A. Restrepo, G. Hoyos, and K.P. Singh. 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.

[57] D.D.G. Gessler and S. Xu. Meiosis and the evolution of recombination at low mutation rates. 2000. Genetics 156:449-456.

[58] S.A. Lee, S. Wormsley, S. Kamoun, A.F.S. Lee, K. Joiner, and B. Wong. 2003. An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. Yeast 20:595610.

[59] B.R. Braun, M. van het Hoog, C. d'Enfert, M. Martchenko, J. Dungan, A. Kuo, D.O. Inglis, M.A. Uhl, H. Hogues, M. Berriman, M. Lorenz, A. Levitin, U. Oberholzer, C. Bachewich, D. Harcus, A. Marcil, D. Dignard, T. Iouk, R. Zito, L. Frangeul, F. Tekaia, K. Rutherford, E. Wang, C.A. Munro, S. Bates, N.A. Gow, L.L. Hoyer, G. Khler, J. Morschhuser, G. Newport, S. Znaidi, M. Raymond, B. Turcotte, G. Sherlock, M. Costanzo, J. Ihmels, J. Berman, D. Sanglard, N. Agabian, A.P. Mitchell, A.D. Johnson, M. Whiteway, A. Nantel. 2005. A human-curated annotation of the *Candida albicans* genome. PLos Genet.1:36-57.

[60] B.L, Granger, M.L. Flenniken, D.A. Davis, A.P. Mitchell, and J.E. Cutler. 2005. Yeast wall protein 1 of *Candida albicans*. Microbiology 151:1631-1644.

[61] A. Nantel, D. Dignard, C. Bachewich, D. Harcus, A. Marcil, A. Bouin, C.W. Sensen, H. Hogues, M. van het Hoog, P. Gordon, T. Rigby, F. Benoit, D.C. Tessier, D.Y. Thomas, and M. Whiteway. 2002. Transcription profiling of *Candida albicans* cells undergoing the yeast-to-hyphal transition. Mol. Biol. Cell 13, 34523465.

[62] J. Staab, S.Bradway, P. Fidel, and P.Sundstrom. 1999. Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* Hwp1. Science 283: 1535- 1538.

[63] B. Wachtler, D. Wilson, K. Haedicke, F. Dalle, and B. Hube. 2011. From attachment to damage: defined genes of *Candida albicans* mediate adhesion, invasion and gamage during interaction with oral epithelial cells. PLos One 6:e17046.

[64] C.J. Nobile, H.A. Schneider, J.E. Nett, D.C. Sheppard, S.G. Filler, D.R. Andes, and H.P. Mitchell. 2008. Complementary adhesin function in *C. albicans* biofilm formation. Curr. Biol. 18:1017-1024.

[65] A.F. Li and S.P. Palecek. 2003. *EAP1*, a Candida albicans gene involved in binding human epithelial cells. Eukariot. Cell 2:1255-1273.

[66] F. Li and S.P. Palecek. 2008. Distinct domains of the *Candida albicans* adhesin Eap1p mediate cellcell and cellsubstrate interactions. Microbiologi 154:1193-1203.

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

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

[69] J. Schmid, E. Voss, and D.R. Soll. 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.

[70] M. Legrand, P. Lephart, A. Forche, F-M.C. Mueller, T. Walsh, P.T. Magee, and B.B. Magee. 2004. Homozygosity at the *MTL* locus in clinical strains of *Candida albicans*: karyotypic rearrangements and tetraploid formation. Mol. Microbiol. 52:1451-1462.

[71] D.J. Sheskin. 2004. Handbook of parametric and nonparametric statistical procedures. 3rd ed. A CRC Press Company.

[72] E.H. Simpson. Measurement of diversity. 1949. Nature 163:688.

[73] J.R. Dillon, M. Rahman, and K. Yeung. 1993. Discriminatory power of typing schemes based on Simpson's index of diversity for *Neisseria gonorrhoeae*. J. Clin. Microbiol. 31:2831-2833.

[74] K.J. Verstrepen, T.B. Reynolds, G.R. Fink. 2004. Origins of variation in the fungal cell surface. Nat. Rev. Microbiol. 2:533540.

[75] K.W. Deitsch, S.A. Lukehart, J.R. Stringer. 2009. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. Nat. Rev. Microbiol. 7:493503.

[76] X. De Bolle et al. 2000. The length of tetranucleotide repeat track in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol. Microbiol. 35:211-222.

[77] Evolutionary analysis. 2004. Pearson Education Inc. New Jersey.

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

[79] F.C. Odds, M.E. Bougnoux, D.J. Shaw, J.M. Bain, A.D. Davidson, D. Diogo, M.D. Jacobsen, M. Lacomte, S.Y. Li, A. Tafanti. Molecular philogenetics of *Candida albicans*. 2007. Eukaryot. Cell 6:1041-1052.

[80] J. Schmid, A.M. Tortorano, G. Jones, C. Lazzarini, N. Zhang, M.J. Bendall, M. Cogliati, S. Wattimena, and L. Klingspor. 2011. Increased mortality in young candidemia patients associated with presence of a *Candida albicans* general-purpose genotype. J. Clin. Microbiol. 49:3250-3256.

[81] X. Zhao, S.H. Oh, and L.L. Hoyer. Deletion of *ALS5*, *ALS6*, or *ALS7* increases adhesion of *Candida albicans* to human vascular endothelial and buccal epithelial cells. 2007. Med. Mycol. 45:429-434.

[82] F. Dalle, L. Dumont, N. Franco, D. Mesmacque, D. Caillot, P. Bonnin, C. Moiroux, O. Vagner, B. Cuisenier, S. Lizard, and A. Bonnin1. 2003. Genotyping of *Candida albicans* oral strains from healthy individuals by polymorphic microsatellite locus analysis. J. Clin. Microbiol. 41:2203-2205.

[83] A.W. Al-Aidan, W. Goessens, N. Lemmens-den Toom, M. Al-Ahdal, and A. van Belkum. 2007. Microevolution in genomic short sequence repeats of Candida albicans in non-neutropenic patients. Yeast 24:155-160.

[84] F.V. Lunel, L. Licciardello, S. Stefani, H.A. Verbrugh, W.J.G. Melchers, J.F.G.M. Meis, S. Scherer, and A. van Belkum. 1998. Lack of consistent short sequence repeat polymorphisms in genetically homologous colonizong and invasive candida albicans strains. J. Bacteriol. 180:3771-3778.

[85] F. Dalle, N. Franco, J. Lopez, O. Vagner, D. Caillot, P. Chavanet, B. Cuisenier, S. Aho, S. Lizard, and A. Bonnin. 2000. Comparative genotyping of *Candida albicans* bloodstream and nonbloodstream isolates at a polymorphic microsatellite locus. J. Clin. Microbiol. 38:4554-4559.

[86] C.Pujol, A.R.Dodgson, D.R.Soll. 2005. Population genetics of ascomycetes pathogenic to humans and animals. In: Xu, J. (Ed.), Evolutionary Genetics of Fungi. Horizon Scientific Press, Norfolk, UK, pp. 149-188.

[87] A. Coste, V. Turner, F. Ischer, J. Morschhauser, A. Forche, A. Selmecki, J. Berman, J. Bille, and D. Sanglard. 2006. A Mutation in Tac1p, a transcription factor regulating CDR1 and CDR2, is coupled with loss of heterozygosity at chromosome 5 to mediate antifungal resistance in *Candida albicans*. Genetics 172: 2139-2156.

[88] 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. 2009. A.C.B.Padovan, G.M.Chaves, A.R.Colombo, M.R.S.Briones. Med. Mycol. 47: 824-835.

[89] J.R.Naglik, F.Fostira, J.Ruprai, J.F. Staab, S.J.Challacombe, and P. Sundstrom. 2006. *Candida albicans HWP1* gene expression and host antibody responses in colonization and disease. J. Med. Microbiol. 55:1323-1327.

[90] J.Chandra, J.D.Patel, J.Li, G.Zhou, P.K.Mukherjee, T.S. McCormick, J.M.Anderson, and M.A.Ghannoum. 2005. Modification of surface properties of biomaterials influences the ability of *Candida albicans* to form biofilms. Appl. Environ. Microbiol. 71:8795-8801.

[91] H.W.Jeng, A.R.Holmes, and R.D.Cannon. 2005. Characterization of two *Candida albicans* surface mannoprotein adhesins that bind immobilized saliva components. Med. Mycol. 43:209-217

[92] C.L.Price, D.W.Williams, M.G.Waters, L.Coulthwaite, J.Verran, R.L.Taylor, D.Stickler, and M.A.Lewis. 2005. Reduced adherence of Candida to silane-treated silicone rubber. J. Biomed. Mater. Res. B 74:481-487.

[93] G.Ramage, S.P.Saville, D.P.Thomas, and J.L.Lopez-Ribot. 2005. Candida biofilms: an update. Eukaryot. Cell 4:633-638.

[94] D.R.Singleton, J.Masuoka, and K.C.Hazen. 2005. Surface hydrophobicity changes of two *Candida albicans* serotype B mnn4 mutants. Eukaryot. Cell 4:639-648.

[95] B.P. Cormack, N. Ghori, and S. Falkow. 1999. An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. Science 285:578-582.

[96] B. Guo, C.A. Styles, Q. Feng, and G.R. Fink. 2000. A Saccharomyces gene family involved in invasive growth, cell-cell adhesion, and mating. Proc. Natl. Acad. Sci. USA 97:12158-121.

[97] A.H.D. Brown, M.W. Feldman, and E. Nevo. 1980. Multilocus structure of natural populations of Hordeum spontaneum . Genetics 96:523536.

[98] B. Haubold, M. Travisano, P.B. Rainey, and R.R. Hudson. 1998. Detecting linkage disequilibrium in bacterial populations. Genetics 150:13411348.

[99] J.M. Smith, N.H. Smith, M. ORourke, and B.G. Spratt. 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA 90:43844388.

[100] P.M. Agapow and A. Burt. 2001. Indices of multilocus linkage disequilibrium. Mol. Ecol. Notes 1:101102.

[101] A. Burt, D.A. Carter, G.L. Koenig, T.J. White, and J.W. Taylor. 1996. Molecular markers reveal cryptic sex in the human pathogen Coccidioides immitis. Proc. Natl. Acad. Sci. USA 93:770773.

[102] A. Edwards, H.A. Hammond, L. Jin, C.T. Caskey, and R. Chakraborty. Genetic variation at five trimeric and tetramerik tandem repeat loci in four human population groups. 1992. Genomics 12:241-253.

[103] C. Pujol, J. Reynes, F. Renaud, M. Raymond, M. Tibayrenc, F.J. Ayala, F. Janbon, M. Mallie, J.M. Bastide. The yeast *Candida albicans* has a clonal mode of reproduction in a population of infected human immunodeficiency virus-positive patients. 1993. Proc. Natl. Acad. Sci. USA 90:9456-9459.

[104] P. Boerlin, F. Boerlin-Petzold, J. Goudet, C. Durussel, J. Pagani, J.P. Chave, J. Bille. 1996. Typing *Candida albicans* oral isolates from human immunodeficiency virus-infected patients by multilocus enzyme electrophoresis and DNA fingerprinting. J. Clin. Microbiol. 34:1235-1248.

[105] M. Tibayrench and F.J. Ayala. 2002. The clonal theory of parasitic protozoa: 12 years on. Trends Parasitol. 18:405-410.

[106] F. Nebavi, F.J. Ayala, F. Renaud, S. Bertout, S. Eholie, K. Moussa, M. Mallie, and T. de Meeus. Clonal population structure and genetic diversity of *Candida albicans* in AIDS patients from Abidjan (Cote d'Ivoire). 2006. Proc. Natl. Acad. Sci. USA 10:3663-3668.

[107] M.J. Wade. Sewall Wright: gene interaction and the shifting balance theory. 1992. Oxf. Surv. Evol. Biol. 8:35-62.

[108] P. Sundstrom, E. Balish, and C.M. Allen. 2002. Essential role of the *Candida albicans* transglutaminase substrate, hyphal wall protein 1, in lethal oroesophageal candidiasis in immunodeficient mice. J. Infect. Dis. 185: 521530.

[109] N. Tsuchimori, L.L. Sharkey, W.A. Fonzi, S.W. French, J.E. Edwards, and S.G. Filler. 2000. Reduced virulence of *HWP1*-deficient mutants of *Candida albicans* and their interactions with host cells. Infect. Immun. 68: 1997 2002.

[110] X. Zhao, S.-H Oh, G. Cheng, C.B. Green, J.A. Nuessen, K. Yeater, R.P. Leng, A.J.P. Brown, and L.L. Hoyer. 2004. *ALS3* and *ALS8* represent a single locus that encodes a *Candida albicans* adhesin; functional comparison between Als3p and Als1p. Microbiology 150:24152428.

[111] L.L. Hoyer, S. Scherer, A.R. Shatzman, and G.T. Livi. *Candida albicans* ALS1: domains related to a *Saccharomyces cerevisiae* sexual agglutinin separated by a repeating motif. 1995. Mol. Microbiol. 15:39-54.

[112] X. Zhao, S. Oh, D.A. Coleman, and L.L. Hoyer. ALS51, a newly discovered gene in the Candida albicans ALS family, created by intergenic recombination: analysis of the gene and protein, and implications for evolution of microbial gene family. 2011. FEMS Immunol. Med. Microbiol. 61:245-257.

[113] D. Metzgar, D. Field, R. Haubrich, and C. Wills. Sequence analysis of a compound coding-region microsatellite in *Candida albicans* resolves homoplasies and provides a high-resolution tool for genotyping. 1998. FEMS Immunol. Med. Micribiol. 20:103-109.

[114] S.R. Lockhart, B.D. Reed, C.L. Pierson, D.R. Soll. Most frequent scenario for reccurent candida vaginitis is strain maintenance with "substrain shuffling": demonstration by sequential DNA fingerprinting with probes Ca3, C1, and CARE2. 1996. J. Clin. Microbiol. 4:767-777.

[115] M.A. Pfaller. Epidomology of candidiasis. 1995. J. Hosp. Infect. 30(Suppl.):329-338.

[116] P. Sampaio, L. Gusmao, C. Alves, C. Pina-Vaz, A. Amorim, and C. Pais. Higly polymorphic microsatellite for identification of *Candida albicans* strains. 2003. J. Clin. Microbiol. 41:552-557.

[117] C.W. Wheat. 2010. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138:433-451.

[118] D. Poulain and T. Jouault. *Candida albicans* cell wall glycans, host receptors and responses: elements for a decisive crosstalk. 2004. Curr. Opin. Microbiol. 7:342-349.

[119] N. Shibata, T. Ichikawa, M. Tojo, M. Takahashi, N. Ito, Y. Ohkubo, and S. Suzuki. 1985. Immunochemical study on the mannans of *Candida albicans* NIH-207, NIH B-792, AND J-1012 strains prepared by fractional precipitation with cetyltrimethylammonium bromide. Arch. Biochem. Biophys. 243:338-348.

[120] N. Shibata, S. Fukasawa, H. Kobayashi, M. Tojo, T. Yonezu, A. Ambo, Y. Ohkubo, and S. Suzuki. 1989. Structural-analysis of phospho-D-mannan-protein complexes from yeast and mold form cells of Candida albicans NIH A-207 serotype-A-strains. Carbohydr. Res. 187:239-253.

[121] H. Kobayashi, S.H. Takahashi, N. Shibata, M. Miyauchi, M. Ishida, J. Sato, K. Maeda, and S. Suzuki. 1994. Structure of cell-wall mannans of pathogenic *Candida-trpicalis* IFO-0199 and IFO-1647 yeast strains. Infect. Immun. 62:968-973.

[122] Y. Okawa, T. Takahata, M. Kawamata, M. Miyauchi, N. Shibata, A. Suzuki, H. Kobayashi, and S. Suzuki. Temperature-dependent change of serological specificity of *Candida albicans* NIH-A-207 cells cultured in yeast extract-added sabouraud liquid-medium disappearance of surface antigenic factor-4, factor-5, and factor-6 at high temperature. 1994. FEBS Lett. 345:167-171.

[123] P.A. Trinela, G. Lepagea, T. Jouaulta, G. Strecken, D. Poulain. Definitive chemical evidence for the constitutive ability of *Candida albicans* serotype A strains to synthesize β-1,2 linked oligomannosides containing up to 14 mannose residues. 1997. FEBS Lett. 416:203-206.

[124] Y. Han and J. Cutler. Antibody-response that protects against disseminated candidiasis. 1995. Infect. Immun. 63:2714-2719.

[125] T. Jouault, C. Delaunoy, B. Sendid, F. Ajana, and D. Poulain. Differential humoral response against alpha- and beta-linked mannose residues associated with tissue invasion by *Candida albicans*. 1997. Clin. Diag. Lab. Immunol. 4:328-333.

[126] R. Li and J. Cutler. Chemical definition of an epitope/adhesin molecule on *Candida albicans*. 1993. J. Biol. Chem. 268:18293-18299.

[127] T. Jouault, G. Lepage,A. Bernigaud, P-A. Trinel, C. Fradin, J-M. Wieruszeski, G. Strecker, and D. Poulain. β 1,2-linked oligomannosides from *Candida albicans* act as signals for tumor necrosis factor alpha production. 1995. Infect. Immun. 63:2378-2381.

[128] C. Fradin, M.C. Slomianny, C. Mille, A. Masset, R. Robert, B. Sendid, J.F. Ernst, J.C. Michalski, and D. Poulain. β 1,2 oligomannose adhesin epitopes are widely distributed over the different families of *Candida albicans* cell wall mannoproteins and are associated through both N- and O-glycosylation processe. 2008. Infect. Immun. 76:4509-4517.