The Functional Significance of Allelic Diversity in *Candida albicans*

Submitted by Sophie Shaw to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Biological Sciences
In July 2014

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

Abstract

Allelic expression imbalance, or AEI, is the term given to differences in the expression levels of the two alleles of a gene. AEI has been previously identified in a number of species using various techniques. Here, the genome-wide extent of allelic expression imbalance in the pathogenic yeast species, *Candida albicans,* was examined through use of RNA sequencing in combination with a novel computational pipeline based around the diploid reference genome. Techniques for validating these results were investigated, and the difficulties surrounding specificity and quantification are discussed.

As *C. albicans* is a highly heterozygous species, it was hypothesised that polymorphisms within alleles lead to differences in allele expression, which are further linked to differences in allele function. The functional consequences of AEI were therefore interrogated through investigation of Gene Ontology, identification of condition specific responses in AEI, and targeted construction and phenotypic screening of heterozygous knockout strains. Together, these results strongly suggest that divergence in allele expression is not linked to differences in allele function.

Investigations of the possible control mechanisms behind the differences in allele expression were considered, with a focus upon structural factors such as chromosomal location, GC content, allele length and codon usage. However, issues with establishing causality are present, and difficulties lie in distinguishing between functional differences and consequences of bias in sequencing technologies.

This piece of research has advanced the understanding of gene expression mechanisms within a medically important pathogen, paving the way for further investigations into the functional consequences of allelic expression imbalance in *Candida albicans.*

Acknowledgements

There are a number of people who have supported me throughout my PhD who I would like to thank. Without you, there would be no thesis.

I would like to thank my supervisor, Dr. Mark Ramsdale, for his constant advice and guidance. I would like to thank both present and past members of the yeast team for teaching me all I know – Dr. Steve Bates, Dr. Jill Cheetham, Dr. Derek Wilkinson, Dr. Stephen Milne and Dr. Deborah Lloyd. Special thanks go to all members of the Haynes group for taking me under their wing and looking after me.

For technical support, I would like to thank the members of the Exeter Sequencing Service for the RNA sequencing, in particular Dr. Konrad Paszkiewicz, who ran the CLCBio analysis, calculated percentage protein identities, and patiently responded to the constant stream of emails and questions. Thanks to Paul O'Neill for modifying my script so that it ran in a nice 12 hours instead of five and half years. Thanks to Prof. Howard Jenkinson and Dr. Lindsay Dutton at the University of Bristol for the co-culture experiments and thanks to Richard Tennant for help with the flow cytometry work.

My friends and family have been a constant source of support and comfort over the past four years. I would like to thank all my girls: Lauren, Hsueh-Lui, Emma, Mel, Josie and Jane for keeping me fed, watered, exercised and smiling. And I'd especially like to thank my husband, Simon, for being there and putting up with me for all this time!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3D	Three dimensional
5-FOA	5-fluoro-orotic acid
6xHis	6 x Histidine residues
AEI	Allelic expression imbalance
AIDS	Acquired immunodeficiency syndrome
ALL	Acute lymphoid leukaemia
Allim	Allelic imbalance metre software
ALS	Agglutinin-like sequence
ASAP	Allele-specific alignment pipeline
ATP	Adenosine triphosphate
BAM	Binary version of a SAM file
BECs	Buccal epithelial cells
bp	Base pairs
BSA	Bovine serum albumin
CAI	Codon adaptation index
cAMP	Cyclic adenosine monophosphate
CBI	Codon bias index
cDNA	Complementary DNA
cells/ml	Cells per millilitre
cfu	Colony forming units
CGD	*Candida* genome database
Ct	Threshold value
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
DEPC	Diethylpyrocarbonate
d.f.	Degrees of freedom
dGTP	Deoxyguanosine triphosphate

DIG	Digoxigenin
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
dN/dS	Ratio of substitution rates at non-synonymous vs. synonymous sites
d.p.	Decimal places
DTT	Dithiothreitol
dTTP	Deoxythymidine triphosphate
EDTA	Ethylenediaminetetraacetic acid
et al.	and others
etc.	Etcetera
EtOH	Ethanol
FISH	Fluorescence *in situ* hybridisation
FUMP	5-fluorouridylic acid
GFP	Green fluorescent protein
GCB	Measure of codon usage bias
GO	Gene Ontology
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	Human immunodeficiency virus
HPLC	High performance liquid chromatography
HRP	Horseradish peroxidase
INCA	Interactive codon usage analysis
INDEL	Insertion or deletion
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kb	Kilobase
kDa	Kilodalton
KO	Knockout

LiAc	Lithium acetate
LB	Luria-Bertani (media)
LOH	Loss of heterozygosity
M	Molar
m/s	Metres per second
MAE	Autosomal monoallelic expression
MAPK	Mitogen activated protein kinase
Mb	Megabase
mg	Milligram
mg/ml	Milligram per millilitre
MIC	Minimum inhibitory concentration
ml	Millilitre
MLST	Multi-locus sequence typing
mM	Millimolar
mm	Millimetre
MOPS	3-(N-morpholino)propanesulfonic acid
MTL	Mating-type like locus
NaOH	Sodium hydroxide
NAT	Nourseothricin
ng	Nanogram
ng/µl	Nanogram per microlitre
nl	Nanolitre
nm	Nanometre
nM	Nanomolar
NMD	Nonsense-mediated decay
OPC	Oral pseudomembranous candidiasis
ORF	Open reading frame
OSCC	Oral squamous cell carcinoma
pBS	pBluescript
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction

PEG	Polyethylene glycol
PKA	Protein kinase
pmol/µl	Picomoles per microlitre
pH	Measure of acidity or basicity of a solution
PVDF	Polyvinylidene fluoride
QconCAT	Quantification concatamer
qPCR	Quantitative real time polymerase chain reaction
QTL	Quantitative trait loci
RBT	Repressed by *TUP1*
RNA	Ribonucleic acid
RNAi	RNA interference
RNA-seq	RNA sequencing
RPKM	Reads per kilobase per million mapped reads
rpm	Revolutions per minute
RPMI	Roswell Park Memorial Institute Media
RT PCR	Reverse transcription polymerase chain reaction
SAM	Sequence alignment/map
SAP	Secreted aspartic proteinase
SC	Synthetic complete media
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide agarose gel electrophoresis
siRNA	Small interfering ribonucleic acid
SNP	Single nucleotide polymorphism
spp.	Species
SSC	Sodium chloride sodium citrate buffer
TAE	Tris-acetate-EDTA
tBOOH	tert-Butyl hydroperoxide
TE	Tris-EDTA
TMM	Trimmed mean of M
tRNA	Transfer ribonucleic acid

U	Units
U/µl	Units per microlitre
U.S.	United States of America
UTR	Untranslated region
VVC	Vulvovaginal candidiasis
v/cm	volts per centimetre
v/v	volume to volume ratio
w/v	weight to volume ratio
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
XTT	2,3-Bis(2-methoxy-4-nitro-5-sulfophenyl)-2*H*-tetrazolium-5-carboxanilide
YE	Yeast extract
YCB	Yeast carbon base
YNB	Yeast nitrogen base
YPD	Yeast extract peptone glucose media
YPE	Yeast extract peptone ethanol media
µg	Microgram
µg/ml	Microgram per millilitre
µl	Microlitre
µM	Micromolar
µm	Micrometre
°C	Degrees Celsius

Chapter 1: Introduction

This body of work aims to investigate the functional consequences of allelic expression imbalance (AEI) in the pathogenic yeast *Candida albicans,* examining the simple hypothesis that alleles which differ in expression level may also differ in function. Here, as an introduction, *C. albicans* will be described alongside its notable pathogenesis related characteristics. Heterozygosity and examples of differing allele function in *C. albicans* will be detailed. Allelic expression imbalance will then be overviewed, with examples from numerous species including *C. albicans.* Finally, suggested control mechanisms of AEI will be discussed.

1.1 *Candida albicans* and Related *Candida* Species – An Overview

The pathogen *Candida albicans* is a commensal ascomycetous yeast which causes both superficial infections in healthy individuals and life-threatening disseminated candidiasis in immune compromised patients (Calderone, 2002b). In healthy patients, *C. albicans* occupies various body cavities including the oral cavity, gastrointestinal tract and vaginal cavity (Soll, 2002). In response to physiological changes in the host, especially compromise of the immune system, *C. albicans* becomes an opportunistic pathogen which can invade a number of human tissues including mucosal tissues and organs. Resultantly, a broad spectrum of diseases are associated with *Candida* infection. At risk groups can be identified based on four predisposing factors: natural, dietary, mechanical and iatrogenic (Calderone, 2002a). Natural predisposing factors include diabetes, microbial infections, pregnancy, infancy, old age and lymphocyte defects. Mechanical predisposing factors include physical changes to the patient such as trauma, burns, wounds or prosthetic devices. A high proportion of infections are within the urinary tract due to the ability of *C. albicans* to form a biofilm upon catheters (Jarvis, 1995). Dietary factors are the least well studied predisposing factor but can include excess of food groups, such as carbohydrates, or deficiencies in vitamins. Iatrogenic factors are due to medical treatment. For example, immunosuppressive treatment given to transplant patients increases their risk of *Candida* infection due to depletion of

lymphocyte cells (Calderone, 2002a). As another example, although anti-microbial drugs prevent bacterial infections, they lead to a drastic change in the microbial flora of the gastrointestinal tract allowing *Candida* species to flourish (Jarvis, 1995). Patients often display a combination of these factors, as an example, a high incidence of *Candida* infections are seen in patients with acute lymphoid leukaemia (ALL) due to a combination of neutropenia in the patient and administration of anti-microbial drugs (Jarvis, 1995). Survival statistics in immune compromised individuals are low; even with rapid administration of antifungal drugs, hematogenously disseminated candidiasis has a mortality rate of 47% (Gudlaugsson *et al.*, 2003).

Candidiasis has emerged as a relatively recent medical condition, with increasing prevalence over the last 50 years (Pfaller and Diekema, 2007). This has coincided with an increase in the number of AIDS patients (Rex *et al.*, 1995) and a general increase in life expectancy (Pfaller and Diekema, 2007). Although more recent reports suggest that the incidence rates may now be declining (Hobson, 2003), this theory is often contested (Pfaller and Diekema, 2007). As *Candida* species are able to infect various body cavities within a wide range of patients, different types of candidiasis are clinically presented. Oral candidiasis is one of the earliest reported *Candida* infections, with oral pseudomembranous candidiasis (OPC) as the most common form (Ruhnke, 2002). Oesophageal candidiasis tends to be seen in patients with chronic disorders and is characterised by ulcers and lesions of the oesophagus (Ruhnke, 2002). Vulvovaginal candidiasis (VVC) is estimated to affect 75% of all women during their lifetime, with recurrent VVC occurring in a smaller population (Ruhnke, 2002). Infections can also be observed upon skin and nails, often referred to as cutaneous candidiasis (Ruhnke, 2002). Skin and mucosal infections can occur in patients who are either immunocompromised or non-immunocompromised. On the other hand, invasive candidiasis is only observed in patients with severe defects in the immune system (Ruhnke, 2002). Invasive candidiasis is used as a broad term which covers several forms of infection, blood stream infections are known as candidemia, hematogenously disseminated candidiasis occurs when infection spreads from candidemia to one or more organs and can be both acute or chronic, less frequently observed is a non-haematogenous infection of a single organ (Kullberg and Filler, 2002).

The most common cause of candidiasis is the species *Candida albicans,* although infections have been reported from a total of 17 different *Candida* species including *C. parapsilosis, C. tropicalis, C. krusei* and the distantly related species *C. glabrata* (Pfaller and Diekema, 2007). Accurately determining the current prevalence of *Candidiasis* is difficult as reports often only detail specific infection types or patient groups (Calderone, 2002a). In 1995, *Candida* infections were the sixth most common nosocomial pathogen (Jarvis, 1995), with higher occurrences of commonly reported bacterial infections such as *Escherichia coli* and *Staphylococcus aureus* observed. In the same study, *Candida* species made up 72.1% of all fungal species found, with *Candida albicans* being the most prominent of these species (76% of *Candida* spp.). By 1999, this statistic had risen with *Candida* species identified as the fourth most common hospital acquired infection in the U.S. (Edmond *et al.*, 1999), overtaking *E. coli* in its prevalence. Estimations of the cost of nosocomial candidiasis just in the U.S. have approached $1 billion per year (Miller *et al.*, 2001). It should be noted that these infection statistics have all been recorded in the U.S. and that rates have been suggested to be lower in other countries, especially within Europe (Hobson, 2003).

These factors of high occurrence, high mortality and high cost make understanding the fundamental mechanics of *C. albicans* infections ever more important so that new therapeutic methods can be developed.

Several phenotypic characteristics are associated with pathogenesis in *Candida* species and their ability to cause infections. A brief description of these characteristics follow.

1.1.1 Morphology

As a pleomorphic organism, *C. albicans* can be found in numerous distinct morphological forms. In laboratory conditions, growth is most commonly found in the unicellular budding yeast (blastospore) form, which readily switches to either long hyphae with parallel sides or elongated pseudohyphae which remain attached to the mother cell and form chains (Figure 1.1) (Berman, 2006). The switch between the yeast and hyphal form is well characterised as a link to virulence (Biswas *et al.*, 2007). Interestingly, although the switch between yeast

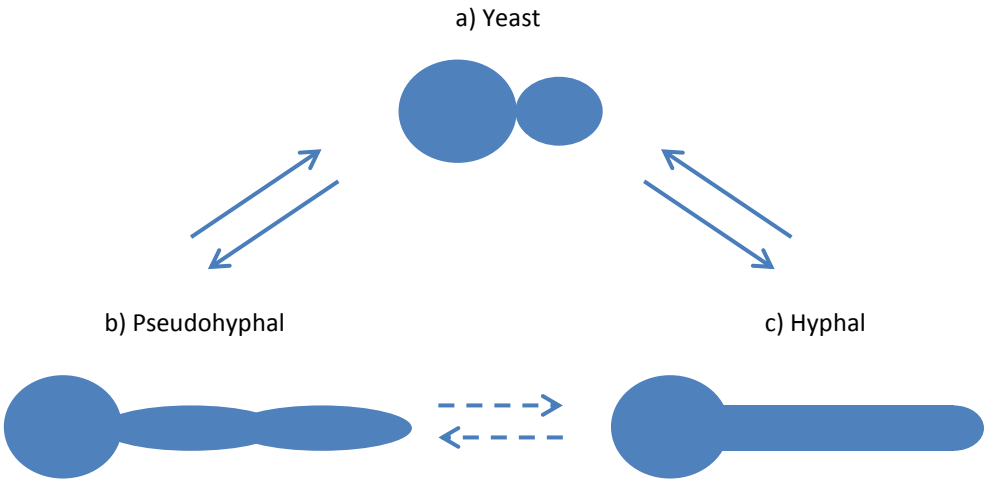


Figure 1.1 Common morphologies of *Candida albicans* a) yeast cells which form both b) pseudohyphal cells and c) hyphal cells. The switch between pseudohyphae and hyphae is less frequent. Figure adapted from (Berman, 2006).

and hyphae occurs readily, the transition between hyphal and pseudohyphal forms is less common (Berman, 2006). As well as observable morphological differences, yeast, pseudohyphae and hyphae differ in other characteristics such as cell cycle, gene expression and the mechanism of polarized growth (Berman, 2006, Kim and Sudbery, 2011). In addition to these morphological forms, under certain nutrient poor growth conditions, *C. albicans* will also form chlamydospores, very large spherical cells with thick cell walls (Figure 1.2). These structures are commonly observed *in vitro* but are rarely observed *in vivo* (Palige *et al.*, 2013). Although the biological purpose of this structure is still unknown, characteristics of this morphology include large liquid droplets and high amounts of RNA (Staib and Morschhäuser, 2007). Interestingly, this phenotypic form is only seen in *C. albicans* and *C. dubliniensis* and not observed in other *Candida* species. A recent study using RNA sequencing of *C. albicans* and *C. dubliniensis* chlamydospores has identified two genes, *CSP1* and *CSP2,* as potential chlamydospore specific cell wall proteins (Palige *et al.*, 2013). Taken together, these morphological characteristics demonstrate the adaptability of *C. albicans* to the ever-changing host environment during the infection process.

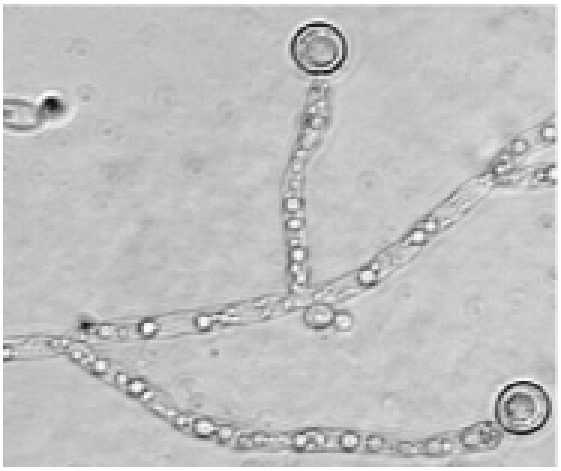


Figure 1.2. Micrograph of chlamydospores in *C. albicans* SC5314 cells after 4 days of growth at 25 °C on rice Tween-80 agar with the use of coverslips to generate microaerophilic conditions. Chlamydospores can clearly be seen as spherical cells at the end of filaments. Figure adapted from (Staib and Morschhäuser, 2007).

The rapid switch in phenotypic morphology is often triggered by changes in environmental cues, such as carbon source or pH. Changes in these environmental signals triggers a rapid switch in the expression of genes, leading to alterations in morphology. This process involves a complex network of interconnected signalling pathways, all triggered by different environmental cues, as demonstrated by Figure 1.3 (Biswas *et al.*, 2007). Two main pathways are utilised in *C. albicans* to trigger morphological changes; the MAPK pathway and the cAMP-PKA pathway. The MAPK pathway acts through a cascade of protein kinases, *CST20, HST7* and *CEK1.* The final protein kinase, *CEK1,* then activates the transcription factor *CPH1,* leading to expression of hyphal specific genes. Alternatively the cAMP-PKA pathway involves activation of the protein kinases *TPK1* and *TPK2,* by cAMP and the gene *BCY1.* The *TPK* genes then in turn activate the transcription factor *EFG1.* Interestingly these pathways are not exclusive and crosstalk may occur between them. Evidence suggests that both of these pathways are also under the control of the “master regulator” gene *RAS1.* As well as activation of hyphal specific genes, under certain conditions the gene *TUP1* acts as a hyphal repressor, ensuring that cells remain in the yeast form. *TUP1* is recruited to many genes including the RBT (repressed by *TUP1*) family, *HWP1* and *WAP1.* This recruitment is carried out by *NRG1* and *RFG1* and resultantly leads to the repression of these seven genes. The receptors to environmental cues are highly specialised for each condition. For

example amino acid sensing is controlled by the gene *CSY1,* which in turn activates the genes *GAP1* and *GPR1,* triggering the cAMP-PKA pathway, whereas sensing of ammonium occurs through *Mep2* which activates the MAPK pathway (Biswas *et al.*, 2007).

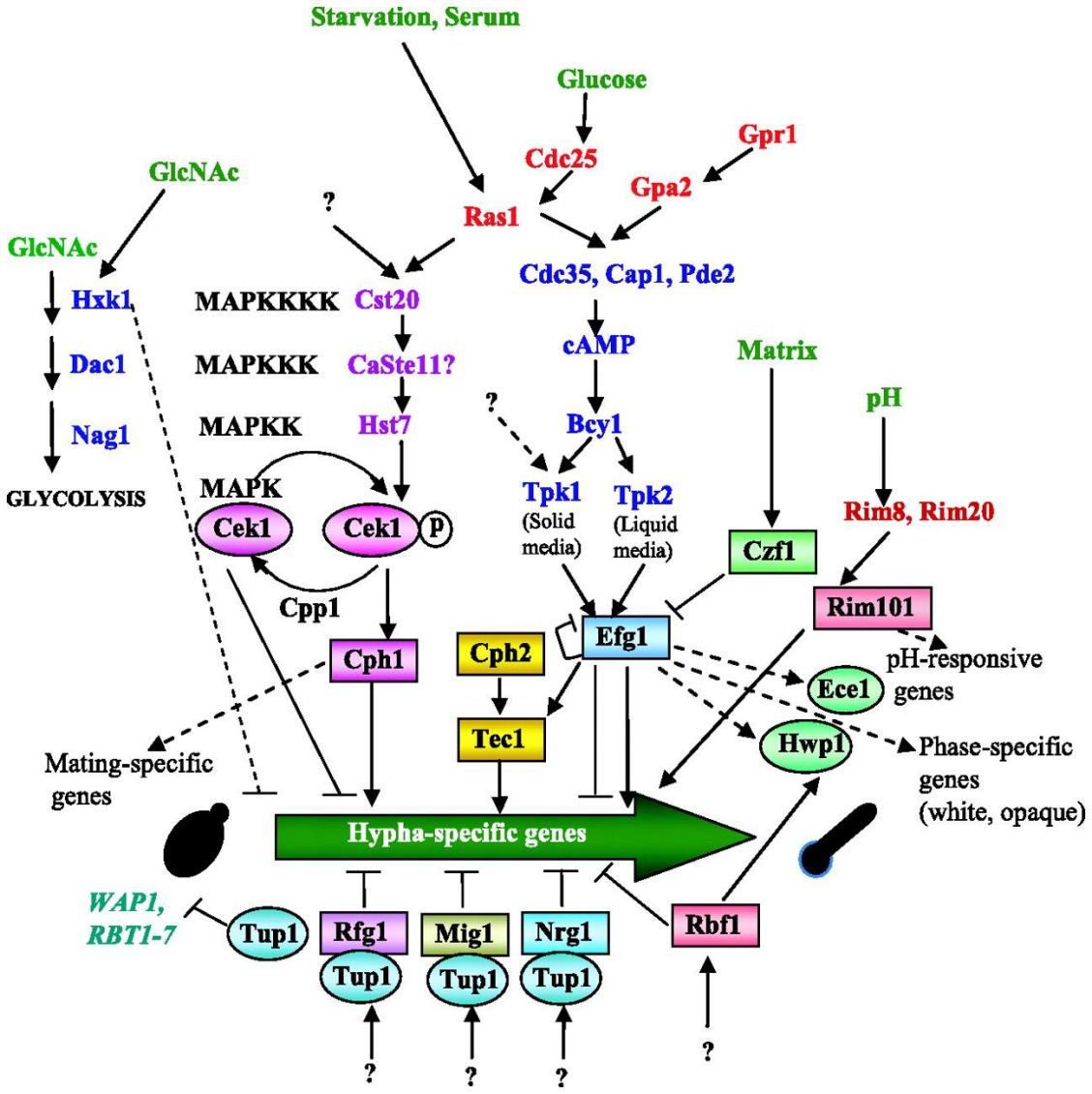


Figure 1.3 The complex network of multiple signalling pathways involved in phenotypic switching in *C. albicans*. Figure adapted from Biswas *et al.*, (2007).

1.1.2 White-Opaque Switching and Mating

A reversible switch from a white to opaque form is an additional phenotypic characteristic seen in certain strains of *C. albicans* on agar at 25 °C, which was first observed in the strain WO-1 (Slutsky *et al.*, 1987). When observed as a colony, the most common phase is a white, smooth dome. Upon microscopic examination these cells appear as normal unicellular yeasts. However after

switching to opaque phase the colony appears grey in colour, often flatter and the cells themselves are elongated and larger (Figure 1.4) (Slutsky *et al.*, 1987) (Kim and Sudbery, 2011). Other differences between the phases include length of generation time, sensitivity to temperature, and an inability to form hyphae in the opaque phase (Slutsky *et al.*, 1987).

Interestingly mating in *C. albicans* was first demonstrated both *in vivo* (Hull et al., 2000) and *in vitro* (Magee and Magee, 2000) using strains that undergo white-opaque switching. *C. albicans* contains a mating-type like (*MTL)* locus which contains orthologues of *S. cerevisiae* mating genes alongside other genes whose functions have diverged. In *C. albicans* two distinct alleles of *MTL* (a and α) are present in either homozygous or heterozygous combinations, reminiscent of a/α strains in *S. cerevisiae.*

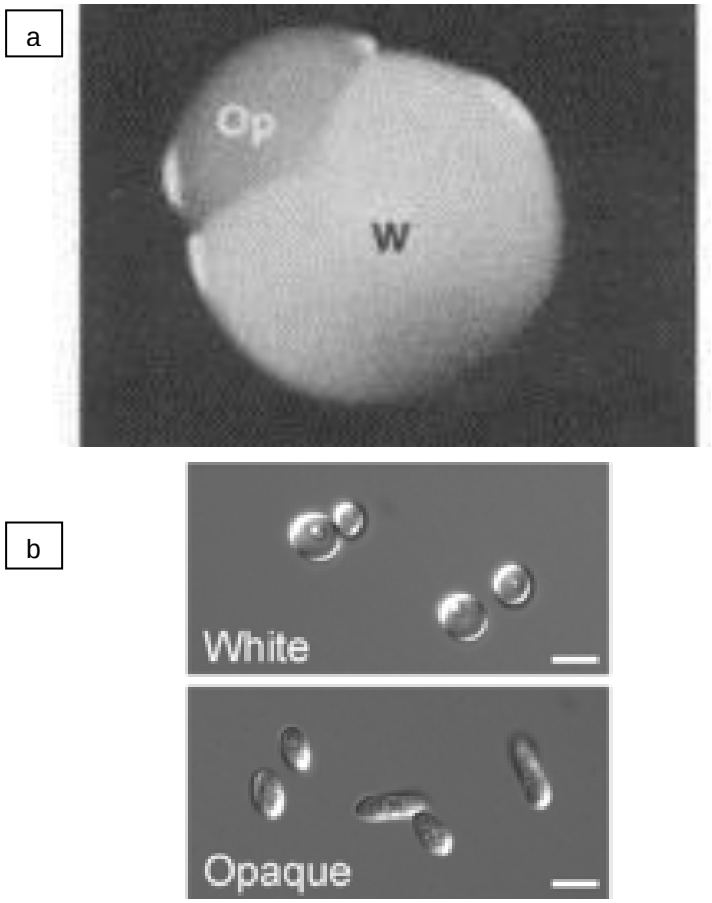


Figure 1.4 Examples of colony and cell morphology in the white-opaque transition from strain WO-1. a) Opaque sector (Op) from a white colony (W). b) Differing cell morphologies with white phase cells showing the common yeast form whereas opaque cells appear elongated. Figures adapted from (Slutsky *et al.*, 1987) and (Zordan *et al.*, 2007).

In an examination of 120 clinical isolates, it was found that the heterozygous strain is the most frequent at 108, with just 12 homozygous strains: seven MTLa and five MTLα (Legrand *et al.*, 2004). Strains homozygous for *MTL,* such as WO-1, undergo white-opaque switching.

It has been elucidated that this switch in appearance is under the control of the gene *WOR1.* In heterozygous strains, the dimer formed by the a1/α2 gene products represses the expression of *WOR1* preventing the switch to opaque phase (Kim and Sudbery, 2011). In homozygous strains this repression is lifted, allowing a shift in phenotype to an opaque phase which is concurrent with the strain's ability to mate through a parasexual cycle. Opaque cells of opposite mating-type form long conjugation tubes which fuse to form tetraploid cells. However unlike a conventional sexual cycle, nuclear fusion does not happen. From the bridge formed by the joined conjugation tubes a new daughter cell is formed. Again as opposed to conventional meiosis, the two nuclei divide asynchronously with one or two nuclei moving to the new daughter cell and one or two nuclei returning to the mother cell (Figure 1.5) (Johnson, 2003). This is then followed by a process of random chromosome loss to return cells to a diploid, or near diploid, state (Bennett and Johnson, 2003).

As this process of concerted chromosome loss is random, formation of genetic diversity can still be achieved despite a lack of meiosis. The strains resulting from this process can contain chromosomes from either parental strain and are often associated with differing levels of aneuploidy. Further genetic diversity is also achieved through homologous recombination of the chromosomes during the parasexual cycle, under the control of the re-programmed meiosis gene *SPO11,* resulting in gene conversion events (Forche *et al.*, 2008). Although this process has been observed under *in vivo* conditions (Lachke *et al.*, 2003, Dumitru *et al.*, 2007), it is a rare event, explaining why *Candida albicans* is found naturally in genetically distinct clades. This is further discussed later in section 1.2.

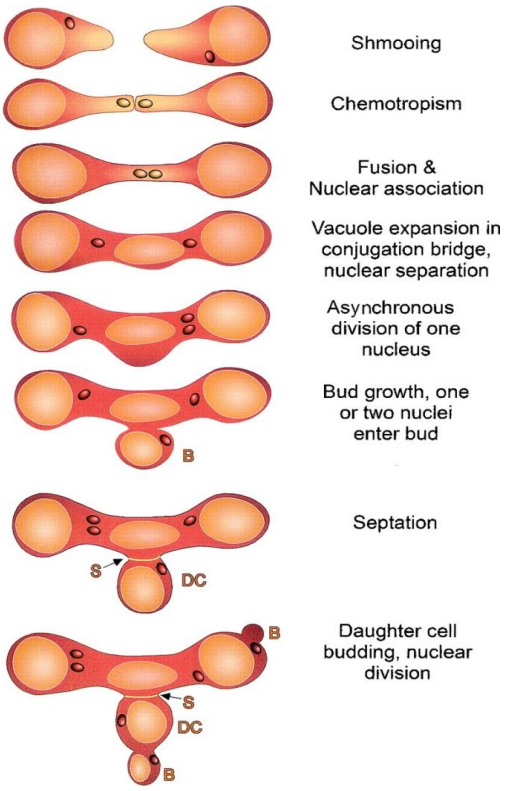


Figure 1.5 The model of nuclear dynamics during the parasexual cycle of *Candida albicans*. Vacuoles are light brown, and nuclei are dark brown. B, bud; S, septum; DC, daughter cell. Figure adapted from Lockhart *et al.,* (2003).

1.1.3 Resistance to Antifungal Drugs

Resistance to antifungal drug treatments is commonly reported in *Candida* species, especially with regards to long-term use of azole-based therapeutics such as fluconazole (Odds, 1993, Rex *et al.*, 1995). Azoles act as a fungistatic drug inhibiting synthesis of an integral constituent of the cell membrane, ergosterol. Most azoles act via inhibition of the 14α-demethylase enzyme, however some earlier azole derivatives such as ketoconazole have a more complex mode of action against various membrane-bound enzymes. The subsequent increase in sterol precursors modifies the structure and function of the plasma membrane (Ghannoum and Rice, 1999). Resistance to azoles has largely been attributed to the activity of efflux transporters, such as the ATP-binding cassette multidrug transporter genes *CDR1* and *CDR2,* which prevent the accumulation of azoles in the cell (Sanglard *et al.*, 1995). Interactions of *trans* modifiers with these genes also further complicate the mechanism of antifungal drug resistance. The transcription factor, *TAC1,* up regulates expression of these ABC-transporter genes, with varying levels of efficacy

dependent upon the sequence of the *TAC1* gene. Azole resistant strains have been shown to have a more active variant of *TAC1* (Coste *et al.*, 2006). Other mechanisms of azole resistance have also been reported including modifications of the affinity of 14α-demethylase (White, 1997). Fluconazole resistance is most commonly reported in *C. albicans* with resistance to other azole derivatives, such as ketoconazole and itraconazole, more commonly reported in related *Candida* species such as *C. glabrata* and *C. tropicalis* (Odds, 1993, Pfaller and Diekema, 2007).

Antifungal resistance isn't restricted to azoles; resistance to the fluorinated pyrimidine, 5-flucytosine, has also been reported. Upon entrance to the cell, 5-flucytosine is converted to 5-fluorouracil. From here, 5-fluorouracil can operate in two separate manners. 5-fluorouracil is converted to 5-fluorouridylic acid (FUMP). FUMP undergoes further phosphorylation and is incorporated within RNA resulting in inhibition of macromolecular synthesis (Ghannoum and Rice, 1999). Alternatively, 5-fluorouracil can also be converted to 5-fluorodeoxyuridine monophosphate, which inhibits the biosynthesis of DNA through blocking production of thymidine (Vermes *et al.*, 2000). Resistance is attributed to mutations in the enzymes of the pyrimidine salvage pathway which convert 5-flucytosine to FUMP (Hope *et al.*, 2004). Other examples of antifungal resistance, such as resistance to both the fungicidal polyene drug, amphotericin B, and the echinocandin antifungal drugs, such as caspofungin, are relatively rare in *C. albicans* but decreased susceptibility has been reported in *C. glabrata* and *C. krusei* for amphotericin B and in *C. guilliermondii* and *C. parapsilosis* for echinocandins (Pfaller and Diekema, 2007). Treatment with antibiotics has also been linked to a significantly increased infection risk, especially in the case of *C. glabrata* infections (Pfaller and Diekema, 2007). This has been attributed to the alteration of bowel flora and an increase of *Candida* growth in intestinal tracts (Jarvis, 1995).

1.1.4 Biofilm Formation

In a clinical setting, both superficial and systemic *Candida albicans* infections can be found in the form of a biofilm. Most commonly these infections are associated with chronic in-dwelling devices such as catheters and dentures. Mortality rates in patients with catheter-associated candidiasis are higher than

that seen for disseminated candidiasis at 41% compared to 34% (Nguyen *et al.*, 1995).

Biofilm formation occurs in three distinct stages termed early, intermediate, and maturation (Chandra *et al.*, 2001). At the early stage, primarily yeast cells and some hyphal cells are present which begin binding to the surface and forming microcolonies. The intermediate stage is characterised by the increase in noncellular material, similar to the polysaccharides found in the cell wall, which form a haze-like film over the microcolonies; and the maturation stage sees this film developing to encase the *C. albicans* cells (Chandra *et al.*, 2001, Ramage *et al.*, 2001).

Several transcription factors which control formation of biofilms have previously been identified; *BCR1* and the hyphal regulator *TEC1* have been shown to genetically interact and influence biofilm formation (Nobile and Mitchell, 2005); and the hyphal regulatory gene *EFG1* has also been identified as an important regulator (Ramage *et al.*, 2002). These studies all underpin the importance of the morphological switch to hyphae during the process of biofilm formation. In 2012, a complete network of regulatory genes responsible for control of biofilm formation was identified through transcriptional studies of homozygous knockout mutants of transcription factors which are unable to form biofilms (Nobile *et al.*, 2012). Six main regulatory genes were identified including *BCR1, TEC1, EFG1* and the newly identified regulators *NDT80, ROB1* and *BRG1.* Together, these transcription factors impact upon the expression of over a 1000 target genes, with eight core genes being identified us up-regulated by all six regulators, including the adhesion gene *ALS1* and the hyphal cell wall gene *HWP1* (Nobile *et al.*, 2012).

Treatment of biofilm infections is often complicated by various factors. For example denture stomatitis is a superficial infection attributed to biofilm formation on dentures. Even with antifungal treatment, the infection is persistence and is often re-established quickly after cessation of treatment (Chandra *et al.*, 2001). Although the suggested form of treatment is removal of the in-dwelling device, especially in the case of catheters, this option isn't always possible leaving treatment with antifungal therapies the only choice

(Chandra *et al.*, 2012). Antifungal resistance of biofilm infections has also been identified as problematic (Chandra *et al.*, 2012), with some evidence suggesting that resistance levels, particularly to azole based treatments, is significantly higher in biofilms when compared to planktonic cells with MICs ranging from 1 µg/ml in planktonic cells to 128 µg/ml in biofilms (Chandra *et al.*, 2001). The mechanisms of resistance change as the biofilm develops, with early phase resistance being attributed to efflux pumps, similar to resistance in planktonic cells, whereas mature biofilm resistance has been shown to be due to reductions in the production of ergosterol intermediates (Mukherjee *et al.*, 2003). Drug resistance is also not restricted to a single group of antifungal drugs; observations of resistance have been seen with exposure to fluconazole and related azoles, 5-flucytosine, amphotericin B (Hawser and Douglas, 1995, Ramage *et al.*, 2001), nystatin and chlorhexidine (Chandra *et al.*, 2001). Research is now being undertaken to investigate the efficacy of drug treatments used in combination with each other and with other agents that target proteins such as cyclosporine A and FK 506, which target the protein phosphatase, calcineurin (Uppuluri *et al.*, 2008, Shinde *et al.*, 2012), and geldanamycin which targets the heat shock protein Hsp90 (Robbins *et al.*, 2011). Recently, it has been shown that the efficacy of fluconazole against *Candida tropicalis* is increased by addition of flavonoids such as catechin and quercetin (da Silva *et al.*, 2013).

1.2 Phylogeny and Clades

Candida albicans falls into the Ascomycota phylum of the fungal kingdom, and then under the sub-phylum Saccharomycotina (otherwise known as the hemiascomycetes) (Scannell *et al.*, 2007). The Saccharomycotina can be split into three “clusters”; *C. albicans* forms part of the *Candida* clade. This consists of closely related infective *Candida* species including *C. parapsilosis, C. tropicalis* and *C. guilliermondii* (Figure 1.6). A distinctive feature shared by all species within the *Candida* clade is the alternative use of the CUG codon, which codes for serine instead of leucine (Butler *et al.*, 2009). However other traits are not well conserved within the clade. The level of ploidy varies with some diploid organisms: *C. albicans, C. dubliniensis, C. tropicalis, C. parapsilosis* and *Lodderomyces elongisporus,* and some haploid organisms: *C. guilliermondii, C. lusitaniae* and *Debaryomyces hansenii* (Butler *et al.*, 2009). The ability to mate

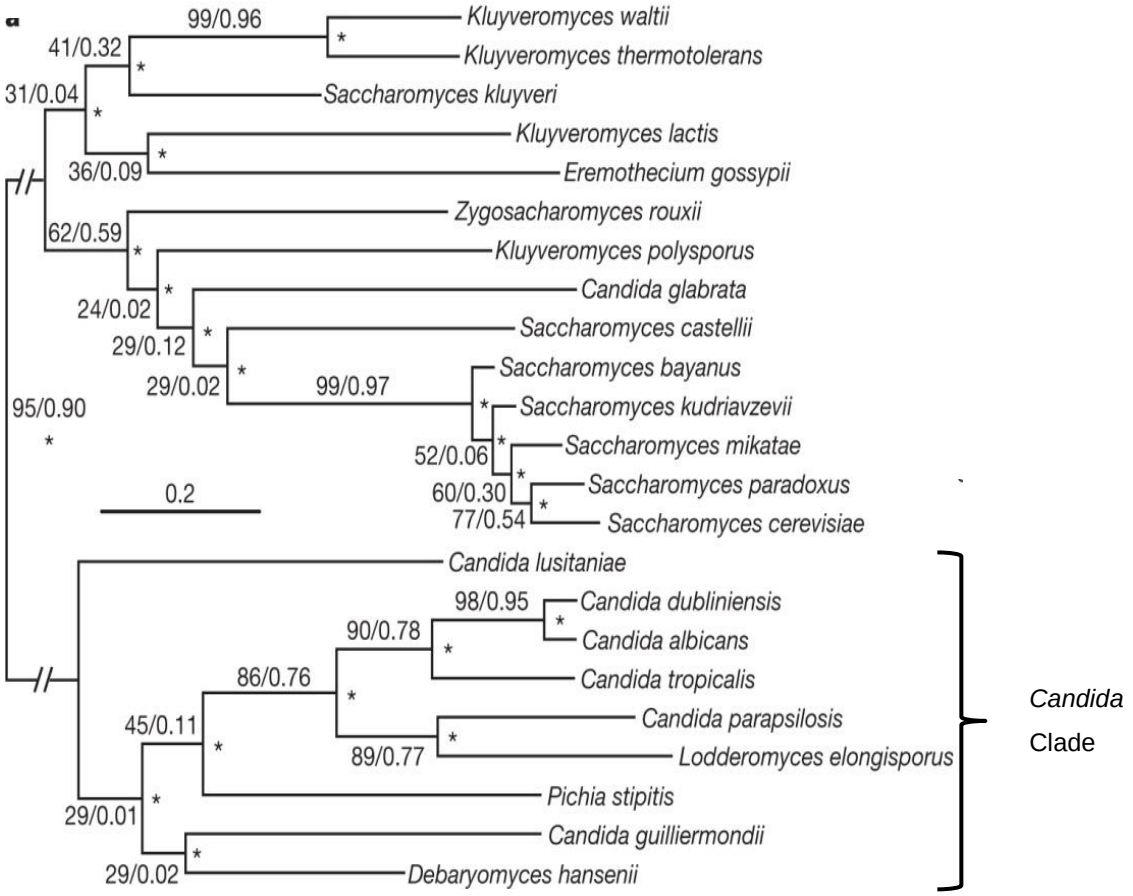


Figure 1.6. Phylogenetic relationships among the sequenced fungal genomes. The yeast species phylogeny recovered from concatenation analysis of 1,070 genes using maximum likelihood. Asterisks denote internodes that received 100% bootstrap support by the concatenation analysis. Values near internodes correspond to gene-support frequency and internode certainty, respectively. The scale bar is in units of amino-acid substitutions per site. Figure adapted from (Salichos and Rokas, 2013).

is also not well conserved; *C. albicans* is primarily a clonal organism with an infrequent parasexual cycle (see section 1.1.2). A similar mating behaviour is observed in *C. tropicalis* (Porman *et al.*, 2011). However *L. elongisporus* is a sexual homothallic organism (Butler *et al.*, 2009). Although *C. glabrata* has historically been named as a *Candida* species and causes a highly similar infection, this species is in fact very distantly related to the other *Candida* species with a genome far more similar to *Saccharomyces cerevisiae.* Interestingly though, unlike *S. cerevisiae, C. glabrata* is also a clonal organism with no known sexual cycle.

C. albicans strains are generally grouped into genetically distinct groups historically termed clades. Different typing methods have been used for this analysis including geographical origins (Odds and Jacobsen, 2008), restriction fragment length polymorphism analysis (A, B, C genotypes) (Scherer and Stevens, 1987, Xu *et al.*, 1999), the presence of an intron within the ribosomal DNA gene *ITS1* (McCullough *et al.*, 1999) and MLST (Odds and Jacobsen, 2008). However the most commonly used criteria for grouping isolates is Ca3 fingerprinting. Five separate genetically distinct groups have been identified; I, II, III (Pujol *et al.*, 1997), SA (Blignaut *et al.*, 2002) and E (Pujol *et al.*, 2002). Different typing methods produce results with some similarities of groupings and some differences (Odds *et al.*, 2007). Although all genetic groups are equally successful as pathogens (Soll, 2002), they have been shown to differ in many phenotypic characteristics including growth rate and growth under salt stress (MacCallum *et al.*, 2009). These phenotypic differences are elucidated to be due to differences in gene sequences. for example a non-synonymous SNP in the *FUR1* gene confers resistance to 5-flucytosine in clade I (Pujol *et al.*, 2004, Odds *et al.*, 2007).

1.3 Genome Architecture

The full diploid genome sequence for *Candida albicans* was completed in 2004 (Jones *et al.*, 2004). Although haploid genome sequencing is commonly carried out on many organisms, the diploid genome has only been assembled for a handful of organisms, mostly humans (Levy *et al.*, 2007, Wang *et al.*, 2008), making *C. albicans* an important model organism for diploid genetics.

Sequencing of the lab strain SC5314 revealed that the haploid genome is 14851 kb in length, containing 6419 open reading frames over 100 codons in length, distributed across eight chromosomes. A gene density of one ORF every 2342 bp is observed. Splicing has been identified in *C. albicans* and therefore 224 genes were identified as containing introns, with most located at 5’ ends (Braun *et al.*, 2005). Although some variations are observed, these results are similar to those identified in other closely related species from the *Candida* group (Butler *et al.*, 2009). 20% of the ORFs identified have no counterpart in other genome sequences (Odds *et al.*, 2004). These *C. albicans*

specific genes may be of particular interest during investigation of pathogenesis-related characteristics only seen in *C. albicans.*

Genes which confer similar functions or have similar sequences are grouped into gene families. In total 23% of all genes were identified as members of a gene family, with 451 families identified computationally (Braun *et al.*, 2005). Virulence-associated gene families include the secreted aspartic proteinase (*SAP)* family (Monod *et al.*, 1994), and the agglutinin-like sequence (*ALS)* family which are involved in adhesion to the host surface (Hoyer, 2001).

1.4 Diploid Nature, Heterozygosity and Divergence in Allele Function

Initially, *Candida albicans* was thought to be a haploid species (Whelan and Magee, 1981). Preliminary studies which suggested that *C. albicans* is a diploid species with heterozygosity used UV exposure and auxotrophic markers to identify strains heterozygous for genes which produce amino acids such as methionine andcysteine, and genes coding for the nucleobase, adenine. At the time it was unclear whether the species was truly diploid or if aneuploidy was present (Whelan *et al.*, 1980, Whelan and Magee, 1981).

As mentioned in section 1.3, the full diploid genome sequence for *Candida albicans* was completed in 2004. The results indicated a high level of polymorphism at a frequency of one SNP approximately every 237 base pairs. This heterozygosity is not distributed evenly across the eight chromosomes, with high levels seen on chromosome five and six covering genes such as the mating type like-locus and the *ALS* gene family. Conversely regions of low heterozygosity/near homozygosity are also present on chromosome three and chromosome seven. These polymorphisms between alleles are not restricted to SNPs, with various numbers of tandem repeat sequences also observed (Jones *et al.*, 2004).

This level of heterozygosity is much higher than that observed in the most closely related species from the *Candida* clade. *Lodderomyces elongisporus* has a similar rate of polymorphisms with one SNP seen every 222 bp, but *C. tropicalis* has one SNP every 576 bp and *C. parapsilosis* has the lowest rate with one SNP observed every 15,553 bp (Butler *et al.*, 2009). This is also higher

than the levels of heterozygosity observed in the human genome (Jones *et al.*, 2004).

Although the wild-type strain SC5314 remains the most widely studied strain of *C. albicans,* investigation of other clinical isolates and strains has shown that a high number of different alleles are present for genes. For example, differing numbers of microsatellites are found in the promoter regions of elongation factor 3 (*EF3*). Each of these differences were identified as a different allele and across 29 reference strains eight alleles were discovered (Bretagne *et al.*, 1997).

1.4.1 Functional Polymorphisms

Due to the high levels of heterozygosity reported in *C. albicans,* SNPs are often observed within coding regions of genes. As a consequence of this, there have been numerous reports of alleles of a single gene differing in function. This concept is a key idea being investigated throughout this work, with the suggestion that both proteins produced by an allele, and therefore that heterozygosity, is fundamental to phenotypes involved in pathogenicity and virulence.

A study using the clinical isolate CA12 found the adenine gene *ADE2* has two alleles, one functional and one non-functional. The non-functional allele has a 1.3 kb deletion which spans the promoter and coding region. Homozygosity of this non-functional allele results in an auxotrophic strain which can no longer produce adenine (Tsang *et al.*, 1999). It is unclear how this clinical isolate, obtained from oral infections of HIV patients, relates to the wild-type strain SC5314, but demonstrates that variability in allele function occurs *in vivo*. In SC5314, a similar case of functional differences in auxotrophic genes is seen in the histidine gene *HIS4* where a single SNP renders allele one inactive (Gómez-Raja *et al.*, 2008).

A number of genes from the *ALS* family span a region of high heterozygosity on chromosome six. These genes encode cell-wall glycoproteins which are involved in adhesion to host cell surfaces (Hoyer, 2001). This family of genes has been reported to have variability in the sequences of alleles, especially

within repeated regions. *ALS9* was discovered to have divergence in allele sequences, both within and outside of the repeated regions. The majority of these differences lie within the N terminal domain and result in amino acid sequences of the two alleles being only 87% identical (Zhao *et al.*, 2003). Mutants of *ALS9* have shown that only allele two can complement function. This suggests that the sequence differences between the alleles leads to differences in function (Zhao *et al.*, 2007).

Variability in allele function has been reported in numerous examples relating to antifungal drug resistance. *TAC1* is the transcription factor responsible for up-regulation of the ABC-transporter genes *CDR1* and *CDR2* (Coste *et al.*, 2006). These genes act as multi-drug transporters and up-regulation confers resistance to azole based antifungals (see section 1.1.3). Investigation of both azole resistant and azole susceptible clinical isolates of *C. albicans* revealed that the alleles of *TAC1* differ in sequence and function impacting upon the resistance status of the isolate. A single amino acid change from asparagine to aspartic acid in the C-terminal activation domain creates a hyperactive allele. Azole resistant strains occur when this allele is homozygous and the *CDR* genes are therefore constitutively expressed (Coste *et al.*, 2006). Interestingly, it has been shown that the *CDR* genes themselves also have differential allele functions. Over-expression of each *CDR2* allele separately showed that allele two has better pumping efficiency than allele one conferring to differences in susceptibility to azoles (Holmes *et al.*, 2006).

A single amino acid mutation in *ERG16,* encoding the lanosterol 14α-demethylase, results in a decreased affinity of the enzyme for azoles and confers resistance to azole drug treatment. Resistance is associated with “loss of allelic variation” with strains homozygous for the mutation demonstrating the most resistance (White, 1997).

Similar results have been found with regards to resistance to the antifungal 5-flucytosine. Homozygous mutations in the pyrimidine salvage pathway enzyme, uracil phosphoribosyltransferase, which lead to an arginine residue becoming a cysteine residue, have been shown to confer resistance. When this mutation is seen in a heterozygous strain, an intermediate phenotype is observed with a

reduced susceptibility to 5-flucytosine but not complete resistance (Dodgson *et al.*, 2004, Hope *et al.*, 2004).

Three genes are present in the *MTL*-locus in *C. albicans* which are not present in the mating type locus in *S. cerevisiae*: a gene for poly(A) polymerase, a gene for an oxysterol binding protein, and a gene for phosphoinositol kinase. Due to high levels of heterozygosity across the *MTL*-locus, the amino acid similarity between the two alleles of these genes is low with an average of just 60%. Therefore different *MTL*-loci might encode genes with different functions (Johnson, 2003), although this observation is currently speculative.

The impact of heterozygosity on the phenotype of *C. albicans* is apparent from the above examples, but conversely homozygosity is also suggested to have important functions. Despite the existence of a parasexual cycle, *C. albicans* is primarily a clonal organism; recombination doesn't occur during meiosis as in other sexual species. A potential mechanism to introduce genetic variation in *C. albicans* is mitotic recombination leading to loss of heterozygosity (LOH).

LOH has been shown to readily occur at the *GAL1* locus in SC5314 during passage through a mouse infection model (Forche *et al.*, 2005) and LOH also occurs when cells are exposed to *in vitro* stress conditions (Forche *et al.*, 2011). Genetic variation produced by LOH events may be used as a mechanism to adapt to a changing host environment. Conversely large regions of LOH may actually be detrimental to fitness and cause increased doubling times (Abbey *et al.*, 2011).

Recent work into *C. albicans* mating has disproved the long standing idea that the species is an obligate diploid. Rare haploid strains have been identified both *in vitro* and *in vivo.* The exact mechanism under which the haploid cells are produced is still unclear; conventional meiosis has been ruled out as chromosomes appear to be from one diploid parent with very little crossover. A possible suggested mechanism is concerted chromosome loss, similar to that seen during the parasexual cycle (Hickman *et al.*, 2013).

As haploid strains are produced, this paper initially disproves the theories of heterozygosity being essential to function. This finding is surprising, as it has been believed that the diploid state of cells may be masking lethal mutations. However certain homologue biases were observed suggesting that some alleles are favoured over others in the haploid state. Haploid strains also have reduced fitness and reduced cell size. Interestingly these phenotypes were also observed in homozygous diploid strains which were produced from auto-diploidisation of haploid cells indicating that not just the diploid state but heterozygosity itself is important. Fitness levels were restored when haploid strains were mated (through a parasexual cycle) to form heterozygous diploid strains (Hickman *et al.*, 2013).

1.5 Differences in Allele Expression

Following on from the idea that alleles may differ in function is the idea that alleles of a single gene may differ in expression levels. This phenomenon can include situations where one allele of a gene is not expressed (monoallelic expression) or the expression levels of alleles are uneven. Commonly these situations are referred to as allelic expression imbalance (AEI). There are a small number of examples of AEI in *Candida albicans* which will be discussed at the end of this section. More generally AEI is widely reported in higher eukaryotic species which will be the main focus of the section.

1.5.1 Monoallelic Expression, Imprinting and X Chromosome Inactivation

Monoallelic expression is an area which has been extensively investigated in higher eukaryotes. It is observed in a number of forms and is generally characterised by one allele in a pair being subjected to silencing.

1.5.1.1 Monoallelic Expression

Monoallelic expression within autosomal cells is where the paternal allele is silenced in some cells, whereas in other cells of the same type, the maternal allele is silenced. This is termed autosomal monoallelic expression (MAE) and has been observed across a large fraction of cell populations. One of the earliest and most widely reported cases of monoallelic expression is seen in the olfactory system in humans. Over 1000 genes are present for odorant receptors but only a single receptor is expressed per neuron. To achieve this fine-tuned

selection of expression, single allele expression and random *cis*-regulation are employed (Chess *et al.*, 1994). A similar system of monoallelic expression is also reported in receptors upon the surface of T cells, where again only a single receptor is expressed per cell.

Despite these two examples of monoallelic expression being within receptor cells, it is a phenomenon that is not restricted to specific cell types. It is also seen across a wide range of formal gene categories (ontologies), but notably there is an excess of genes which encode cell surface markers and thus contribute to unique cellular identities (Chess, 2012).

Monoallelic expression has been seen to have functional consequences through haploinsufficiency. In mice, the toll-like receptor 4 gene (*Tlr4*) is expressed from just one allele in lymphocytes. Heterozygous mice have two types of lymphocyte cells, ones which express a functional copy and therefore respond to lipopolysaccharides, and others which express the non-functional copy and are therefore not responsive, causing the mice to be more sensitive to infections by Gram-negative bacteria (Pereira *et al.*, 2003).

As well as humans and mice, monoallelic expression has been found in more simple eukaryotes species as reviewed by Borst and Chaves (1999). Two examples based upon antigenic variation include the *var* gene in the malarial species *Plasmodium falciparum* and surface glycoproteins in *Trypanosoma brucei. P. falciparum* have around 50 *var* genes, responsible for antigenic variation. To ensure that only one *var* gene is expressed per cell, a system of selective transcription similar to that seen in the olfactory system in humans is adopted (Scherf *et al.*, 1998). *T. brucei* encode a large number of surface glycoproteins but again only one is expressed per cell. This is achieved by some alleles containing a modified DNA base, beta-D-glucosyl-hydroxymethyluracil or J, rendering some inactive and others active (van Leeuwen *et al.*, 1997).

1.5.1.2 Imprinting

Parent of origin specific monoallelic expression, also known as imprinting, occurs when just one allele is expressed based on the parental origin. Currently

there are around 50 genes in the human genome which are classically defined as imprinted (Morison *et al.*, 2005). A commonly cited example of imprinting in the human genome is the *H19/Igf2* locus. The *H19/Igf2* locus is made up of two genes which are 7 kb apart on chromosome 11p15.5. Each gene is expressed from opposite chromosomes. Expression is under control of a region of the *H19* promoter which contains differentially methylated regions (DMRs). *H19* is expressed from the unmethylated maternal allele. Lack of methylation allows a zinc finger protein, CTCF, to bind and suppress expression of *Igf2* and activate expression of *H19.* Whereas on the paternal allele, the *H19* promoter is methylated, causing silencing of *H19* and prevention of binding of CTCF leading to *Igf2* expression (Pidsley *et al.*, 2012) (Figure 1.7).

A more complex example of imprinting in mice is *Mest.* Here only the paternal allele is active, with differential methylation of the promoter CpG island rendering the maternal allele silent (Figure 1.8a). Expression of an additional longer form of the gene, termed *MestXL,* has been found in cells of the central nervous system. The longer RNA occurs due to alternative polyadenylation. Imprinting in these cells is then further complicated by the longer *MestXL* transcript. As well as *Mest* being only transcribed from the paternal allele,

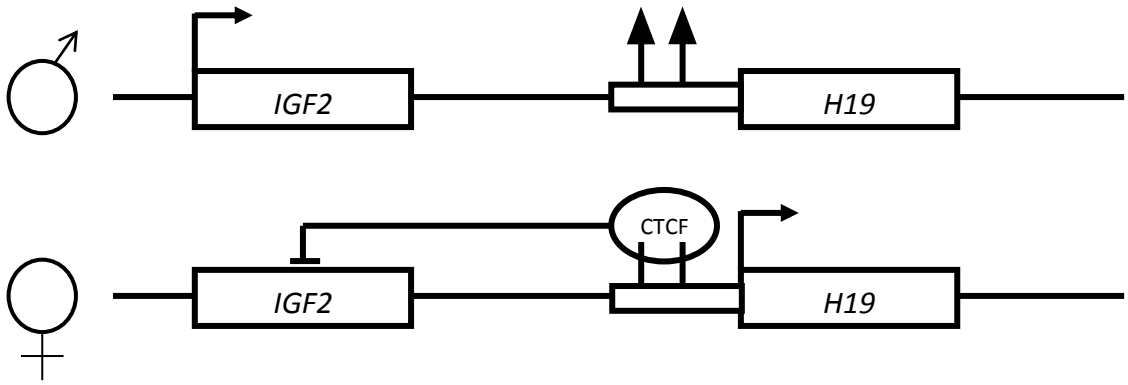


Figure 1.7 A simplified schematic map of the human *H19/Igf2* locus on chromosome 11p15.5, with paternal allele (♂) on the top and maternal allele (♀) on the bottom. Expression of *H19* is repressed in the paternal allele due to methylation (filled black triangles), allowing activation of *Igf2*. Conversely expression of *H19* is active in the unmethylated maternal allele, allowing CTCF to bind and repress *Igf2*.

MestXL overlaps with the RNA of the adjacent antisense gene *Copg2* causing paternal suppression via transcriptional interference, leaving *Copg2* expression only from the maternal allele (Figure 1.8b) (MacIsaac *et al.*, 2011).

Imprinting on a genome-wide scale in eukaryotes has been investigated in mouse brains using RNA sequencing. In total, divergence in allele expression levels was found at 1308 loci, with the expressed allele showing inconsistencies across different developmental stages (Gregg *et al.*, 2010). A conclusion from this study was that imprinting wasn't solely due to monoallelic expression and that AEI also has a role to play. However, DeVeale *et al.,* (2012) argued that the results found by Gregg *et al.,* (2010) were a massive over representation of monoallelic expression and they could not be repeated. False positives were

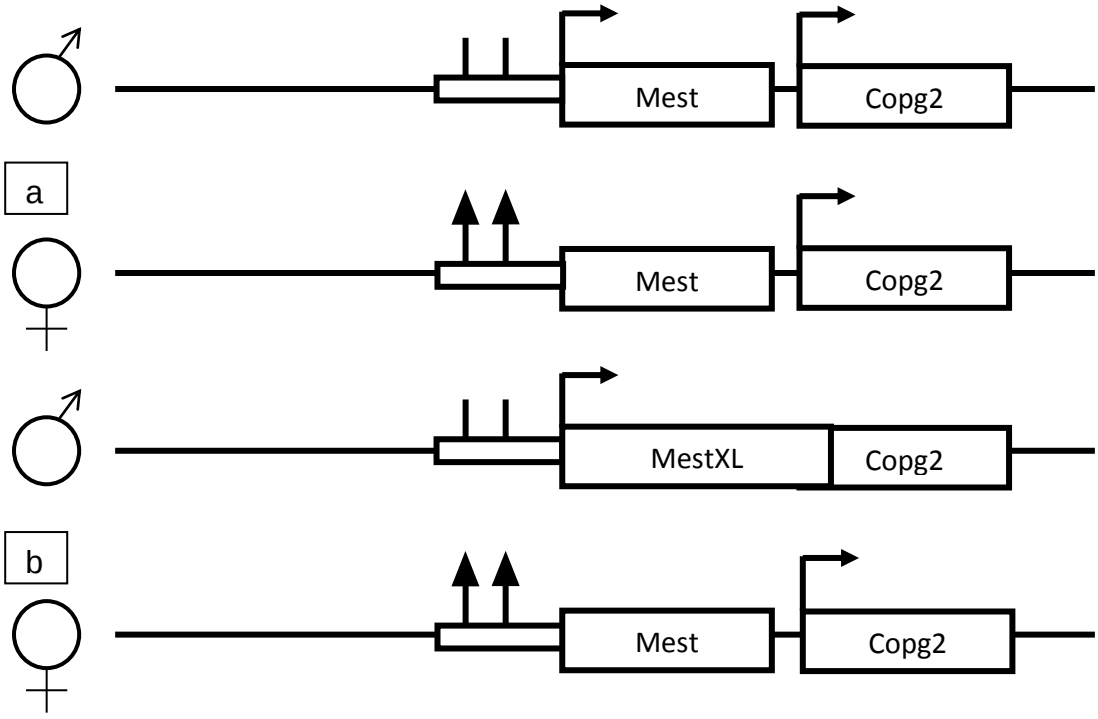


Figure 1.8 A simplified schematic map of the mouse *Mest* locus. a) Shows "normal" regulation with paternal allele (♂) on the top being expressed and maternal allele (♀) on the bottom being silenced due to methylation (filled black triangles). *Copg2* is expressed from both alleles. b) Shows expression of the longer *MestXL* transcript from the paternal allele (♂) leading to repression of *Copg2.* The maternal allele (♀) has *Mest* silenced but *Copg2* is expressed.

attributed to technical and biological variation as well as systemic errors in the sequencing (DeVeale *et al.*, 2012).

Imprinting has also been identified on a genome-wide scale in various plant species. RNA sequencing has been used to examine triploid maize endosperm to identify genes with at least a fivefold difference between the two paternal alleles and single maternal allele. On this basis, a total of 179 genes were identified as imprinted (Zhang *et al.*, 2011). However this study only identified imprinting in early endosperm and suggested that this was lost after 12 days post pollination and therefore cannot be used to determine imprinting levels in all cell types and at all life cycle stages.

1.5.1.3 X-chromosome Inactivation

X-chromosome inactivation is adopted by female mammals to prevent gene dosage effects caused by females carrying a second X chromosome. Heteropyknosis (condensation) of one X chromosome occurs in female cells, resulting in gene inactivation across that chromosome (Lyon, 1961). Therefore this mechanism could be seen as large scale monoallelic expression. The process of lyonization or silencing renders one X chromosome active and condenses the second X chromosome into a transcriptionally silent structure known as a Barr body. This silencing is achieved using the non-coding RNA *XIST* which binds to the inactive X chromosome, recruiting the Polycomb repressive complex 2 (PRC2) which triggers inactivation (Simon *et al.*, 2013). The exact mechanisms of silencing are still unclear but suggestions have been made that both DNA methylation and histone deacetylation are involved (Boumil and Lee, 2001). In most mammals, the silenced X-chromosome is determined randomly in each cell during embryo development, with approximately half of cells inactivating the paternal X and the other half inactivating the maternal X. This can lead to differential expression of phenotypes across an organism, accounting for phenomena such as the coat colour mosaicism seen in tortoiseshell cats.

X-chromosome inactivation has been shown to have functional consequences on cell phenotype. In women who are heterozygous for the X-linked tumour suppressor gene *FOXP3,* breast cancer arises from cells where the X-

chromosome containing the functional copy of *FOXP3* is inactivated, leaving expression of a non-functional allele (Zuo *et al.*, 2007).

1.5.2 Genome-Wide Allelic Expression Imbalance

AEI is not restricted to imprinted genes or to cases of monoallelic expression. There are now a large number of studies showing that both alleles of a gene can have significantly different levels of expression. The recent development of microarrays and sequencing technologies has allowed for the identification of genes with significant AEI on a genome-wide scale.

For example, Lo *et al.* (2003) used microarray analysis based upon known SNP locations to investigate the variability of human allele expression. Here it was demonstrated that at least 326 genes in the human genome showed preference for the expression of one allele. Similar studies in humans include use of the Illumina Allele-specific Expression BeadArray platform which identified that 20% of 1380 genes exhibit AEI across 2968 SNPS (Serre *et al.*, 2008), and use of the Affymetrix Human Mapping 500 K array set revealed that 2.2% of genes have monoallelic expression across multiple SNPs (Gimelbrant *et al.*, 2007). Outside of humans, a study using SNP chip arrays with *Drosophila simulans*, found AEI in 37% of the probe sets tested (Yang *et al.*, 2011).

RNA sequencing has also been used to identify genes with AEI on a genome-wide scale in humans. Using CD4+ T cells, it was found that 4.6% of heterozygous SNP-sample pairs have evidence of imbalance in allele expression levels, which was validated for four genes using bacterial cloning (Heap *et al.*, 2010). Other methods employed to identify genes with AEI in humans include use of expression sequence tags (Ge *et al.*, 2005), and use of allele-specific serial analysis of gene expression where 25% of human genes were identified as having AEI (Vidal *et al.*, 2011). These studies don't discuss in detail the functional consequences of this AEI, but instead focus upon the increased complexity of gene expression observed.

Contrary to the above examples in humans, (Yan *et al.*, 2002b) developed an RT-PCR based method which could detect a 20% difference in allele expression. Although the study only investigated 13 genes, expression levels

were assessed across 96 individuals with this RT-PCR method and very little AEI was identified. Interestingly, a genome-wide study using chip based methods identified only five genes with monoallelic expression and a further 125 genes with AEI, but did not identify any known imprinted genes (Song *et al.*, 2011).

It should be noted here that all of these papers identified different percentages of genes with AEI. A review by Pastinen *et al.* (2006) found that correlations between different studies of AEI were low, possibly due to differential probe location in array-based studies. This demonstrates the lack of consistency in the methods used. In most cases, different cell types or cell lines have been used, adding a level of uncertainty. There are also inconsistencies in the reference genome used across studies. The downstream bioinformatic analysis of expression lacks uniformity, with the criteria for a gene with AEI often differing. Teare *et al.* (2011) discusses the discrepancies between different methods used to identify AEI, including restriction fragment length polymorphism, RT PCR, differential hybridisation to oligo arrays and RNA sequencing. However a study by Cheung *et al.* (2010) looking at the regulation of gene expression identified AEI using both microarrays and RNA sequencing and found a significant strong positive correlation between the methods (Cheung *et al.*, 2010).

AEI has also been exploited as an alternative method to quantitative trait loci (QTL) in human genetic studies for use in identifying *cis*-acting polymorphisms and regions of regulatory DNA which cause genetic variation (Teare *et al.*, 2011). This method works on the principle that alleles with expression imbalance must differ by at least one regulatory SNP which is in linkage disequilibrium. However this method requires knowledge of haplotypes and therefore can be prone to phasing errors, especially where regulatory regions are far from the transcript. A study by Lefebvre *et al.* (2012) developed a genotype-based method to negate the need for haplotype information in these type of studies and improved reliability of results for regulatory regions with increasing distances from the transcript (Lefebvre *et al.*, 2012). A general comment about using AEI in this way is that the method assumes that the

disparity in expression of alleles is solely due to *cis*-acting polymorphisms and that no other elements, such as differences in genome structure, contribute.

Plants have been used as model organisms for genome-wide identification of AEI with further discussion upon functional impact. The highly heterozygous maize *Zea mays* showed AEI in 11 of 15 genes analysed using a method which combined RT-PCR and denaturing HPLC. This study linked AEI to the functional response to abiotic stresses such as drought and density (Guo *et al.*, 2004). Similar results have been seen in the tomato species *Solanum peruvianum* and *Solanum chilense* where AEI was suggested to have a role in adaptation to abiotic stress (Mboup *et al.*, 2012). Other higher eukaryotic species where allelic expression imbalance has been identified include the identification of AEI in two species of pig using RNA sequencing (Esteve-Codina *et al.*, 2011) and in a wild population of yellow baboons (Tung *et al.*, 2011).

Allelic expression imbalance in constructed diploid hybrid strains has been used with numerous species for various purposes. In yeast, detection of AEI by RNA sequencing of a hybrid of *Saccharomyces cerevisiae* and *Saccharomyces bayanus* was used to identify genes whose expression was favoured in each of the parental species (Bullard *et al.*, 2010b). Whereas a study by Tirosh *et al.* (2009) used a hybrid of *S. cerevisiae* and *S. paradoxus* to identify the contribution of *cis* and *trans* regulatory factors on genes. *Cis* factors can be identified using AEI, whereas *trans* factors are identified by comparing the level of AEI in a diploid hybrid to the ratio of gene expression between parental strains. Similar work has also been carried out in *S. cerevisiae* strains BY4716 and RM11-1a, with AEI identified through pyrosequencing (Sung *et al.*, 2009), and in *Drosophila* using hybrid strains of *D. melanogaster* and *D. simulans* (Wittkopp *et al.*, 2004, Main *et al.*, 2009). The use of a constructed “super-hybrid” of rice species demonstrated that levels of AEI can significantly change between developmental stages, in this case tillering and heading (Zhai *et al.*, 2013).

1.6 Mechanisms of Control of AEI

The exact regulatory mechanisms behind allelic expression imbalance are yet to be fully determined and it is unknown if it is controlled by a single mechanism or combination of several. Many different factors have been suggested to play a role including CpG methylation patterns, heterochromatin blocks, distances to enhancers, and replication asynchrony (reviewed by Ohlsson *et al.* (1998)). Differences in promoter sequences, inter-chromosomal interactions, and post-translational histone modifications may also play a role. There are suggestions from some groups that epigenetic mechanisms are the most influential factors (Pastinen and Hudson, 2004), but this is still speculative. Examples of the contributions of some of these factors are discussed below.

1.6.1 Regulation of AEI via Methylation

Methylation of DNA is an epigenetic marker generally associated with gene silencing. Repression is achieved through association of methylated DNA at CpG dinucleotides with methyl-CpG-binding proteins. In humans, these proteins are known as MBD1-4 and MeCP2. These proteins further recruit other enzymes, such as histone deacetylases, which leads to modification of the chromatin structure and therefore silencing of the gene (Newell-Price *et al.*, 2000). Methylation is commonly associated with imprinting as a mechanism of silencing just one allele. All but one imprinted gene in humans has been shown to exhibit allele-specific methylation (Brannan and Bartolomei, 1999).

For example in the case of the *H19/Igf2* locus in mice (discussed in section 1.5.1.2), the lack of methylation on the *H19* maternal allele causes the DNA to bind to the transcriptional repressor CTCF. This changes the three-dimensional arrangement of the DNA preventing enhancers accessing *Igf2* and ensuring only *H19* is expressed (Hou and Corces, 2011).

Lister *et al.* (2008) studied the relationship between the genome-wide methylome (methyl-seq) and transcriptome (RNA-seq) in *Arabidopsis thaliana* showing that methylation was associated with altered transcript abundance of hundreds of genes. However, a smaller scale study of methylation in maize found that not all imprinted genes have methylation differences (Zhang *et al.*, 2011). Little information is available on genome-wide levels of allele-specific

methylation patterns due to difficulties in using standard bisulphite sequencing methods. As methylcytosines occur at higher frequencies than SNPs, it is often difficult to identify which allele is associated with the methylation pattern using conventional next generation sequencing techniques. Due to the fragmentation used in library preparation, reads which contain methylcytosines may not contain a SNP and therefore cannot be assigned to a specific allele.

1.6.2 Regulation of AEI via Promoter Region Differences

As promoters are directly linked to gene expression levels, it can be speculated that differences in promoter regions of alleles may cause differential allele expression. Definitions of exact promoter regions are often unclear, especially within non-human species. Therefore the role of SNPs in upstream regions adjacent to genes with AEI has often been investigated.

On a genome-wide level, Gagneur *et al.* (2009) found a significant positive correlation between the polymorphism density in the promoter region and the level of AEI observed. However, these polymorphisms are yet to be determined as the sole *cis*-acting factor contributing to AEI.

On a single gene basis, a study on the human *RPTOR* gene showed that a SNP in the upstream region of one allele was associated with people originating from cooler climates. The SNP (T to C) affected the binding ability of the transcription factor POU2F1 and therefore lead to a decrease in expression of the *RPTOR* allele (Sun *et al.*, 2010).

1.6.3 Regulation of AEI via Other Mechanisms

As well as DNA methylation and promoter differences, other mechanisms have been suggested to play a role in the control of allele-specific expression. In cases of imprinting, a possible role for inter-chromosomal interactions has been put forward. Imprinted loci are overrepresented in regions of chromosomal interactions, and X-chromosomes interact through an unknown mechanism during X-chromosome inactivation (Gartler and Goldman, 2005). Positioning of the DNA within the cell itself may also impact upon expression levels. It has been shown that in *S. cerevisiae* regions of heterochromatic DNA, especially telomeric regions, are tethered to the nuclear envelope leading to transcriptional

silencing (Andrulis *et al.*, 1998). Although this mechanism has not been linked to allelic expression imbalance, it is sensible to infer that if alleles differ in their perinuclear tethering, the expression levels may also differ.

Evidence has been gathered which suggests that chromatin structure and therefore gene expression rates do differ between homologous chromosomes in the same cell (Gaur *et al.*, 2013). How these chromatin differences are achieved is yet to be fully determined, but this could link back to differences in DNA methylation patterns as mentioned in section 1.6.1.

Fluorescent *in situ* hybridisation analysis was used with the monoallelic odorant receptor genes (see section 1.5.1.1) to show that the active and inactive alleles replicate asynchronously; a trend commonly reported in imprinting and X chromosome inactivation (Chess *et al.*, 1994). Although asynchronous replication has been shown to occur in every case of human monoallelic expression, there is no actual link between whether an allele is replicated early or late and whether is it expressed or not (Gimelbrant *et al.*, 2007) and there is yet to be an explanation of how asynchronous replication controls the expression levels of alleles directly.

These examples demonstrate the complexity of the control mechanisms behind AEI and how there is still a long way to go before this phenomenon is clearly understood.

1.7 AEI and Disease

AEI has been implicated as a causative factor in various human diseases, most significantly cancer. For example loss of imprinting at the *H19/Igf2* locus resulting in biallelic expression has been seen in many cancer types (Feinberg, 1993), and a complete switch in monoallelic expression at this locus has been observed in oral squamous cell carcinoma (OSCC) (Tuch *et al.*, 2010a). In addition, the OSCC tissues demonstrated that genes with AEI were enriched for cancer related functions and were associated with copy number mutations, implying a role for allelic expression imbalance in cancer aetiology (Tuch *et al.*, 2010a). A study using TaqMan qPCR showed that AEI levels of *BRCA1* were significantly increased in lymphocytes from familial breast cancer patients when

compared to cancer free patients, but the exact mechanisms and contributions to disease phenotypes are yet to be determined (Chen *et al.*, 2008). SNP chips have been used to monitor AEI in colorectal cancer cells. Differences in the patterns of expression were also observed between cancer cells and non-cancerous B cells suggesting colorectal cancer-specific AEI occurs (Lee *et al.*, 2013). Examples in non-cancer related diseases include (not exclusively) reduction in the expression of one allele of the APC gene which has been linked to familial adenomatous polyposis (Yan *et al.*, 2002a).

Imprinting associated congenital disorders in humans such as Prader-Willi, Angelman, Beckwith-Wiedemann and Silver-Russell syndromes are caused by both genetic mutations and epigenetic alterations, which contribute to changes in the methylation status of either the imprinted gene or the control region. Each syndrome is associated with unique phenotypes dependent upon the imprinted gene affected. For example, Silver-Russell syndrome causes postnatal growth retardation due to loss of methylation at the *H19/Igf2* locus, and subsequent loss of *Igf2* expression. Conversely, postnatal overgrowth is observed in Beckwith-Wiedemann syndrome due to hypermethylation at the *H19/Igf2* locus which leads to increased expression of *Igf2* (Girardot *et al.* 2013). Prader-Willi and Angelman syndromes are both associated with loss of imprinting at chromosome 15. Prader-Willi is due to maternal uniparental disomy (both copies of the chromosome are from the mother) whereas Angelman syndrome is due to loss of maternal expression. This implies that both maternal and paternal copies of chromosome 15 are required for normal development in humans (Girardot *et al.*, 2013).

1.8 AEI and *Candida albicans*

Although AEI has been identified in some *Saccharomyces* yeast species hybrids (see section 1.5.2), investigations on a genome-wide scale in *Candida albicans* are still in their rudimentary stages. Muzzey *et al.* (2013) demonstrated the advantages of using a phased diploid genome reference whilst investigating AEI on a genome-wide scale using existing RNA sequencing data, but did not look into the functional consequences of the AEI discovered. In a follow up study, the level of AEI at both the transcriptional and translational level in *C. albicans* was assessed. Again, very little functional inference was made to

explain the levels of allelic expression imbalance, although an over-representation of genes with mitochondrial functions was observed (Muzzey *et al.*, 2014). However some examples of the functional impact of allelic expression imbalance in *C. albicans* have been seen on a gene-by-gene basis.

Staib *et al.* (2002) identified differences in promoter regions of the *SAP2* alleles resulting in differential regulation in expression of the two alleles. In the case of the drug-resistance gene *MDR1,* it has been found that the promoters of the alleles differ in DNA sequence conferring to a difference in expression level (Bruzual and Kumamoto, 2011). A similar study into the chitin synthesis gene *CHS7* found that although the alleles themselves did not differ in sequence, the promoter regions differed in length. This directly impacted upon the allele expression levels. A heterozygous knockout containing just the allele with the shorter promoter had similar characteristics to the wild-type whereas a knockout strain with only the allele with the long promoter remaining suffered from reduced chitin and moderate morphological differences during hyphal growth (Sanz *et al.*, 2007).

As previously described in section 1.4.1, evidence has been gathered showing that the alleles of a gene can differ in function, for example the *ADE2* gene has a functional and non-functional allele (Tsang *et al.*, 1999), and phenotypic differences in the knockout of the *ALS9* gene can only be complemented by allele two (Zhao *et al.*, 2007). During the sequencing of the *C. albicans* diploid genome, SNPs were identified in 3579 ORFs, just over approximately half of all genes. In 78% of these genes, the SNPs lead to an alteration in predicted protein sequence (Jones *et al.*, 2004). This suggests that differences in function may occur at a high rate across the genome. Although differences in allele function have not yet been linked to differences in the expression levels, it would be logical to suggest that this may be the case.

The above examples have shown that SNPs within promoter regions are leading to AEI. However, the epigenetic control factors mentioned in section 1.6 may also be impacting upon AEI in *C.* albicans. The genome-wide extent of DNA methylation, and the subsequent impact upon expression, has been identified in *C. albicans.* Unlike other fungal species, where DNA methylation

occurs at the highest frequency over repeat sequences, in *C. albicans* methylation is centred on 150 genes with an over representation of genes involved in environmentally cued pathways such as the switch to hyphae and drug resistance. It was shown that methylation levels are fluid and vary significantly between the yeast and hyphal form, and that this methylation is directly causing transcriptional repression in a set of these genes (Mishra *et al.*, 2011).

1.9 Aims and Objectives

The body of work presented here aims to investigate the functional consequences of allelic expression imbalance in *Candida albicans.* Due to the high levels of heterozygosity identified during genome sequencing, it is sensible to infer that both coding regions and regulatory regions of alleles may differ in sequence. Dependent upon the extent of these polymorphisms, alleles of a single gene may therefore be divergent in virulence-related functions. A striking example of this can be seen when observing the 149 genes with the Gene Ontology (GO) term “pathogenesis”. 48.3% of these genes contain non-synonymous substitutions, in comparison to just 23.1% seen across all genes. As function plays a role in determining gene expression levels, genes with differing allele function may also experience a difference in allele expression.

As mentioned above, to date, there has been little evidence of genome-wide allelic expression imbalance in *C. albicans,* although examples have been demonstrated on a gene-by-gene basis. Here RNA sequencing has been used to identify the extent of AEI in wild-type *C. albicans* grown under common laboratory conditions. Unlike previous studies which use SNP identification with RNA sequencing to investigate AEI, here the availability of the diploid reference genome enabled reads to be aligned directly to each allele in a novel manner.

From the subsequent list of genes with significant AEI, the functional and phenotypic consequences of differing allele expression levels were investigated. The contribution of each allele to pathogenesis-related phenotypes, such as morphology and antifungal drug resistance, was tested by comparing constructed heterozygous knockout strains where only one allele remains. Using genes within this list, attempts were made to verify the AEI using various

methods including allele-specific qPCR and western blotting. Due to complications in the methods surrounding the issue of allele-specificity, this was unfortunately not achieved, but has presented an opportunity to discuss the difficulties presented whilst studying allele-specific expression.

Structural factors, such as chromosomal location, GC content, gene length and codon usage, have previously been shown to impact upon gene expression levels. However, the role of these structural factors in control of AEI is yet to be investigated. Therefore, in this study, the contribution of each factor to allele-specific expression levels was explored.

Finally, the method of computational analysis used to identify AEI was critically assessed and therefore modified to identify genes with allele-specific expression from existing RNA sequencing data which is publically available. This enabled a consideration of AEI in a condition specific manner. The functional consequences of which were again investigated through the use of heterozygous knockout strains.

1.9.1 Aims and Objectives Summarised

The main research objectives which are investigated in this piece of work are as follows:

- Identification of AEI in the wild-type strain of *Candida albicans,* SC5314, using RNA sequencing.
- Analysis of the impact of structural factors upon levels of AEI.
- Development of an alternative method to identify and verify AEI.
- Construction and phenotypic screening of heterozygous knockout mutants to assess the functional consequence of AEI.
- Development of a computational pipeline to allow for condition-specific identification of significant AEI using RNA sequencing data.
- Further construction and phenotypic screening of heterozygous knockout mutants to assess the functional consequences of condition-specific AEI.

Chapter 2: General Materials and Methods

2.1 Strains Used

Strains used in this study are listed in Table 2.1 and Table 2.2. All strains were stored in 50% (v/v) glycerol at -80 °C. Strains were streaked onto appropriate solid media (see section 2.3) and stored at 4 °C for no longer than one month.

2.2 Plasmids Used

Plasmids used in this study are listed in Table 2.3.

2.3 Growth Conditions

Yeast strains were grown in YPD media (2% (w/v) Bacto-peptone, 2% (w/v) glucose, 1% (w/v) yeast extract) unless otherwise stated. When grown on solid media, 2% (w/v) agar was added. For nourseothricin selection, Sabouraud dextrose agar (Melford laboratories) was used (65 mg/ml) with nourseothricin (Werner Bioagents) at 100 µg/ml, filter sterilised using a 0.2 µm filter disc (Sartorius Stedim). Media were autoclaved at 121 °C for 15 minutes. All strains were grown at 30 °C unless otherwise stated. 90 mm diameter sterile plastic Petri dishes (triple-vented) (Sterilin) were used as standard.

Escherichia coli strains were grown in LB (Luria-Bertani) media (1% (w/v) bacto-tryptone, 0.5% (w/v) yeast extract, 1% (w/v) sodium chloride, pH 7.5). When grown on solid media 2% (w/v) agar was added. For selection of ampicillin resistant strains, ampicillin (Sigma Aldrich) was added to the media at 100 µg/ml, filter sterilised using a 0.2 µm filter disc (Sartorius Stedim). Media were autoclaved at 121 °C for 15 minutes. All strains were grown at 37 °C unless otherwise stated. 90 mm diameter sterile plastic Petri dishes (triple-vented) (Sterilin) were used as standard.

Table 2.1 Yeast strains used in this study

Strain Name	Species	Genotype	Source
SC5314	*C. albicans*	Ura^{+} ancestor of CAI4	Gillum *et al.*, 1984
SC3	*C. albicans*	SC5314 *rbt4-1::NAT1 RBT4-2*	This Study
SC4	*C. albicans*	SC5314 *RBT4-1 rbt4-2::NAT1*	This Study
SC5	*C. albicans*	SC5314 *RBT4-1 rbt4-2::NAT1*	This Study
SC6	*C. albicans*	SC5314 *vps1-1::NAT1 VPS1-2*	This Study
SC7	*C. albicans*	SC5314 *VPS1-1 vps1-2::NAT1*	This Study
SC8	*C. albicans*	SC5314 *CDC6-1 cdc6-2::NAT1*	This Study
SC9	*C. albicans*	SC5314 *cdc6-1::NAT1 CDC6-2*	This Study
SC10	*C. albicans*	SC5314 *cdc6-1::NAT1 CDC6-2*	This Study
SC12	*C. albicans*	SC5314 *RPS1-1 rps1-2::CIp10-NAT1*	This Study
SC13	*C. albicans*	SC5314 *VPS1-1::*V5-6xHis-*NAT1 VPS1-2*	This Study
SC14	*C. albicans*	SC5314 *VPS1-1::*V5-6xHis-*NAT1 VPS1-2*	This Study
SC16	*C. albicans*	SC5314 *vps1-1::NAT1 VPS1-2*	This Study
SC17	*C. albicans*	SC5314 *vps1-1::NAT1 VPS1-2*	This Study
SC18	*C. albicans*	SC5314 *VPS1-1 vps1-2::NAT1*	This Study
SC27	*C. albicans*	SC5314 *VPS1-1 VPS1-2*-V5-6xHis-*NAT1*	This Study
SC28	*C. albicans*	SC5314 *VPS1-1 VPS1-2*-V5-6xHis-*NAT1*	This Study
SC30	*C. albicans*	SC5314 *SMI1-1 smi1-2::NAT1*	This Study
SC32	*C. albicans*	SC5314 *VPS1-1 vps1-2::NAT1*	This Study
SC33	*C. albicans*	SC5314 *smi1-1::NAT1 SMI1-2*	This Study
SC34	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC35	*C. albicans*	SC5314 *CDC6-1*-V5-6xHis-*NAT1 CDC6-2*	This Study
SC36	*C. albicans*	SC5314 *CDC6-1*-V5-6xHis-*NAT1 CDC6-2*	This Study
SC41	*C. albicans*	SC5314 *CDC6-1 CDC6-2*-V5-6xHis-*NAT1*	This Study
SC42	*C. albicans*	SC5314 *erb1-1::NAT1 ERB1-2*	This Study
SC44	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC45	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC46	*C. albicans*	SC5314 *erb1-1::NAT1 ERB1-2*	This Study
SC47	*C. albicans*	SC5314 *ERB1-1 erb1-2::NAT1*	This Study
SC48	*C. albicans*	SC5314 *ERB1-1 erb1-2::NAT1*	This Study
SC49	*C. albicans*	SC5314 *erb1-1::NAT1 ERB1-2*	This Study
SC51	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC52	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC53	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC54	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study

SC55	*C. albicans*	SC5314 *rck2-1::NAT1 RCK2-2*	This Study
SC56	*C. albicans*	SC5314 *ERB1-1 erb1-2::NAT1*	This Study
SC61	*C. albicans*	SC5314 *CDC6-1 CDC6*-2-V5-6xHis-*NAT1*	This Study
SC62	*C. albicans*	SC5314 *CDC6-1 CDC6*-2-V5-6xHis-*NAT1*	This Study
SC66	*C. albicans*	SC5314 *ADH2-1 adh2-2::NAT1*	This Study
SC67	*C. albicans*	SC5314 *ADH2-1 adh2-2::NAT1*	This Study
SC68	*C. albicans*	SC5314 *ADH2-1 adh2-2::NAT1*	This Study
SC69	*C. albicans*	SC5314 *adh2-1::NAT1 ADH2-2*	This Study
SC70	*C. albicans*	SC5314 *rps7a-1::NAT1 RPS7A-2*	This Study
SC71	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC72	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC73	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC74	*C. albicans*	SC5314 *gpx1-1::NAT1 GPX1-2*	This Study
SC75	*C. albicans*	SC5314 *gpx1-1::NAT1 GPX1-2*	This Study
SC76	*C. albicans*	SC5314 *gpx1-1::NAT1 GPX1-2*	This Study
SC81	*C. albicans*	SC5314 *GPX1-1 gpx1-2::NAT1*	This Study
SC85	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC86	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC87	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC88	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC90	*C. albicans*	SC5314 *adh2-1*::*NAT1 ADH2-2*	This Study
SC92	*C. albicans*	SC5314 *adh2-1*::*NAT1 ADH2-2*	This Study
SC93	*C. albicans*	SC5314 *rps7a-1*::*NAT1 RPS7A-2*	This Study
SC94	*C. albicans*	SC5314 *rps7a-1*::*NAT1 RPS7A-2*	This Study
SC96	*C. albicans*	SC5314 *rps7a-1*::*NAT1 RPS7A-2*	This Study
SC97	*C. albicans*	SC5314 *rps7a-1*::*NAT1 RPS7A-2*	This Study
SC98	*C. albicans*	SC5314 *rps7aA-1*::*NAT1 RPS7A-2*	This Study
SC99	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC100	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC101	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC102	*C. albicans*	SC5314 orf19.5648-1 orf19.5648-2::*NAT1*	This Study
SC107	*C. albicans*	SC5314 *GPX1-1 gpx1-2*::*NAT1*	This Study
SC108	*C. albicans*	SC5314 *GPX1-1 gpx1-2*::*NAT1*	This Study

Table 2.2 *Escherichia coli* strains used in this study

Strain Name	Genotype	Source
DH5α	F- Φ80*lacZ*ΔM15 Δ(*lacZ*YA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rK-,mK+) *pho*A *sup*E44 λ- *thi*-1 *gyr*A96 *rel*A1	Lab Stock
DH5α (pJK795)	*NAT1* cassette pBS	Shen *et al.*, 2005
DH5α (Clp10-*NAT1*)	Clp10 + Ag*TEF1*p-Ca*NAT1*-Ag*TEF1*t	Bates, S., Personal Communication
topo(SB168)	V5-6xHis-*NAT1* cassette	Milne *et al.*, 2011

Table 2.3 Plasmids used in this study

Plasmid	Description	Source
pJK795	*NAT1* cassette pBS	Shen *et al.*, 2005
Clp10-*NAT1*	Clp10 + AgTEF1p-Ca*NAT1*-AGTEF1t	Bates, S., Personal Communication
SB168	V5-6xHis-*NAT1* cassette	Milne *et al.*, 2011
pGEM-T		Promega

2.4 Purification of Plasmid DNA from *E. coli*

A single colony of *E. coli* was inoculated in 5 ml of LB medium in a 30 ml universal tube (Greiner Bio-One) with ampicillin and grown overnight at 37 °C, 180 rpm. Plasmids were then extracted using either an NBS Biologicals mini prep kit, a Promega plasmid mini prep kit or an Omega EZNA plasmid mini prep kit, as per manufacturer's instructions. Plasmids were resuspended in 1 x TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0), separated by gel electrophoresis (see section 2.5) to check the size and purity of constructs, and stored at -20 °C until use. DNA concentrations were subsequently measured using a NanoDrop Spectrophotometer.

2.5 DNA Gel Electrophoresis

Gel electrophoresis was carried out using 1% (w/v) agarose in TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 8.0) containing 15 nl/ml ethidium bromide. For loading, samples were diluted in a 5:1 ratio in DNA loading buffer

(0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol, 30% (v/v) glycerol). Results were visualised using a G:Box system (SynGene) set with a UV filter for use with ethidium bromide.

2.6 Gel Extraction

Extraction and purification of DNA fragments from an agarose gel was carried out using either an NBS DNA purification spin column kit or a Qiagen QIAquick gel extraction kit following the manufacturer's instructions.

2.7 Polymerase Chain Reaction

Polymerase chain reactions (PCR) were set up as follows: 1 x Thermostart PCR mastermix (0.625 U ThermoPrime *Taq* DNA polymerase, 75 mM Tris-HCl (pH 8.8 at 25 °C), 20 mM ammonium sulphate, 1.5 mM magnesium chloride, 0.01% (v/v) Tween® 20, 0.2 mM each of dATP, dCTP, dGTP and dTTP; Thermo scientific), 10 µM forward primer, 10 µM reverse primer and 1 µl of template DNA.

Thermal cycling conditions were as follows:

Initial denaturation:	94 °C for 3 minutes	
Denaturation:	94 °C for 1 minute	30 cycles
Annealing:	52 °C for 1 minute	
Extension:	72 °C for 1 minute per 1 kb amplified	
Final Extension:	72 °C for 10 minutes	
Hold:	4 °C	

2.8 Cloning

2.8.1 PCR Amplification of Gene

The desired gene for cloning was amplified using polymerase chain reaction as described in section 2.7. For a full list of oligonucleotides (Invitrogen) used to amplify the genes of interest see Table 2.4. The correct band was isolated using gel extraction as described in section 2.6 and quantified using a NanoDrop Spectrophotometer.

Table 2.4 Oligonucleotides used to amplify genes of interest for cloning

Name	Sequence 5' – 3'	Position[1]	Strain Sequenced
RCK2-F	AGCTTTTATTCGACATGGGAAG	18 – 39	SC5314
RCK2-R	GCTGCTTTGAATAGGAATCTGTT	1746 – 1767	SC5314
RPS7A-F	CCAAGGATCAAGCTTCATC	-26 – -8	SC5314
RPS7A-R	CAAATTCTCTTCTGACGGATG	598 – 618	SC5314
5648-F	GTATGATAAACAGTGGTAATGG	-2 – 22	SC5314
5648-R	CATATCTGCTTCATTTGCC	640 – 658	SC5314

1. Relative to A bp of ATG codon = 1. Negative numbers represent primers placed before the gene start position.

2.8.2 Ligation

DNA fragments were ligated into the pGEM-T vector (Promega); a linear vector that allows for easy insertion of *Taq*-polymerase amplified PCR products with no prior enzymatic digest. The plasmid backbone allows for two levels of selection based on both ampicillin resistance and use of the *lacZ* operon (successful colonies are white). Ligation reactions were set up as follows: 5 µl of 2x Rapid Ligation Buffer (Promega), 1 µl of pGEM-T vector, 1 µl of T4 DNA ligase, an appropriate volume of gel extracted PCR product (see below), or for the positive control 2 µl of control DNA, or for the negative control 2 µl of sterile water. The volume was then made up to 10 µl using sterile water and incubated overnight at 4 °C.

Appropriate volumes of PCR fragment were calculated as follows:

$$\text{ng of insert} = \frac{\text{ng of vector} \times \text{size of insert (kb)}}{\text{size of vector (kb)}} \times \text{insert:vector molar ratio}$$

In all cases, the insert:vector molar ratio was always 3:1.

2.8.3 Preparation of Competent *E. coli* Cells

E. coli cells competent for transformation were prepared as follows with all centrifugations carried out at 4 °C. A single colony of the appropriate *E. coli* strain was grown overnight in 5 ml of LB at 37 °C, 180 rpm. The culture was diluted 1:20 in 100 ml of pre-warmed LB and grown at 37 °C, 180 rpm, until the optical density at 550 nm reached 0.48. The culture was then chilled on ice for

five minutes, split into two 50 ml falcon tubes and pelleted at 4000 rpm for five minutes. Each pellet was resuspended in 20 ml of ice cold TfbI solution (30 mM potassium acetate, 100 mM potassium chloride, 10 mM calcium chloride, 50 mM manganese chloride, 15% glycerol, pH 5.8). Cell suspensions were cooled on ice for five minutes and then pelleted at 4000 rpm for five minutes. Each pellet was resuspended in 2 ml of ice-cold TfbII solution (10 mM MOPS, 75 mM calcium chloride, 10 mM potassium chloride, 15% glycerol, pH 6.5) and left to cool on ice for 15 minutes. Cells were then stored in 100 µl aliquots at -80 °C.

2.8.4 Transformation of *E. coli* Cells

The protocol for transformation of plasmid DNA into *E. coli* cells was taken from the manufacturer's instructions for the pGEM-T vector. 2 µl of ligation reactions were combined with 50 µl of competent *E. coli* cells, flicked to mix, and incubated on ice for 20 minutes. The cells were then heat shocked at 42 °C for 45 seconds followed by two minutes on ice. 950 µl of SOC media (2 % (w/v) tryptone, 0.5% (w/v) yeast extract, 10 mM sodium chloride, 10 mM potassium chloride, 40 mM magnesium solution (1 M magnesium sulphate heptahydrate, 1.7 M magnesium sulphate), 20 mM glucose) was added and the cells were left to incubate at 37 °C, 180 rpm, for 1.5 hours. 100 µl was then taken and plated onto LB agar plates containing 100 µg/ml ampicillin, 0.1 mM IPTG and 40 µg/ml X-Gal. Plates were incubated at 37 °C overnight.

Transformation efficiencies of the competent cells were calculated using the control DNA supplied with the vector as a positive control and the following calculation:

$$\text{Transformation efficiency} = \frac{\text{cfu on control plate}}{\text{ng of vector used}} \times \text{final dilution plated}$$

2.8.5 Screening Successful Transformants

Colonies which had the ability to grow on the selective media showed successful transformation of the plasmid due to the resistance to ampicillin. Colonies which were white also showed a further level of selection demonstrating that the *lacZ* operon has been interrupted, and therefore the insert DNA has been successful ligated into the plasmid. These successful

colonies were checked using colony PCR with the same oligonucleotides used for initial gene amplification (see section 2.7 and section 2.11). Plasmids were extracted as described in section 2.4 and stored at -20 °C.

2.9 Construction of DNA Cassettes for Transformation

DNA cassettes for transformation into *C. albicans* were constructed using PCR as described in section 2.7 with a modification of the annealing temperature to 55 °C. 1 µl of Plasmid DNA (200 ng/µl) containing an appropriate cassette (see Table 2.3) was used as the DNA template. Oligonucleotides (Invitrogen) were designed containing 60 bp – 100 bp of target gene sequence and 18 bp – 20 bp of cassette (Figure 2.1). For a full list of oligonucleotides and template plasmid DNA combinations used see Table 2.5. Reactions were checked using DNA gel electrophoresis (see section 2.5). This was followed by precipitation as follows: four separate PCR reactions per cassette were combined and added to 15 µl of 3M sodium acetate and 450 µl of 100% ethanol and frozen at -20 °C for 30 minutes. DNA was pelleted at 13,000 rpm for 15 minutes, washed in 70% ethanol, pelleted at 13,000 rpm for three minutes and resuspended in 10 µl of sterile water. Cassettes were stored at -20 °C until required for *Candida albicans* transformations (see section 2.10).

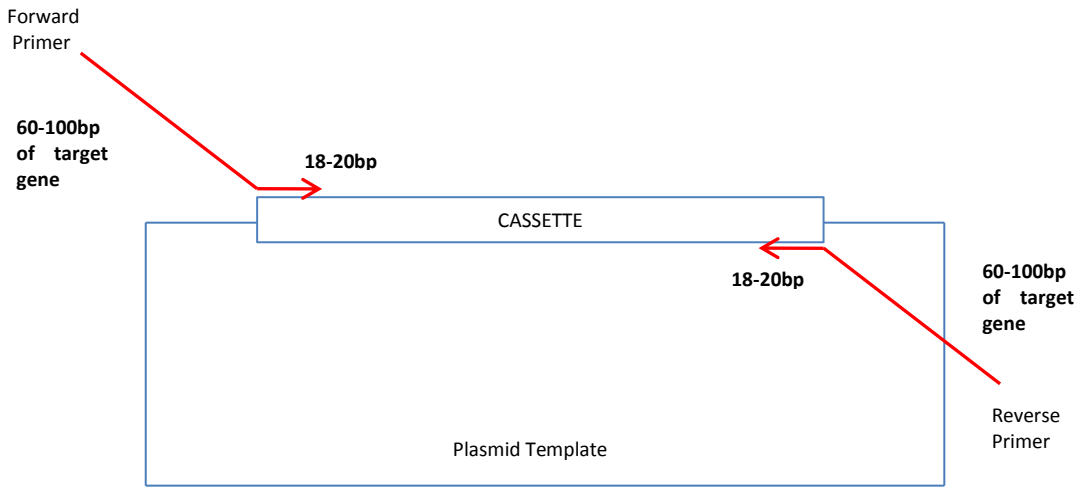


Figure 2.1 Construction of DNA cassette for transformation into *Candida albicans.* Each primer contains 60 – 100 bp of DNA corresponding to the gene of interest and 18 – 20 bp of DNA corresponding to the cassette.

Table 2.5 Oligonucleotides used to construct DNA cassettes

Name	Sequence[1] 5' – 3'	Position[2]	Plasmid Template
CDC6-KO-F	GGTCCACCAGGGACGGGTAAGACTGCTCAAGTTCAATTAATCCTACAACCTTATCAACAGAATTCAAGAATACGGCGGGATATCAAGCTTGC	375 – 448	pJK795
CDC6-KO-R	CCCGATATCTTGAAACCATATTCCGCGGGTACTTTTTTGGTATTGTTTTGAAAGTCACCGTGGGTACCGAATTCGAGC	1545 – 1604	pJK795
ERB1-KO-F	CAGTGATGATGATGATGATGACGATGATGATGACGATGACAACAACTCAGAAGCAGATTCTGGCGGGATATCAAGCTTGC	219 – 280	pJK795
ERB1-KO-R	TATTCATCTATGTAGTCCAAAGACGAGCAGTTCCATCAGCACCAGCACTAAATAACCAAGGTGGGTACCGAATTCGAGC	2497 – 2557	pJK795
RBT4-KO-F	CGCCTATGTCACCCAGACTCGTGGTGTTACTGTTGGTGAAACTGCCACCGTTGCTACAACTGTTACCGGCGGGATATCAAGCTTGC	63 – 130	pJK795
RBT4-KO-R	GCCATATAAGATTTACCAGTCTTTGGATCAGTACCCATAACGTTACCAGCTGGGTCGTAGGAACAAACAACGTGGGTACCGAATTCGAGC	982 – 1053	pJK795
RCK2-KO-F	CGACATGGGAAGCAAGCCAATGATATGAAAAGAAAGCAACAACAGCAGCCACAGCAATATCAACAACCGCGGGATATCAAGCTTGC	28 – 95	pJK795
RCK2-KO-R	GATGTCTATGACGATGAGTTCTTAGGACTTGGTTGTAGGTATCCTGTACAACTCTGCCATCCTGGGTACCGAATTCGAGC	1605 – 1666	pJK795

SMI1-KO-F	GACGATAATGAACCAATTGGTACCAATTCTCATAGGTCATCAACTAATGACTC AGCATTACCGCGGGATATCAAGCTTGC	88 – 149	pJK795
SMI1-KO-R	GCGTATCCACTTTGTCATGGTCTTGTAGATGATTTTGAAAACCAGCCACTAGT CTCGTTGCTGGGTACCGAATTCGAGC	1959 – 2019	pJK795
VPS1-KO-F	GGCTCCTTTAGGTGGAGGGTCATCCTCGCCAGTAGATTTGCCTCAAATCACT GTTGTTGGATCCCGCGGGATATCAAGCTTGC	48 – 112	pJK795
VPS1-KO-R	CTTCGGTTTCCATAGTTTCTCTTTCACTCATAGTACCTGTGGCTCTCAATACT GGAGGTGGGGCTGGGTACCGAATTCGAGC	1762 – 1825	pJK795
CDC6-V5-F	TTGATATAGTGAAAAGTGTTGAAAATATTGGAATCTTGAAAAAAATTTTACAAA AACCAAATAAGGGCGAGCTTCGAGGTCA	1384 – 1445	SB168
CDC6-NAT1-R	CCTACTATCTATCTATCTCTGTCTATCTATTTGTCTGTCTATTTACTTGTATCAT TAATCGTTAGTATCGAATCGACAG	1452 – 1510	SB168
VPS1-V5-F	AAGGAATGTGTTAGAATGGTTGAGGTGTTGAGAAATGCTAGTGAAATTGTTT CTAGTGTTAAGGGCGAGCTTCGAGGTCA	2020 – 2079	SB168
VPS1-NAT1-R	AAGAAGATAAATATATACCACCGACTTTCTGAAATAAAAAGAATTACTCTACT CTATAATACGTTAGTATCGAATCGACAG	2086 – 2146	SB168
ADH2-KO-F	CTATAATCACGAATCAATTGATACTTACCCACTTTTTATTAAATCTAACTCAAT TACACCATGTCTGTCGCGGGATATCAAGCTTGC	-60 – 9	pJK795
ADH2-KO-R	CACTCATTATTATCGTACTTGGCATGAATGCGCTTATTTGTCGTTGTCCAAGA CATATCTATGGGTACCGAATTCGAGC	1020 – 1080	pJK795

RPS7A-KO-F	GGAATAGTCAACCAACAGCAAATAGCCAAGGATCAAGCTTCATCATTAATCAT GTCCTCTAAGATCGCGGGATATCAAGCTTGC	-51 – 15	pJK795
RPS7A-KO-R	GACGGATGATGAAATGGAAAGGTTATTTTTGGGGGGATGTTAATCTAATGAG ATTCACCTGGTGGGTACCGAATTCGAGC	544 – 605	pJK795
5648-KO-F	GAATCTGAATAATCTAATAATTCTTCTTGACCTTCGTGAGACATGATTGGTTG TATGTTTGTATGATAAACAGGCGGGATATCAAGCTTGC	-62 – 11	pJK795
5648-KO-R	GGATTAAGAAGTGCTATTGCTAGAGGTGTTGAAGAGGCTGCTAACATATCTG CTTCATTTGCCTGGGTACCGAATTCGAGC	640 – 702	pJK795
GPX1-KO-F	CTAAATATGGTGAAAAGCAACGTCGAGCTGGCATGGGAAACCATGGAAGATA CATTTCTGGCGGGATATCAAGCTTGC	-6 – 54	pJK795
GPX1-KO-R	CTATGTCTAGCTTTCTAGCAACTGTTCAATCCTTGGTGTTATTGCCACGGGTC TAGTAAACGTATGGGTACCGAATTCGAGC	633 – 696	pJK795

1. Red indicates sequence of plasmid template.
2. Relative to A base pair of ATG codon = 1. Negative numbers represent primers placed before the gene start position.

2.10 *Candida albicans* Transformation

Transformation of exogenous DNA into *C. albicans* was performed using a modified protocol from Cheetham (2008). DNA cassettes were inserted into the genomic DNA via a process of homologous recombination. The sites of recombination were selected during cassette construction (see section 2.9).

To summarise, a single colony of the appropriate *C. albicans* strain was grown overnight in 150 ml of YPD at 30 °C, 180 rpm in a 250 ml conical flask. 50 ml was taken and cells were pelleted at 2500 rpm for two minutes, washed in 20 ml LiAc/TE solution (100 mM LiAc pH 7.0, 1 x TE (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0)), pelleted at 2500 rpm for two minutes, and resuspended in 1 ml LiAc/TE. 10 µl of salmon sperm carrier DNA (10 mg/ml, Invitrogen) was heated to 100 °C for five minutes and allowed to cool. This was then added to 10 µl of transforming DNA and 100 µl of prepared *C. albicans* cells. 750 µl of PEG/LiAc/TE solution (50% PEG 3350, 0.1 mM LiAc pH 7.0, 1 x TE) was added and vortexed to mix. This was followed by a minimum of four hours incubation at 30 °C, 180 rpm. The cell mixture was then heat-shocked at 42 °C for one hour. Cells were pelleted at 8000 rpm for 15 seconds, resuspended in 300 µl sterile water and spread onto selective media. Plates were incubated for 2 days at 30 °C. Colonies that developed on the selective media were further streaked onto selective plates and incubated at 30 °C for a further 24 hours to fully ensure the presence of the antibiotic resistance cassette. Single colonies were then checked for correct insertion and positioning of transformed DNA using colony PCR (see section 2.11).

In the case of heterozygous knockout construction, the remaining allele was amplified via colony PCR (see section 2.11) and sent for sequencing (see section 2.12) to identify the genotype of the knockout constructed. Alternately, in the case of allele tagging, the tagged allele was amplified via colony PCR and sequenced for identification.

2.11 Colony PCR

Colony polymerase chain reaction (PCR) was used to check for correct insertion of DNA cassettes and for amplification of DNA for sequencing. A single colony of the correct strain was put into 10 µl of sterile water and 2 µl of

lyticase solution (25 U/µl lyticase (Sigma Aldrich), 0.1M sodium phosphate buffer at pH 7.5, 10% glycerol). This was incubated at 37 °C for 10 minutes followed by freeze fracturing at -80 °C for 10 minutes. 2 µl of this mixture was then used as the DNA template in a polymerase chain reaction as described in section 2.7. Reactions were checked using DNA gel electrophoresis (see section 2.5).

When designing primers to verify genotypes, oligonucleotides (Invitrogen) of around 18 – 25 bp in length were used to amplify fragments of around 500 bp in length. For a full list of oligonucleotides used for checking insertion of DNA cassettes see Table 2.6 and oligonucleotides used for sequencing DNA see Table 2.7.

2.12 Sequencing

PCR fragments of amplified regions were diluted 1:10 in sterile water and sent, alongside the appropriate the primer (diluted to 3.2 pmol/µl), to either Source Bioscience Lifesciences or GATC Biotech for sequencing.

Table 2.6 Oligonucleotides used to check correct insertion of DNA cassette

Name	Sequence 5' – 3'	Position[1]	Strain Checked
NAT1-CH-R	CCGTAATTTTTGCTTCGCG	190 – 208	SC1 – SC11, SC16 – SC18, SC30, SC32 – SC34, SC42, SC44 – SC49, SC51 – SC56, SC66 – SC76, SC81, SC85 – SC90, SC92 – SC94, SC96 – SC102, SC107
CDC6-CH-F	ACTGGGAACCCTTCATGTGT	128 – 147	SC8 – SC10
ERB1-CH-F	GAGATGAGGTTGGCGCACA	-143 – -125	SC42, SC46 – SC49, SC56
RBT4-CH-F	TCCCATCAACTGTCCATCC	-224 – -206	SC3 – SC5
RCK2-CH-F	GAGCGTGTGTGTGAAGAGAGAA	-282 – -261	SC34, SC44 – SC45, SC51 – SC55
SMI1-CHSEQ-F	GACGACGACGATGGAAAAAC	-254 – -235	SC30, SC33
VPS1-CH-F	CAGTCTAGTTCAATGGAGGCTGG	-276 – -254	SC6 – SC7, SC16 – SC18, SC32
Clp10-IS	GATATCGAATTCACGCGTAG	2936 – 2955	SC12
RP10-GS	GTACATTCCTACTCCGTTCG	1376 – 1395	SC12
V5-S	GAGGGCGTGAATGTAAGCG	140 – 158	SC13 – SC14, SC27 – SC28, SC35 – SC37, SC41, SC57 – SC58, SC60 – SC61, SC65, SC95, SC104 – SC106
VPS1-HIS-C	GCTGGACGTGTCATCCCAT	739 – 757	SC13 – SC14, SC27 – SC28
CDC6-SEQ-F	GGCTAGTATAAATTGCATCCC	686 – 706	SC35 – SC36, SC41, SC61
ERB1-SEQ-F	CCCGATAGTAAAAACACTGCG	1963 – 1983	SC65, SC95
ADH2-CHSEQ-F	GATCACTCTTGCAAGCTAATCTCC	-221 – -198	SC66 – SC69, SC90, SC92
RPS7A-CH-F	GCATGCATCGGAATTCTTTC	-331 – -312	SC70, SC93, SC94, SC96 – SC98
5648-CH-F	CGCCCTCCTTTAGTCTATTCAC	-272 – -251	SC71 – SC73, SC85 – SC88, SC99 – 102
GPX1-CH-F	CCCACTACACCACAAAGGAAAG	-305 – -284	SC74 – SC76, SC81, SC107

1. Relative to A bp of ATG codon = 1. Negative numbers represent primers placed before the gene start position.

Table 2.7 Oligonucleotides used to amplify DNA for sequencing

Name	Sequence 5' – 3'	Position[1]	Strain Sequenced
CDC6-SEQ-F	GGCTAGTATAAATTGCATCCC	686 – 706	SC8 – SC10, SC35 – SC36, SC41, SC61
CDC6-SEQ-R	CCGTAAGAGTGGTAGTAGTAGC	1179 – 1200	SC8 – SC10
ERB1-SEQ-F	CCCGATAGTAAAAACACTGCG	1963 – 1983	SC42, SC46 – SC49, SC56, SC65, SC95
ERB1-SEQ-R	GCCAAATTAAATCCAATATCCC	2464 – 2485	SC42, SC46 – SC49, SC56
RBT4-SEQ-F	CCCAGACTCTACTAAAGACGC	534 – 554	SC3 – SC5
RBT4-SEQ-R	CCCAGTTTTGAGCACGAC	957 – 974	SC3 – SC5
RCK2-SEQ-F	TGTGGGCGTTAGGATGTGTA	1082 – 1101	SC34, SC44 – SC45, SC51 – SC55
RCK2-SEQ-R	CAAACCTTCAATTGGGGCTT	1544 – 1563	SC34, SC44 – SC45, SC51 – SC55
SMI1-CHSEQ-F	GACGACGACGATGGAAAAAC	-254 – -235	SC30, SC33
SMI1-SEQ-R	CCCTGCTGCACCAGTAGAAT	191 – 210	SC30, SC33
VPS1-SEQ-F	TGTCAACGCTGCTAATACGG	597 – 616	SC6 – SC7, SC16 – SC18, SC32
VPS1-SEQ-R	GGATCAATGGCATTAACCCC	1186 – 1205	SC6 – SC7, SC16 – SC18, SC32
V5-S	GAGGGCGTGAATGTAAGCG	115 – 133	SC13 – SC14, SC27 – SC28, SC35 – SC37, SC41, SC57 – SC58, SC60 – SC61, SC65, SC95, SC104 – SC106
VPS1-HIS-C	GCTGGACGTGTCATCCCAT	739 – 757	SC13 – SC14, SC27 – SC28
VPS1-F-1	TGAGTCGGACCAGCCAAATA	-618 – -599	SC6, SC7
VPS1-R-1	CCACTTACACACGACCATCG	-220 – -201	SC6, SC7
VPS1-F-2	CGATGGTCGTGTGTAAGTGG	-220 – -201	SC6, SC7
VPS1-R-2	CAAGGGCCTTCTGGTAACAA	182 – 201	SC6, SC7
VPS1-F-3	TTGTTACCAGAAGGCCCTTG	182 – 201	SC6, SC7

VPS1-R-3	CCGTATTAGCAGCGTTGACA	597 – 616	SC6, SC7
VPS1-F-5	GGGGTTAATGCCATTGATCC	1186 – 1205	SC6, SC7
VPS1-R-5	GGGGCTTCCATTTGTTGTAA	1747 – 1766	SC6, SC7
VPS1-F-6	TTACAACAAATGGAAGCCCC	1747 – 1766	SC6, SC7
VPS1-R-6	CAGGCCCACTTACTCTACGC	2210 – 2229	SC6, SC7
ADH2-CHSEQ-F	GATCACTCTTGCAAGCTAATCTCC	-221 – -198	SC66 – SC69, SC90, SC92
ADH2-SEQ-R	GCACCTGATTGACAGTATTCACAG	300 – 323	SC66 – SC69, SC90, SC92
RPS7A-SEQ-F	CCACCACCAAGTTTACAAGCTTAC	175 – 198	SC70, SC93 – SC94, SC96 – SC98
RPS7A-SEQ-R	TGAATCTTTAGAATCCAACAAGAC	442 – 465	SC70, SC93 – SC94, SC96 – SC98
5648-SEQ-F	ATCGTCGATTTACCACTACCC	269 – 289	SC71 – SC73, SC85 – SC88, SC99 – 102
5648-SEQ-R	GCTTGTTGACCATGGAAGATC	705 – 725	SC71 – SC73, SC85 – SC88, SC99 – 102
GPX1-SEQ-F	CTGATGATTCGACACTCTCAG	152 – 172	SC74 – SC76, SC81, SC107
GPX1-SEQ-R	CTGGATCTGCTTGTTCACC	502 – 520	SC74 – SC76, SC81, SC107

[1.] Relative to A bp of ATG codon = 1. Negative numbers represent primers placed before the gene start position.

2.13 Genomic DNA Extraction for *Candida albicans*

A modified protocol taken from Hoffman and Winston (1987) was used to extract genomic DNA from *Candida albicans* cells. A colony of the appropriate strain was grown overnight in 10 ml of YPD at 30 °C, 180 rpm. Cells were pelleted at 4000 rpm for five minutes, resuspended in 1 ml of sterile water and transferred to a screw-capped microfuge tube. Cells were then pelleted at 13,000 rpm for one minute and resuspended in 200 µl of glass bead buffer (10 mM Tris-HCl pH 8, 100 mM sodium chloride, 1 mM EDTA, 2% Triton X-100, 1% SDS). 200 µl of phenol/chloroform/isoamyl alcohol (25:24:1) and 300 µl of 0.4 – 0.6 mm acid-washed glass beads were added, followed by cell disruption by vortexing for three minutes. 200 µl of 1x TE buffer (10 mM Tris-HCl pH 8.0, 1mM EDTA) was added and mixed by inversion. Cell debris was pelleted at 13,000 rpm for five minutes and the aqueous phase was then transferred to a fresh tube. To this, 1 ml of 100% ethanol was added and mixed by inversion. The precipitate was pelleted at 13,000 rpm for five minutes and resuspended in 400 µl 1x TE buffer. To remove any contaminating RNA, 10 µl of DNAse-free RNase A (10 mg/ml) was added and the solution was incubated at 37 °C for one hour. To this, 10 µl of 3 M sodium acetate, pH 5.2, and 1 ml of 100% isopropanol was added and mixed by inversion. This was incubated at room temperature for 10 minutes, followed by pelleting of the DNA at 13,000 rpm for 10 minutes. The pellet was air-dried before resuspension in 50 µl 10 mM Tris-HCl pH 8.

DNA quality was checked via DNA gel electrophoresis (see section 2.5) and concentrations were checked using a NanoDrop spectrophotometer. Samples were stored at -20 °C until use.

2.14 Phenotypic Screening

The following sections list the phenotypic assays used in both chapter four and chapter five.

2.14.1 Constructing an *RPS1::NAT1* strain

As a control to show the effect of the nourseothricin cassette alone on phenotypes of *Candida albicans*, a strain was constructed where the *RPS1* locus was replaced by the nourseothricin cassette (Murad *et al.*, 2000). To do

this the appropriate strain of *E. coli* (see Table 2.2) was grown up in 5 ml of LB with ampicillin overnight at 37 °C, 180 rpm. The plasmid DNA was extracted as described in section 2.4.

25 µl of plasmid DNA was digested at 37 °C for two hours with the restriction enzyme StuI (Promega) according to the manufacturer's instructions. Correct digestion was checked using DNA gel electrophoresis (see section 2.5). The product was then precipitated by addition of 15 µl 3 M sodium acetate and 450 µl 100% ethanol, followed by freezing at -20 °C for 30 minutes. The DNA was pelleted at 13,000 rpm for five minutes, washed with 70% ethanol, pelleted again at 13,000 rpm for three minutes and resuspended in 10 µl of sterile water. The plasmid DNA was then transformed into *Candida albicans* (see section 2.10) at the *RPS1* locus. Correct insertion was checked using colony PCR (see section 2.11) with primers CIp10-IS and RP10-GS (see Table 2.6).

2.14.2 Growth Rate

Growth rates of *Candida albicans* strains were measured using a liquid assay in a 96-well plate format. A single colony of the appropriate strain was grown overnight in 10 ml of YPD at 30 °C, 180 rpm. 10 µl of this was taken and diluted into 1 ml of fresh YPD. 100 µl as this solution was then transferred to a single well of the 96-well plate. Each strain was plated in technical quadruplicate and sterile YPD was used as a control. The experiment was then carried out in either biological duplicate or triplicate.

Optical density at 650 nm was measured every 3½ minutes using a kinetic spectrophotometer held at either 30 °C or 37 °C for a total of 24 hours (Molecular Devices VersaMax Microplate Reader). Plates were shaken for three minutes in between reads.

For each strain three measurements were calculated: an average end-point optical density (taken at 16 hours), time to maximum inflection (OD at 650 nm > 0.3), and generation time. ANOVA followed by Dunnett's *post-hoc* test was used to statistically compare these measures. Generation times were calculated as follows:

$$g = \frac{Log_{10}N_{T1} - Log_{10}N_{T0}}{Log_{10}2}$$

$$\text{generation time (minutes)} = \left(\frac{T1-T0}{g}\right) \times 60$$

Where:

g = generations in T1 – T0

T0 = First time point

T1 = Second time point

NT_0 = Optical density at first time point

NT_1 = Optical density at second time point

2.14.3 Antifungal Sensitivity

Sensitivity to the antifungal compounds fluconazole (Sigma Aldrich), 5-flucytosine (Sigma Aldrich) and amphotericin B (solubilized, Sigma Aldrich) was tested using a liquid assay in a 96-well format. Drug stocks were prepared in sterile water and stored at -20 °C until use.

The appropriate *Candida albicans* strain was grown overnight in 5 ml of YPD at 30 °C, 180 rpm. Cells were counted using a haemocytometer and then diluted to a concentration of 1 x 10^6 cells/ml. 20 µl of cells were taken and added to a well containing 160 µl of YPD and 20 µl of the appropriate drug at a final concentration ranging from 1 – 1024 µg/ml. The plate contained drug concentrations increasing two fold in each column, with sterile water as a control.

Plates were incubated at 30 °C, 180 rpm, and growth was assayed by measuring optical density at 595 nm using a spectrophotometer (Bio-Rad iMark Microplate Reader) at 0, 24 and 48 hours. Assays were carried out in technical quadruplicate and biological duplicate or triplicate.

2.14.4 Growth Under Stress Conditions

To test the ability of strains to grow under stress conditions, serial dilutions on solid media containing different compounds were used. A full list of conditions

used and strains tested are listed in Table 2.8. The appropriate strains were grown overnight in 5 ml of YPD at 30 °C, 180 rpm. Cell concentrations were counted with a haemocytometer and adjusted to 10x fold dilutions in YPD ranging from 1 x 10^7 cells/ml to 1 x 10^3 cells/ml. A 48-pronge replicator was used in a sterile fashion to spot the strains onto the correct media. Suitable concentrations of compounds were selected by pre-screening SC5314 on a range of concentrations. Plates were incubated for three days at 30 °C and observed every 24 hours.

2.14.5 Hyphal Induction

To test the ability of a strain to switch from the yeast to the hyphal form, a single colony was first grown for a minimum of 24 hours in 5 ml of YPD at 30 °C, 180 rpm, to ensure that cells were in stationary phase. Medium containing 45 ml of YPD + 5 ml foetal calf serum was pre-warmed at 37 °C for a minimum of one hour. 1 ml of the stationary phase culture was then taken and added to the pre-warmed YPD + foetal calf serum and the combined solution was incubated at 37 °C for a maximum of three hours. Every 15 minutes 80 µl was taken and combined with 20 µl of 70% ethanol, in technical triplicate. This fixed the sample and allowed for storage at 4 °C. To quantify the induction of hyphae, cells were pelleted at 13,000 rpm for five minutes and resuspended in 10 µl of sterile water. Using a light microscope, observations were made as to the ability of a strain to induce hyphae. For strains which appeared to have hyphal induction defects, the experiment was repeated in biological triplicate.

Table 2.8 Stress conditions tested

Condition	**Compound (and final concentration)**	**Functional Implication[1]**	**Gene**	**Strain**
Ethanol sensitivity	Ethanol (6 %)	General protein defect	*SMI1* *ADH2*	SC5314, SC12 SC30, SC33 SC66, SC67, SC68, SC69, SC90, SC92
Cell wall damaging	Calcofluor white (40 µg/ml)	Defects in cell wall biogenesis	*SMI1*	SC5314, SC12 SC30, SC33
Hygromycin B	Hygromycin B (100 mg/ml)	Antifungal sensitivity	*SMI1*	SC5314, SC12, SC30, SC33
Divalent Cation sensitivity	Calcium chloride (0.5 M)	Altered expression of plasma membrane ATPases and defects in other biological processes	*SMI1*	SC5314, SC12 SC30, SC33
Cell membrane targeting	SDS (0.1%)	Defects in cell membrane and cell wall	*SMI1*	SC5314, SC12 SC30, SC33
Respiration inhibitor	Antimycin A (1 µg/ml)		*ADH2*	SC5314, SC12 SC42, SC46, SC47, SC48, SC49, SC56

Oxidative Stress	Hydrogen peroxide (2 mM)	Oxidative stress sensitivity		SC5314, SC12
			ADH2	SC42, SC46, SC47, SC48, SC49, SC56
			GPX1	SC74, SC75, SC76, SC81, SC107
Oxidative Stress	Menadione (50 µM)	Oxidative stress sensitivity		SC5314, SC12
			ADH2	SC42, SC46, SC47, SC48, SC49, SC56
			GPX1	SC74, SC75, SC76, SC81, SC107
	tBOOH (1 mM)	Oxidative stress sensitivity		SC5314, SC12
			ADH2	SC42, SC46, SC47, SC48, SC49, SC56
			GPX1	SC74, SC75, SC76, SC81, SC107
Fluconazole	Fluconazole (64 µg/ml)	Antifungal sensitivity		SC5314, SC12
			VPS1	SC6, SC7, SC16, SC17, SC18, SC32
			ADH2	SC42, SC46, SC47, SC48, SC49, SC56
			GPX1	SC74, SC75, SC76, SC81, SC107
Flucytosine	5-Flucytosine (32 µg/ml)	Antifungal sensitivity		SC5314, SC12
			VPS1	SC6, SC7, SC16, SC17, SC18, SC32
Amphotericin B	Amphotericin B (1 µg/ml)	Antifungal sensitivity		SC5314, SC12
			GPX1	SC74, SC75, SC76, SC81, SC107

1. Functional implications as stated in (Hampsey, 1997)

2.14.6 Buccal Epithelial Cell Adhesion

To measure the adhesive capabilities of *Candida albicans* strains, buccal epithelial cells (BECs) were used as described by Odds and Webster (1988). BECs were isolated from the inner cheek using a sterile cotton bud and suspended in 10 ml of PBS. Cells were then pelleted at 3000 rpm for five minutes, washed twice in 10 ml of PBS, and resuspended in 1 ml of PBS. This produced an appropriate number to test.

A single colony of the appropriate *Candida albicans* strain was grown overnight in 10 ml of YPD at 30 °C, 180 rpm. Cells were pelleted at 3000 rpm for 5 minutes, washed twice in 10 ml PBS and resuspended in 5 ml of PBS.

200 µl of BECs were then combined with 200 µl of *C. albicans* cells and incubated at 30 °C for one hour. At this point, 40 µl of 37% formaldehyde was added to crosslink any adhesion and the number of *C. albicans* cells adhered to 150 different BECs was calculated using a light microscope. Three measurements were calculated from these results:; percentage of BECs adhered to *C. albicans* cells, number of *C. albicans* cells per BEC, and number of *C. albicans* cells per BEC discounting BEC with no cells adhered. These measurements were statistically compared to the wild-type strain SC5314 using a Student's t-test. The analysis was performed in biological triplicate.

2.14.7 Virulence with *Galleria mellonella* Model

Galleria mellonella (Greater Wax Moth) larvae are commonly used as a model for virulence of *Candida albicans.* The results have been shown to be consistent with those found using mammalian models, but *G. mellonella* are cheaper, easier to manipulate, and avoid unnecessary mammalian suffering (Cotter *et al.*, 2000).

A single colony of the appropriate strain of *C. albicans* was grown overnight in 5 ml of YPD at 30 °C, 180 rpm. The cells were pelleted at 4000 rpm for 5 minutes, washed with 10 ml of PBS, and resuspended in 5 ml of PBS. Cell numbers estimated using a haemocytometer and were diluted in PBS to 1×10^7, 2×10^7 and 5×10^7 cells/ml. 10 µl of cell suspension was injected into the right pro-leg of *G. mellonella,* before incubation at 37 °C. Survival rates were monitored

every 24 hours for a total of 72 hours. PBS alone was used as a control to monitor natural survival rates. Kaplan-Meier survival statistics were then calculated using IBM SPSS version 21 (IBM, 2012). *Galleria* larvae were injected in batches of 10 and repeated in biological triplicate.

2.15 Southern Blotting

Southern blotting was used to check that all heterozygous knockout strains had just one copy of the *NAT1* cassette integrated into the genomic DNA and that this copy was in the correct position. The full protocol was based upon the methodology in Southern (1975): genomic DNA is digested using restriction enzymes, DNA is separated using gel electrophoresis, transferred to a blot, hybridised with a specific probe and then an antibody, and finally the blot is developed and visualised. Details of each step follow.

2.15.1 Synthesis of Digoxigenin Probe

For non-radioactive Southern blotting, digoxigenin (DIG) probes were synthesised using PCR DIG Probe Synthesis kit (Roche). Reactions were prepared as follows: 0.75 µl of Vial 1 enzyme mix, 5 µl of Vial 2 DIG probe synthesis mix, 5 µl of Vial 3 PCR buffer without magnesium chloride, 1 µl of forward primer, 1 µl of reverse primer, 1 µl of template DNA, 36.25 µl of sterile water. For a list of oligonucleotides (Invitrogen) used see Table 2.9.

Thermal cycling conditions were set as follows:

Initial denaturation:	95 °C for 2 minutes	
Denaturation:	95 °C for 30 seconds	30 cycles
Annealing:	50 °C for 30 seconds	
Extension:	72 °C for 40 seconds	
Final Extension:	72 °C for 7 minutes	
Hold:	4 °C	

Successful reactions were checked using DNA gel electrophoresis (see section 2.5).

Table 2.9 Oligonucleotides used to amplify *NAT1* probe for Southern blotting

Name	Sequence 5' – 3'	Position[1]	Plasmid Template
NAT1-S-F	CTGCTACTGGTGATGGTTTC	524 – 542	pJK795
NAT1-S-R	AAACCACACAAAGTGAAACC	889 – 907	pJK795

1. Relative to gene start position = 1.

2.15.2 Restriction Enzyme Digests

Restriction enzyme digests were used to fragment genomic DNA into fragments of known size predicted from the sequence of the *C. albicans* genome. This information was sourced from the *Candida* genome database (www.candidagenome.org) (Inglis *et al.* 2012). A list of restriction enzymes used for each heterozygous knockout can be found in Table 2.10. Genomic DNA was extracted from the appropriate strain as described in section 2.13. Digests were set up as follows: 3 µg of DNA, 3 µl of the appropriate buffer (See Table 2.10), 3 µl of restriction enzyme and 23 µl of sterile water. Reactions were incubated at 37 °C for 24 hours and separated using DNA gel electrophoresis (see section 2.5) with a 0.7% (w/v) agarose gel without ethidium bromide. Staining was carried out after the gel had run by washing with 1x TAE containing 15 nl/ml ethidium bromide with gentle agitation for 30 minutes.

Table 2.10 Restriction enzymes used to digest genomic DNA for Southern blotting

Restriction Enzyme[1]	Buffer	Gene	Fragment Size (bp)	Strains Checked
BglII	3	*SMI1*	4588	SC30, SC33
BsaHI	4	*GPX1*	5342	SC74 – SC76, SC81, SC107
BspHI	4	*VPS1*	2931	SC6 – SC17, SC16 – SC18, SC32
BsrGI	2	*RBT4*	3635	SC3 – SC5
ClaI	4	*ADH2*	7145 (allele 1) 1910 (allele 2)	SC66 – SC69, SC90, SC92
MspA1I	4	*CDC6*	2758	SC8 – SC10
PciI	3	*ERB1*	2573	SC42, SC46 – SC49, SC56

1. All restriction enzymes were purchased from New England BioLabs UK.

2.15.3 Blotting

After visualisation of the digested DNA, the gel was depurinated in 250 mM hydrochloric acid for 15 minutes with gentle agitation, followed by washing in sterile water for 10 minutes. The gel was then denatured twice in denaturation solution (1.5 M sodium chloride, 500 mM sodium hydroxide) for 15 minutes with gentle agitation, followed by washing in sterile water for 10 minutes. Neutralisation was then carried out twice in neutralisation solution (500 mM Tris, 1.5 M sodium chloride, 4.3% (v/v) hydrochloric acid, pH 7.2) for 15 minutes with gentle agitation, followed by washing in sterile water for 10 minutes. The gel was then equilibrated in 20x SSC buffer (2 M sodium chloride, 300 mM sodium citrate, pH 7.0) for 10 minutes with gentle agitation. The blot was then set up in 10x SSC buffer (1M sodium chloride, 150 mM sodium citrate, pH 7.0) and left overnight to transfer the DNA from the gel to a positively charged nylon membrane (Roche).

2.15.4 Hybridisation

The nylon membrane was placed DNA side up in 2x SSC (200 mM sodium chloride, 30 mM sodium citrate, pH 7.0). The DNA was cross-linked to the membrane using exposure to ultraviolet light (HL-2000 Hybrilinker, UVP).

The range of appropriate hybridisation temperatures (° C) for the probe was determined using the following calculations:

$$\text{Low } T_{hyb} = (49.82 + 0.41(\text{GC content})) - \frac{600}{\text{Length (bp)}} - 25$$

$$\text{High } T_{hyb} = (49.82 + 0.41(\text{GC content})) - \frac{600}{\text{Length (bp)}} - 20$$

$$\text{Optimal } T_{hyb} = \frac{\text{Low } T_{hyb} + \text{High } T_{hyb}}{2}$$

30 ml of DIG Easy Hyb solution (Roche) was heated to the required temperature. The membrane was then pre-hybridised in 20 ml of this warmed solution at the optimal hybridisation temperature using a hybridisation oven (HL-2000 Hybrilinker, UVP) for a minimum of 30 minutes to a maximum of 3 hours. 20 µl of the Digoxigenin probe (as prepared in section 2.15.1) was taken, combined with 50 µl of sterile water, heated at 100 °C for five minutes, and then added to the remaining 10 ml of warmed DIG Easy Hyb solution (Roche). (In the instance of probe re-use, the probe hybridisation buffer was prepared by heating to 68 °C for 10 minutes). The pre-hybridisation buffer was removed from the membrane and replaced with the DIG Easy Hyb solution (Roche) containing the probe. The membrane was then hybridised overnight at the optimal hybridisation temperature using a hybridisation oven (HL-2000 Hybrilinker, UVP).

2.15.5 Development of the Blot

The membrane was taken and washed twice in low stringency wash buffer (2x SSC (200 mM sodium chloride, 30 mM sodium citrate, pH 7.0), 0.1% (v/v) SDS) for five minutes with gentle agitation. This was followed by washing twice in high stringency wash buffer (0.5x SSC (50 mM sodium chloride, 7.5 mM sodium citrate, pH 7.0), 0.1% (v/v) SDS) for 15 minutes at 65 °C in a hybridisation oven

(HL-2000 Hybrilinker, UVP). The membrane was washed in washing buffer (0.3% (v/v) Tween-20 in maleic acid buffer (0.1 M maleic acid, 0.15 M sodium chloride, pH 7.5)) for two minutes with gentle agitation before blocking in blocking reagent (1% (w/v) blocking reagent, Roche, in maleic acid buffer) for a minimum of 30 minutes to a maximum of 3 hours with gentle agitation. The antibody (anti-digoxigenin-AP, Roche) was prepared by centrifugation at 10000 rpm for five minutes. 1 µl of antibody per 10 ml of fresh blocking reagent was taken from the top layer. The membrane was then incubated with the blocking reagent containing antibody at room temperature for 30 minutes with gentle agitation. This was followed by washing twice in washing buffer for 15 minutes with gentle agitation. The membrane was equilibrated in detection buffer (0.1 M Tris, 0.1 M sodium chloride, pH 9.5) for three minutes with gentle agitation before being moved to a development cassette. Chemiluminescence substrate (CDP-star, Roche) was added to the membrane and left in the dark for five minutes before development and visualisation using a G:Box system (SynGene) set with a chemiluminescence filter for detection of ECL substrates.

2.15.6 Stripping the Blot

To store and re-use the membrane, any bound antibody was removed. This was done by washing the blot twice with stripping buffer (0.2 M sodium chloride, 1% (v/v) SDS) at 37 °C in a hybridisation oven (HL-2000 Hybrilinker, UVP). The membrane was then washed in 2x SSC (200 mM sodium chloride, 30 mM sodium citrate, pH 7.0) for 10 minutes with gentle agitation and stored in 2x SSC at -20 °C. For re-use the protocol was restarted from the hybridisation stage after UV crosslinking.

Chapter 3: Identification of Allelic Expression Imbalance

3.1 Introduction

In this chapter, RNA sequencing techniques have been used to identify genes with significant levels of allelic expression imbalance in the wild-type *Candida albicans* strain SC5314. From here, the list of significant genes is investigated for trends which may be of biological relevance such as Gene Ontology and chromosomal location, with a focus upon alleles that may have divergent functions. Expanding upon this point, structural factors which have previously been shown to influence gene expression levels, such as GC content and gene length, are explored in relation to AEI. The patterns observed within these structural factors are then discussed alongside a brief description of other factors such as DNA methylation and RNA decay rates, which are outside of the scope of this project, but may still be of importance. Finally, validation of AEI identified through RNA sequencing is attempted using qPCR, restriction enzyme digests and western blotting. Since these methods proved to be largely unsuccessful a discussion of issues surrounding allele specificity and sensitivity is presented. Firstly, a brief introduction to gene expression analysis with RNA sequencing technologies is given, followed by an introduction into the relationship between structural factors and gene expression levels.

3.1.1 RNA sequencing

The development of next generation sequencing technologies has paved the way for advances in RNA sequencing; a technique which uses quantification of mRNA levels to determine gene expression patterns. This is achieved through a simple process in which RNA is extracted from the tissue of interest and then, according to the first RNA sequencing workflows, converted from RNA to cDNA followed by fragmentation of the cDNA and library preparation (Nagalakshmi *et al.*, 2008). More modern techniques directly fragment the RNA and the next generation sequencing is undertaken and reads are aligned to a reference

genome with quantification at each gene location. The resulting information on gene expression is termed the transcriptome (Figure 3.1) (Wang *et al.*, 2009).

RNA sequencing as a method to monitor gene expression levels has distinct advantages over existing methods. Approaches such as northern blotting and qPCR determine expression levels on a gene-by-gene basis whereas RNA sequencing and microarrays provide information on a genome-wide scale, allowing for a much more in depth analysis of overall patterns of gene expression. Sequencing of RNA also has benefits over the use of microarrays including a higher sensitivity than microarray hybridisation allowing for identification of genes with the lowest expression levels (Wang *et al.*, 2009, Tuch *et al.*, 2010a), greater specificity, low background noise with no upper limit for detection, high accuracy (Wang *et al.*, 2009),

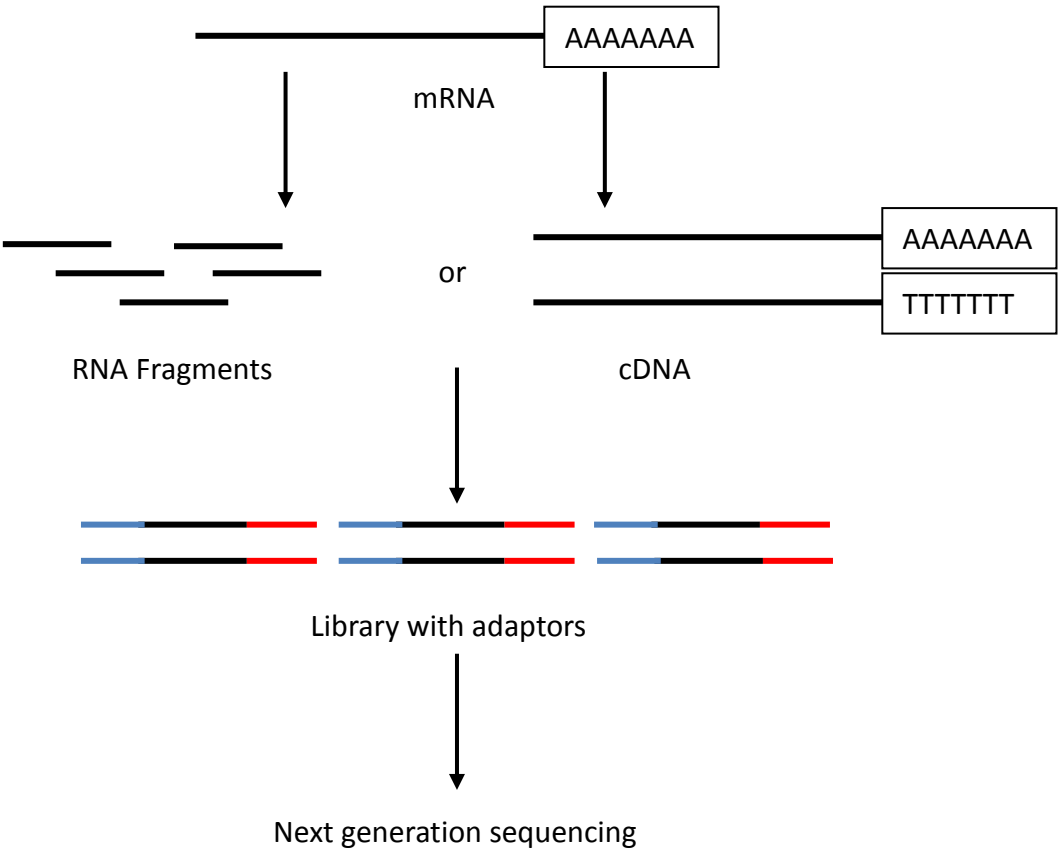


Figure 3.1. Schematic of key steps in RNA sequencing process. RNAs are extracted and converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are added to each cDNA fragment, creating the cDNA library. Short sequences are obtained from each cDNA using high-throughput sequencing technology. Figure adapted from Wang *et al.* 2009.

and high levels of reproducibility (Marioni *et al.*, 2008, Wang *et al.*, 2009). Validation of expression levels determined by RNA sequencing is seen through a high correlation between RNA sequencing and qPCR in various species including yeast and human tissues (Marioni *et al.*, 2008, Nagalakshmi *et al.*, 2008, Bloom *et al.*, 2009, Bruno *et al.*, 2010, Tuch *et al.*, 2010a) as well as a high correlation in results observed between RNA sequencing and microarrays (Marioni *et al.*, 2008, Bloom *et al.*, 2009, Esteve-Codina *et al.*, 2011, Guida *et al.*, 2011). Unlike microarrays, RNA sequencing assumes no prior knowledge of coding sequence, creating the possibility for identification of novel transcripts and introns. With the recent development of *de novo* transcriptome assembly, the need for a reference genome is also overcome, vastly increasing the number of organisms in which RNA sequencing can be undertaken (Wang *et al.*, 2009). As microarrays are susceptible to cross-hybridisation of homologous DNA fragments, such as those produced by pairs of alleles or paralogous genes (Tuch *et al.*, 2010a), RNA sequencing has a distinct advantage for detection of events such as allelic expression imbalance (AEI).

This study aims to exploit these advantages of RNA sequencing to determine allelic expression imbalance in the yeast *Candida albicans*. As detailed in section 1.5, RNA sequencing has been used in numerous species to identify allele-specific expression. This includes identification of imprinting within mouse brain tissue (Gregg *et al.*, 2010) and within triploid maize endosperm (Zhang *et al.*, 2011), identification of AEI in human T cells (Heap *et al.*, 2010), human cancer cells (Tuch *et al.*, 2010a), in two pig species (Esteve-Codina *et al.*, 2011), in yellow baboons (Tung *et al.*, 2011) and determination of *cis*-acting polymorphisms in hybrid yeast strains (Bullard *et al.*, 2010b).

Throughout these studies, a lack of consistency is seen with respect to the exact method used to determine AEI. Although all studies use RNA sequencing, the downstream processing of the results varies dramatically, from the choice of alignment software, to the choice of a control, to the statistical analysis of allele expression levels. One consistent factor is that all studies have used a haploid reference genome and incorporated SNP identification into their analysis. Using this method for identification of AEI has proven to be problematic, as mapping bias towards the reference genome, and against the alternative SNP, even with

SNP masking, is often reported (Degner *et al.*, 2009, Stevenson *et al.*, 2013). For *Candida albicans,* a diploid reference genome is available. Therefore this allows us to take a new approach and align reads directly to each allele at unique SNP positions, avoiding issues of bias towards the reference genome.

RNA sequencing has been used in numerous yeast species to determine overall expression levels including the fission yeast *Schizosaccharomyces pombe* (Wilhelm *et al.*, 2008), where expression of more than 90% of the genome was detected; and *S. cerevisiae* (Nagalakshmi *et al.*, 2008), where a lower level of gene expression, at 74.5% of the genome, was detected. In *C. albicans,* various studies have used RNA sequencing to explore gene expression levels. In 2010, Bruno *et al.* investigated the differential transcriptional response of the wild-type *C. albicans* strain SC5314 to nine different infection specific *in vitro* conditions. Haploid alignments were carried out and expression was detected for 97% of previously annotated ORFs plus 602 newly identified transcripts. Analysis of the UTRs was also carried out alongside intron discovery using TopHat (Bruno *et al.*, 2010). Other RNA sequencing reports in *C. albicans* include a study of the transcriptional profile of *C. albicans* cells and bone marrow derived dendritic cells from mice (Tierney *et al.*, 2012), a study looking at the transcriptional control of biofilm formation (Nobile *et al.*, 2012), a comparison of the transcriptome of *C. albicans* and *C. dubliniensis* chlamydospores to identify chlamydospore specific genes (Palige *et al.*, 2013), and a strand-specific RNA sequencing study looking at the transcriptional differences between white and opaque cells (Tuch *et al.*, 2010b). Closely related *Candida* species have also undergone RNA sequencing including *C. parapsilosis* under both normal and hypoxic growth conditions (Guida *et al.*, 2011).

As mentioned in section 1.8, a recent paper by Muzzey *et al.* used the RNA sequencing data published by Bruno *et al.* (2010) alongside an improved phased reference genome to demonstrate that identification of allelic expression imbalance in *Candida albicans* is achievable (Muzzey *et al.*, 2013). As mentioned above, the diploid reference genome negated the need for SNP identification before AEI is determined. However, this paper only investigated the levels of AEI from a single growth condition and very little biological or

functional inference was made from the results. In a follow up study, the genome-wide extent of allelic expression imbalance at both the transcriptional and translational levels in *C. albicans* was evaluated (Muzzey *et al.*, 2014). In the work presented here, a similar method of identification of AEI has been adopted, but with a focus upon the biological impact of these expression levels.

3.1.2 The Relationship Between Structural Factors and Gene Expression Levels

Gene expression levels are often attributed to the control of promoters and transcription factors. However, there is evidence that structural factors of the genome may have an impact upon transcription. These factors include chromosomal location (Muller, 1930, Gottschling *et al.*, 1990), overlap with a neighbouring open reading frame (Cullen *et al.*, 1984, MacIsaac *et al.*, 2011), GC content (Goncalves *et al.*, 2000, Urrutia and Hurst, 2003, Versteeg *et al.*, 2003), gene length (Coghlan and Wolfe, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003) and codon usage (Sharp and Li, 1987, Morton, 1993, Merkl, 2003).

3.1.2.1 Chromosomal Location

The chromosomal location of a gene has been shown to impact upon expression levels due to the uneven distribution of heterochromatin across chromosomes. This concept was first identified in the early 1930s by Muller, who showed that the phenotypic characteristics of *Drosophila melanogaster* change when a gene is relocated to a region of denser heterochromatin. This phenomenon was titled ‘Position Effect Variegation’ (Muller, 1930). Dense heterochromatin is associated with silent regions of DNA due to the inaccessibility of DNA binding factors and transcriptional machinery.

In some yeast species, silencing of expression due to heterochromatin is seen in genes located in sub-telomeric regions. In *S. cerevisiae,* it has been shown that repositioning of genes to telomeric regions results in disruption of expression (Gottschling *et al.*, 1990) and in *C. glabrata* a group of paralogous adhesin genes known as the *EPA* genes are transcriptionally repressed due to their sub-telomeric positioning. This silencing is achieved through recruitment of silencing machinery to the chromatin structure (Castaño *et al.*, 2005). In

Candida albicans some cases of genes from within a gene family that cluster to the same chromosomal locations have been observed (Braun *et al.*, 2005), however little evidence has been seen for sub-telomeric silencing.

Contradictory evidence has been seen for the relationship between chromosomal location and allelic expression imbalance. Savova *et al.* (2013) claim that there is little clustering in terms of autosomal monoallelic genes, however some special cases, such as the olfactory genes, do cluster to a specific chromosomal location. On the other hand, a study in maize showed that imprinted genes do cluster in location when compared to the overall frequency of genes across the entire genome, where a definition of at least two genes within 1 Mb of each other was used (Zhang *et al.*, 2011). In mammals, allele-specific heterochromatin patterns have been observed, along with allele-specific DNA methylation. In this case, these factors have been elucidated as the control mechanism behind imprinting of genes (Singh *et al.*, 2011). Looking at allelic expression imbalance as opposed to imprinting, Lo *et al.* (2003) showed that some genes with AEI clustered in chromosomal location but most were randomly distributed across the genome.

3.1.2.2 Overlapping Genes

The distance between neighbouring open reading frames has been shown to influence expression levels of both of the genes. Studies in various organisms have shown that when genes on the same strand with the same orientation overlap with each other, transcriptional interference occurs where the transcription of one gene is repressed by the overlapping gene (Cullen *et al.*, 1984, Bateman and Paule, 1988, Irniger *et al.*, 1992). Although these studies detailed the impact of overlapping genes on the same strand, transcriptome analysis of the yeast *S. cerevisiae* revealed that overlapping of transcripts on either strand could contribute to observed impacts upon gene expression (Nagalakshmi *et al.*, 2008).

In terms of allelic expression imbalance, the impact of overlapping transcripts has been examined using strand specific microarrays in a heterozygous strain of *S. cerevisiae*. In total 196 pairs of transcripts overlapped on opposite strands (sense-antisense), and 36 of these demonstrated AEI. Both symmetric and

asymmetric patterns of AEI were observed in terms of the strands. For example *FET4* showed symmetric expression with both sense and antisense strands being expressed on one chromosome and silenced upon the other. Conversely *DAP2* showed asymmetric expression with one chromosome showing strong expression of the sense strand and weak antisense expression whereas the other chromosome had the opposite expression pattern (Gagneur *et al.*, 2009). However, it should be noted here that 335 genes with AEI showed no evidence of overlap with their neighbouring open reading frame, suggesting that overlap is not the sole cause of allelic expression imbalance.

Another example of overlapping transcripts influencing AEI is the imprinted gene *Mest* (as described in section 1.5.1.2). *Mest* has two different transcripts, *Mest* and the longer transcript *MestXL.* In both cases the paternal allele is expressed and the maternal allele is silenced. Additionally as *MestXL* is longer, overlap with the antisense gene *Copg2* occurs causing silencing of the paternal allele via transcriptional interference (MacIsaac *et al.*, 2011). The exact mechanism of transcriptional interference is unclear but possible suggestions include collision of the elongation complexes of both mRNAs, overlapping causing RNA editing of adenosine to inosine marking the *Copg2* RNA for degradation, or the termination complex of *MestXL* directly interfering with the chromatin of the *Copg2* promoter.

3.1.2.3 GC Content and Gene Length

GC content and gene length are structural factors determined by the DNA sequence of the genes themselves. These factors have been correlated with gene expression in humans (Urrutia and Hurst, 2003, Versteeg *et al.*, 2003) and yeast (Marín *et al.*, 2003) with mixed results.

Some studies have found a negative correlation between GC content and gene expression (Goncalves *et al.*, 2000). The causal relationship is not yet determined, but this correlation supports the hypothesis that a sequence with higher GC content requires more energy to unwind and is therefore transcribed less efficiently. Alternatively, this negative correlation between GC content and gene length may be due to chromatin organisation. It has been demonstrated in *S. cerevisiae* that low GC content tracts of dA:dT are rigid and therefore unable

to bend around and bind nucleosomes. These nucleosome free regions are then associated with increased accessibility for the transcriptional machinery and have higher gene expression (Mavrich *et al.*, 2008). On the other hand, some studies have also indicated a positive relationship, with expression increasing as GC content is increased (Urrutia and Hurst, 2003, Versteeg *et al.*, 2003). Contrarily it has been suggested that differences in chromatin structure in GC low areas result in lower transcription levels (Marín *et al.*, 2003). The relationship between ORF length and expression is far more straightforward with the general conclusion being that shorter genes are more transcriptionally efficient and are therefore expressed to higher levels (Coghlan and Wolfe, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003).

Although the relationships between GC content, gene lengths and gene expression levels have been investigated, it has yet to be elucidated what impact these structural factors have upon allelic expression imbalance.

3.1.2.4 Codon Usage

Codon usage is another structural factor which has been linked to gene expression. In yeast, it was discovered that genes with higher expression levels tend to use codons corresponding to the most abundant tRNA species (Bennetzen and Hall, 1982). Since then, various measures of codon usage have been developed which claim to predict gene expression levels including the Codon Adaptation Index (CAI) (Sharp and Li, 1987), the Codon Bias Index (CBI) (Morton, 1993), and GCB (Merkl, 2003). Another factor which complicates codon usage in *Candida albicans* is the CUG codon. Typically this is translated as a leucine, but in certain *Candida* species, including *C. albicans,* the CUG codon is now translated as serine (as discussed in section 1.2) (Ohama *et al.*, 1993).

3.1.3 Aims of this Chapter

This chapter aims to address three main research questions:

1. Does AEI occur on a genome-wide scale in SC5314, the wild-type strain of *C. albicans*?
2. Are patterns present in the structural factors of genes identified with AEI?

3. Can a method be developed to validate AEI identified by RNA sequencing?

3.2 Materials and Methods

3.2.1 RNA Sequencing

3.2.1.1 Cell Harvests

Colonies of SC5314 were inoculated in 10 ml of YPD and grown overnight at 30 °C, 180 rpm, in biological triplicate. Cells were then diluted to an optical density at 600 nm of 0.25 in 10 ml fresh YPD. Cell cultures were incubated at 30 °C, 180 rpm, until the optical density at 600 nm equaled 1.0. Cell density was then estimated using a haemocytometer (mean 5 x 10^7 cells/ml) and cells from each replicate were harvested by centrifugation at 4000 rpm for one minute before re-suspension in 200 µl of sterile water, cells were then frozen in liquid nitrogen and stored at -80 °C.

3.2.1.2 RNA Preparation

Total RNA samples were obtained using a Qiagen RNeasy mini-kit according to the manufacturer's instructions for yeast using mechanical disruption including an on column DNase digestion using RNase-free DNase set (Qiagen). 5 x 10^7 cells were disrupted with 0.4 – 0.6 mm glass beads using a FASTPREP-24 bead beater (MP) (3 x 20 seconds at 6 m/s). RNA was quantified using a NanoDrop spectrophotometer and purity assessed using an Agilent 2100 Bioanalyser (samples prepared according to the manufacturer's instructions). Double stranded cDNA was created using reverse transcription with random hexamers using a Thermoscript™ RT-PCR system. Samples were sonicated (bioruptor sonicator) and prepared for Illumina GA2 sequencing using standard NEB protocols by the University of Exeter Sequencing Service.

3.2.1.3 Illumina Base Calling and Pipeline

The Illumina 1.4 pipeline was used to analyse images from the GA2 instrument. Base calling was carried out using the Bustard module with a standard chastity filter applied. 76 bp reads were obtained and trimmed to 60 bp to minimise overall error rates. This yielded 16,166,757 reads for replicate 1, 17,369,675 for replicate 2 and 16,853,362 for replicate 3. Raw sequence data is available from

the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) accession number [GEO:GSE35233].

3.2.1.4 Alignment and Identification of Allelic Expression Imbalance

CLCBio software (www.clcibio.com) was used to align reads to the reference Assembly 19 diploid genome sequence (Jones *et al.*, 2004). Standard parameters were set to allow up to two mismatches per read. Non-unique reads were discarded so that only reads which could unambiguously differentiate between alleles were counted. From this data, the normalised measure of expression, RPKM (reads per kilobase per million mapped reads) (Mortazavi *et al.*, 2008) was calculated for each allele. Allele expression of a single gene was compared using Fisher's-Exact test, with a cut off of 2x fold difference in expression at a p-value of <0.0000077 (set using Bonferroni correction). Alongside this, a haploid alignment against Assembly 21 (Van Het Hoog *et al.*, 2007) was undertaken based on both unique and non-unique reads indicating overall gene expression levels. Methods from 3.2.1.1 to 3.2.1.4 were kindly carried out by the University of Exeter Sequencing Service.

3.2.2 Gene Ontology (GO) Analysis

Analysis for over representation of Gene Ontology (GO) terms within genes identified to have AEI was carried out using "CGD GO Term Finder" at the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012). Lists of GO terms were created using "CGD Gene Ontology Slim Mapper" also available from the *Candida* genome database.

3.2.3 Calculation of Differences in Promoter Sequence

As promoter sequences in *C. albicans* are currently undefined, the 1000 bp of DNA sequence upstream of each allele were downloaded from the *Candida* Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012). If the neighbouring open reading frames were within the 1000 bp upstream, the sequence up to the neighbouring ORF was taken. Sequences of each pair of alleles of a gene were aligned used ClustalW (http://www.genome.jp/tools/clustalw/) (Kyoto University Bioinformatics Center, 2010) and sequences were recorded as different if one of more SNPs or INDELs were observed. For comparison, an equivalent set of genes with equal

allele expression (fold difference ≈ 1) (Appendix I Table II) were also analysed. The probability of observing SNPs across a region of 1000 bp was calculated based upon the observed average level of heterozygosity of one SNP in every 237 bp (Jones *et al.*, 2004). The observed and expected number of promoter sequences with SNPs was compared statistically using a chi-square test.

3.2.4 Calculation of Percentage Protein Identity

Protein sequences for all genes with allelic expression imbalance identified in this study and for an equivalent set of genes with equal allele expression (fold difference ≈ 1) (Appendix I Table II) were downloaded from the *Candida* Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012).

To assess if AEI could have a link to differences in allele function, percentage protein identity comparisons between alleles were carried out using a Needleman-Wunsch algorithm for global alignment (Rose and Eisenmenger, 1991). Arc-sine transformations were carried out to normalise the data (Osborne, 2005) before statistical analysis was undertaken, using a Student's t-test, to compare the percentage protein identities of genes with AEI and genes with equal allele expression.

3.2.5 Analysis of Structural Factors

3.2.5.1 Chromosomal Locations and Identification of Overlapping Genes

Chromosomal coordinates and strand information were obtained from the 'Chromosomal Features File' available from the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012). Gene locations were calculated by converting start coordinates to a percentage of overall chromosome length. Clusters were identified by a significant deviation from a Poisson distribution. Over- or under-representation of genes with AEI on each chromosome was statistically analysed using a chi-square test. As the 'Chromosomal Features File' only lists the chromosomal coordinates for "allele one" of an allele pair, this was carried out for just one allele, with the assumption that the second allele will have the same chromosomal location and strand orientation.

Chromosomal coordinates were also used to calculate the number of genes across the entire genome which overlap with the neighbouring open reading

frame. Stop coordinates of a feature were subtracted from the start coordinate of the next feature to determine the distance between the features in base pairs. If this number was negative, the features were classified as overlapping. This included all features of the *C. albicans* genome, therefore centromeres, repeat regions etc. were included as well as ORFs. Over- or under-representation of features which overlap on each chromosome was statistically analysed using a chi-square test. The RPKM values, and therefore expression levels, of a pair of overlapping genes were statistically compared using a Fisher's Exact test in R version 2.12.0 (The R Foundation for Statistical Computing, 2010) at a p-value < 0.000568 (set using Bonferroni correction). The relationship between strand and expression difference in a pair of overlapping genes was statistically compared using a chi-square test.

For comparison of overlapping genes between pairs of alleles with AEI, chromosomal coordinates were manually obtained for both alleles and neighbouring alleles using "GBrowse for *C. albicans* SC5314 Assembly 19" available from the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012). Hypergeometric distribution analysis was used to statistically assess the frequency of genes with AEI which overlap with their neighbouring open reading frame.

3.2.5.2 GC Content, Gene Length and Codon Usage

FASTA files containing the genomic open reading frame sequences, as annotated in Assembly 19 (Jones *et al.*, 2004), were downloaded from the *Candida* Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012).

Open reading frame lengths were obtained from the 'Chromosomal Features File' available from the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012) for "allele one" and calculated manually for "allele two". GC content was calculated using the gene sequences and BioEdit version 7.0.9.0 (Hall, 1999). Codon usage frequencies for the entire genome were calculated using INCA version 2.1 (Supek and Vlahovi ek, 2004). Two codon usage measures were used in this study, GCB (Merkl, 2003) and the Codon Adaptation Index (CAI) (Sharp and Li, 1987). CAI values for each allele were calculated manually based on the formulae in Sharp and Li (1987) using the

codon usage frequencies calculated by INCA. CUG usage for each ORF was calculated using in-house PERL scripts (Appendix II.I).

Before the analysis of the impact of each structural factor upon AEI was undertaken, overall gene expression levels were correlated with each structural factor to assess the relationship between each factor and expression. Overall gene expression levels were determined by alignment of reads against the haploid reference genome, as described in section 3.2.1.4, and normalised to RPKM values. For each gene, the GC content, ORF length, CAI value, GCB value and CUG usage was calculated as the average of the measure from each allele. Correlations were statistically analysed using a Spearman's correlation in SPSS version 21 (IBM, 2012).

To determine the contribution of each structural factor on AEI, the values obtained from the allele with the lowest and highest expression were statistically compared using a Student's t-test. Equivalent comparisons were undertaken using a set of 210 genes with equal allele expression (fold difference ≈ 1) (Appendix I Table II) to reveal if any effects were restricted to genes with AEI. ORF lengths were normalised using a logarithmic transformation prior to statistical analysis so that a more stringent parametric test could be used (Osborne, 2005).

In general, for genes with AEI and genes with equal allele expression, the difference in the values obtained from each allele for all structural factors was calculated by subtracting the allele one value from the allele two. However for ORF length the fold difference was calculated by dividing the length of the longest allele by the length of the shortest allele. The difference in variance between the data obtained from genes with AEI and genes with equal allele expression was statistically analysed using an F-test. Correlations between differences in structural factors and percentage protein identities (section 3.2.3), and between fold difference in ORF length and all other structural factors were statistically analysed using a Spearman's correlation in SPSS version 21 (IBM, 2012).

3.2.6 Validation of Allelic Expression Imbalance

3.2.6.1 Allele-Specific qPCR

3.2.6.1.1 Cell Extractions

To accurately validate expression levels, cells were grown and harvested in the same manner as when obtained for RNA sequencing, as follows. A single colony of the wild-type strain SC5314 was grown overnight in 5 ml of YPD at 30 °C, 180 rpm. Cell concentrations were calculated using a haemocytometer and 5 x 10^6 cells/ml were inoculated, in triplicate, in 50 ml of YPD and grown at 30 °C, 180 rpm, until optical density at 600 nm was equal to one.

To harvest the cells, an amount equal to an optical density at 600 nm of 0.2 were taken, pelleted at 4000 rpm for five minutes and resuspended in 1 ml of PBS. Cells were transferred to a 1.5 ml screw-topped microfuge top and again pelleted at 13000 rpm for five minutes. The supernatant was removed, cells were flash frozen using liquid nitrogen and stored at – 80 °C until use. The cells were then immediately ready for RNA extraction with the correct cell number.

3.2.6.1.2 RNA Extractions

Total RNA was extracted using the RNeasy Mini Kit (Qiagen) as per manufacturer's instructions for yeast, using mechanical disruption including an on column DNase digestion using RNase-free DNase set (Qiagen). RNA concentration and quality was quantified using a NanoDrop Spectrophotometer, formaldehyde agarose gel electrophoresis (see section 3.2.6.1.3) and an Agilent 2100 Bioanalyser (samples prepared as per manufacturer's instructions).

3.2.6.1.3 Formaldehyde Agarose Gel Electrophoresis

To assess the integrity of RNA samples, formaldehyde agarose gel electrophoresis was used. To summarise, 1.2% (w/v) agarose was melted in 10% (v/v) formaldehyde agarose gel buffer (200 mM MOPS, 50 mM sodium acetate, 10 mM EDTA, pH 7.0). 1.8% (v/v) 37% formaldehyde and 1 µl of ethidium bromide were added and mixed. The gel was then poured, allowed to set and equilibrated in 1x formaldehyde agarose gel running buffer (10% (v/v) formaldehyde agarose gel buffer, 2% (v/v) 37% formaldehyde) for 30 minutes.

RNA samples were diluted in a 4:1 ratio in 5x RNA loading buffer (16 µl saturated aqueous bromophenol blue solution, 80 µl 500 mM EDTA, pH 8.0, 720 µl 37% formaldehyde, 2 ml 100% glycerol, 3.084 ml formamide, 4 ml formaldehyde agarose gel buffer, 100 µl RNase-free water). This was followed by heating to 65 °C for five minutes before cooling on ice. Samples were then run at 5 v/cm in 1x formaldehyde agarose gel running buffer for four hours.

3.2.6.1.4 cDNA Preparation

cDNA was prepared using the Superscript VILO cDNA synthesis kit (Invitrogen). Reactions were prepared as follows: 4 µl of 5x VILO reaction mix, 2 µl of 10x Superscript enzyme mix and 2.5 µg of total RNA. Diethylpyrocarbonate (DEPC) -treated water was used to make the final volume of the reaction 20 µl. The reactions were carried out using a PCR machine with the following cycle: 25 °C for 10 minutes, 42 °C for 2 hours, 85 °C for 5 minutes, and then stored at – 20 °C.

3.2.6.1.5 qPCR using TaqMan Probes

Allele-specific qPCR was attempted using TaqMan Genotyping Assays-By-Design (Applied Biosystems) with two genes, *CDC6* and *VPS1.* The probes consisted of a pair of forward and reverse oligonucleotides which amplified a region of 150-200 bp and two fluorophore tagged oligonucleotides which bound in between this region, one FAM tagged and specific to one allele, the other VIC tagged and specific to the other allele. This technique had previously been proved successful by Harries *et al.* (2006) and Tuch *et al.* (2010a). For a full list of oligonucleotides used see Table 3.1. These probes were diluted to a 20x working concentration using 1 x TE buffer and stored in 20 µl aliquots at -20 °C.

To analyse the specificity and efficiency of the probes, the qPCRs were first tested using genomic DNA (see section 2.13) from the wild-type strain SC5314, where both probes should be equally detected, and from appropriate heterozygous knockout strains (see Table 2.1), where only one probe should be detected. Heterozygous strains were constructed using transformation and homologous recombination of a cassette as detailed in sections 2.8 – 2.11. DNA was tested in 10x fold dilutions ranging from 100 ng to 0.0001 ng in triplicate. Sterile water was used as a negative control.

Reactions were set up as follows: 5 µl of TaqMan Universal PCR mastermix (Applied Biosystems), 0.5 µl of 20x TaqMan Assay (Applied Biosystems) and 4.5 µl of DNA diluted in 1 x TE buffer.

qPCR was carried out using a Stratagene Mx3005P set to detect fluorescence from FAM and HEX (equivalent to VIC) fluorophores. The cycle was set as follows:

95 °C for 10 minutes
92 °C for 15 seconds } 40 cycles
60 °C for 1 minute }

To calculate the efficiency of the primer sets, ΔCt values were plotted against the concentration of DNA standards from SC5314. Efficiency was calculated using the gradient of this line (Δx/Δy) and the following calculation taken from Applied Biosystems:

$$\text{Efficiency} = \left(\left(10^{-\frac{1}{\text{gradient}}} \right) - 1 \right) \times 100$$

Table 3.1 TaqMan genotyping assay-by-design oligonucleotides used for allele-specific qPCR

Name	Sequence 5' – 3'	Assay	Position[1]	Probe
CDC6_F	TTTGCCATACAATGCTGATCAAATTAAATCA	*CDC6* 1	818 – 838	-
CDC6_R	TGTATAGCACCCGGGTGGAA		888 – 907	-
CDC6_V ALLELE 1	ATTATCTAATCTTAAACAAGAAAT		863 – 886	VIC
CDC6_M ALLELE 2	ATTATCTAATCTTAAACAAGAGAT		863 – 886	FAM
VPS1_F	AGAGTACTTTACCTGACATCAAGATGAGA	*VPS1* 1	947 – 975	-
VPS1_R	AGATGCCAGTGTAATCTTTGGAGAAAT		1073 – 1099	-
VPS1_V ALLELE 1	AAGCATTGATAATTCCTG		1003 – 1020	VIC
VPS1_M ALLELE 2	AGCATTGATAATTCTTG		1003 – 1019	FAM
CDC6_F_2	GGTGATTTGAGAAAGGCATTTGATATATG	*CDC6* 2	939 – 967	-
CDC6_R_2	TGATTTTCATTATTGCCAAAAGATGTCATACAAA		1051 – 1084	-
CDC6_V_2 ALLELE 1	AGTTGTCAAGGTACTGATACG		996 – 1016	VIC
CDC6_M_2 ALLELE 2	TTGTCAAGGTACCGATACG		998 – 1016	FAM
VPS1_F_2	ACCCATCATACAGAGCCAAAGC	*VPS1* 2	863 – 884	-
VPS1_R_2	CTTGATGTCAGGTAAAGTACTCTTGATGT		941 – 969	-
VPS1_V_2 ALLELE 1	CTGTGGTACGCCTTAC		891 – 906	VIC
VPS1_M_2 ALLELE 2	TTCTGTGGTACTCCTTAC		889 – 906	FAM
VPS1_F_3	AAAGTCCCCGTTGGTGATCAG	*VPS1* 3	502 – 522	-
VPS1_R_3	TGACAGACAAGATAATGGCGTTAGG		577 – 601	-
VPS1_V_3 ALLELE 1	AAGATATTGAAAGGCAAATC		527 – 546	VIC
VPS1_M_3 ALLELE 2	AGATATTGAAAGACAAATC		528 – 546	FAM

1. Relative to gene start position = 1.

3.2.6.1.6 qPCR using SYBR® Green

Allele-specific qPCR was attempted using a conventional SYBR® green technique with primers which amplify each allele specifically.

Oligonucleotides were designed for three genes: *VPS1, CDC6* and *RBT4.* Specificity for each allele was achieved by designing the oligonucleotides to bind at SNP positions. Specificity of a range of oligonucleotide combinations were first computationally tested using BLASTn (http://blast.ncbi.nlm.nih.gov/Blast.cgi) before testing using PCR (as described in section 2.7). Table 3.2 lists all of the oligonucleotides tested. Genomic DNA extracted from either the wild-type strain SC5314 or from an appropriate heterozygous knockout strain (see Table 2.1) was used as template DNA. Heterozygous strains were constructed using transformation and homologous recombination of a cassette as detailed in sections 2.8 – 2.11. PCR products were analysed for specificity using DNA gel electrophoresis (see section 2.5).

Table 3.2 Oligonucleotides analysed for allele-specificity for use in allele-specific qPCR

Name	Sequence 5' – 3'	Position[1]	Gene
VPS1-F	CAGGGACGCATTGAAAGA	822 – 839	*VPS1*
VPS1-1-R	GAGATTCGGCCATTTCTG	1025 – 1042	*VPS1*
VPS1-2-R	GGAGACTCAGCCATTTCAG	1025 – 1043	*VPS1*
VPS1-2-R-2	AGGAGACTCAGCCATTTCA	1026 – 1044	*VPS1*
VPS1-F1-1A	GGTGGAGGGTCATCCTCG	58 – 75	*VPS1*
VPS1-F1-2A	GGGGGAGGATCATCCTCG	58 – 75	*VPS1*
VPS1-R1-1A	TTCCCTGGCAAATGCAAA	327 – 344	*VPS1*
VPS1-R1-2A	TTCCCCGGCAAATGCA	329 – 344	*VPS1*
VPS1-F23-1A	TTGCATTTGCCAGGGAA	328 – 344	*VPS1*
VPS1-F23-2A	TGCATTTGCCGGGGA	329 – 343	*VPS1*
VPS1-R2-1A	CTTTGATTTGCCTTTCAATATC	479 – 500	*VPS1*
VPS1-R2-2A	CATATCTTTGATTTGTCTTTCAATATC	479 – 505	*VPS1*
VPS1-R3-1A	GGCGTTAGGCTTGGAAA	569 – 585	*VPS1*
VPS1-R3-2A	AATGGCGTTAGGCTTAGAAATA	567 – 590	*VPS1*
VPS1-F45-12A	CTATCAGGGACGCATTGAAAG	818 – 838	*VPS1*
VPS1-R4-1A	ATTTCTTCAATGAATGTTCGATTC	974 – 997	*VPS1*
VPS1-R4-2A	ACTTCTTCAATGAATGCTCAATTC	974 – 997	*VPS1*
VPS1-R5-1A	GATTCGGCCATTTCTGGTC	1022 – 1040	*VPS1*
VPS1-R5-2A	GACTCAGCCATTTCAGGTCC	1021 – 1040	*VPS1*
VPS1-F6-1A	GAATCGAACATTCATTGAAGAAAT	974 – 997	*VPS1*
VPS1-F6-2A	GAATTGAGCATTCATTGAAGAAGT	974 – 997	*VPS1*
VPS1-R678-12A	TGGCATTAACCCCATTCTTG	1179 – 1198	*VPS1*
VPS1-F7-1A	CCAGAAATGGCCGAATCTC	1024 – 1042	*VPS1*
VPS1-F7-2A	CCTGAAATGGCTGAGTCTCC	1024 – 1043	*VPS1*
VPS1-F8-1A	TACCAACAGGAATTATCAATGCTT	997 – 1020	*VPS1*
VPS1-F8-2A	TATCAACAAGAATTATCAATGCTTG	997 – 1021	*VPS1*
CDC6-F1-1A	CCAATAACTCACGAGTAGATAAT	89 – 111	*CDC6*
CDC6-F1-2A	TCCAAAAAACTCAAAATAGATAAT	90 – 111	*CDC6*
CDC6-R12-1A	CATTAGAATAGTTCACGTTGG	256 – 276	*CDC6*
CDC6-R12-2A	CATTAGAATGACTCACGTTGG	256 – 276	*CDC6*
CDC6-F2-1A	TCTGTTGACTACTGGGAACC	118 – 137	*CDC6*
CDC6-F2-2A	CTCTTTTGATACAGGGAAACC	117 – 138	*CDC6*
CDC6-F3-1A	CCAAAATACCTTCAACTTCAC	180 – 200	*CDC6*
CDC6-F3-2A	CCAAAAAACCTTCAACTTCAC	180 – 200	*CDC6*
CDC6-R3-1A	GCAGTCTTACCCGTCCC	384 – 400	*CDC6*
CDC6-R3-2A	GCAGTTTTACCAGTGCCTGG	381 – 400	*CDC6*

CDC6-F4-1A	CCAAGATTAGTGAGAAATAATATTC	774 – 798	*CDC6*
CDC6-F4-2A	CCAAGATTAGTTAGAAATAATATTC	774 – 798	*CDC6*
CDC6-R4-1A	CAAATCACCAGATATTGATGC	927 – 947	*CDC6*
CDC6-R4-2A	ATCACCGGATATTGATGC	927 – 944	*CDC6*
CDC6-F5-1A	CAAGAAATTTTCCACCCG	879 – 896	*CDC6*
CDC6-F5-2A	CAAGAGATCTTCCACCCG	879 – 896	*CDC6*
CDC6-R5-1A	TTTGCCACATGTTGAATCA	1030 – 1048	*CDC6*
CDC6-R5-2A	TTTAGCCACATGTTGAATCA	1030 – 1049	*CDC6*
CDC6-F6-1A	AATATCTGGTGATTTGAGAAAG	932 – 953	*CDC6*
CDC6-F6-2A	AATATCCGGTGATTTGAGAAAG	932 – 953	*CDC6*
CDC6-R6-1A	AGTTTTGTTGGAACATTTCG	1156 – 1175	*CDC6*
CDC6-R6-2A	AGTTTTGTTGGAACGTTTCG	1156 – 1175	*CDC6*
CDC6-F7-1A	CAAGGTACTGATACGATTAATAAA	1002 – 1025	*CDC6*
CDC6-F7-2A	CAAGGTACCGATACGATTAATAAA	1002 – 1025	*CDC6*
CDC6-R7-1A	TTTTAACAGTCCAATCAAATTA	1244 – 1265	*CDC6*
CDC6-R7-2A	TTTCAACAGTCCAATCAAATCA	1244 – 1265	*CDC6*
CDC6-F8-1A	CGAAATGTTCCAACAAAACT	1156 – 1175	*CDC6*
CDC6-F8-2A	CGAAACGTTCCAACAAAACT	1156 – 1175	*CDC6*
CDC6-R8-12A	CAACAAAATACAACTACTACTTTCC	1248 – 1272	*CDC6*
CDC6-R1M-1A	AGTTCACGTTGGATCCAC	250 – 267	*CDC6*
CDC6-R1M-2A	GACTCACGTTGGATCCAC	250 – 267	*CDC6*
CDC6-F5M-1A	AATTTTCCACCCGGGT	884 – 899	*CDC6*
CDC6-F5M-2A	GATCTTCCACCCGGGT	884 – 899	*CDC6*
CDC6-R7M-2A	CAACAGTCCAATCAAATCAT	1243 – 1262	*CDC6*
RBT4-F1-1A	CAACTGTCACTGGTGGTGAC	137 – 156	*RBT4*
RBT4-F1-2A	ACTGTCACAGGTGGTGGC	139 – 156	*RBT4*
RBT4-R1-12A	CCATTACCACCATCAGCAT	278 – 296	*RBT4*
RBT4-F2-1A	GCTGGTGATATTCAACAATC	196 – 215	*RBT4*
RBT4-F2-2A	GCTGATGATATCCAACAATC	196 – 215	*RBT4*
RBT4-R2-12A	CCATAACTGTAATAAACACCG	354 – 374	*RBT4*
RBT4-F3-1A	CAGCTGTTCCAGAAGCTG	248 – 265	*RBT4*
RBT4-F3-2A	CAGTTGTTCCAGAAGCTGA	248 – 266	*RBT4*
RBT4-R3-12A	GGATATTGGTCATCTGAAGG	436 – 455	*RBT4*
RBT4-F4-1A	CCAATTTGCTCAACAAATC	627 – 645	*RBT4*
RBT4-F4-2A	CAAATTTGCTCAACAAATTT	627 – 646	*RBT4*
RBT4-R4-1A	CAAAGTTTCACCATATGTACC	776 – 796	*RBT4*
RBT4-R4-2A	CAAGTTTTCACCATATTTACC	776 – 796	*RBT4*
RBT4-F5-1A	CTACTGGTTACGAATATGCTC	704 – 724	*RBT4*
RBT4-F5-2A	CTACTGTTTACCAATATGCTCA	704 – 725	*RBT4*
RBT4-R56-12A	GATTTCCAGACAACTTGAGT	905 – 924	*RBT4*

RBT4-F6-1A	CGTGATCAATCAAGTTGTC	733 – 751	*RBT4*
RBT4-F6-2A	GCTGATCAATACAGTTGTTCT	733 – 753	*RBT4*
RBT4-F7-1A	CTGCACACTCTGAGTGGTACA	760 – 781	*RBT4*
RBT4-F7-2A	TTGCAACACTCTGGTGGTAA	760 – 780	*RBT4*
RBT4-R78-12A	CCAGTTTTGAGCACGACAA	955 – 973	*RBT4*
RBT4-F8-1A	CATATGGTGAAACTTTGGCT	780 – 799	*RBT4*
RBT4-F8-2A	AATATGGTGAAAACTTGGCT	780 – 799	*RBT4*
RBT4-R4M-2A	GTTTTCACCATATTTACCACC	772 – 793	*RBT4*

[1.] Relative to gene start position = 1. Negative numbers represent primers placed before the gene start position.

Oligonucleotide pairs indicating possible allele specificity (Table 3.3) were taken forward and tested using gradient PCR under a range of annealing temperatures to optimise specificity: 50.1 °C, 50.4 °C, 51.0 °C, 52.2 °C, 53.5 °C, 55.1 °C, 56.6 °C, 58.3 °C, 60.1 °C, 61.2 °C, 61.8 °C and 62.2 °C.

Table 3.3 Oligonucleotides optimised for allele-specificity using gradient PCR

Name	Sequence 5' – 3'	Position[1]	Gene
CDC6-F7-1A	CAAGGTACTGATACGATTAATAAA	1001 – 1025	*CDC6*
CDC6-F7-2A	CAAGGTACCGATACGATTAATAAA	1001 – 1025	*CDC6*
CDC6-R7-1A	TTTTAACAGTCCAATCAAATTA	1244 – 1265	*CDC6*
CDC6-R7-2A	TTTCAACAGTCCAATCAAATCA	1244 – 1265	*CDC6*
RBT4-F4-1A	CCAATTTGCTCAACAAATC	627 – 645	*RBT4*
RBT4-F4-2A	CAAATTTGCTCAACAAATTT	627 – 646	*RBT4*
RBT4-R4-1A	CAAAGTTTCACCATATGTACC	776 – 796	*RBT4*
RBT4-R4-2A	CAAGTTTTCACCATATTTACC	776 – 796	*RBT4*

Efficiency of the allele-specific oligonucleotide combinations were then tested using 2 x SYBR® Green Jumpstart *Taq* Ready Mix (Sigma Aldrich) with 10x fold dilutions of genomic DNA from either the wild-type strain SC5314 or from an appropriate heterozygous knockout strain (see Table 2.1). To ensure that the polymerase was working optimally, genomic DNA extraction using phenol was avoided. Instead DNA was extracted using a Masterpure™ Yeast DNA Purification Kit (Epicentre Biotechnologies) according to the manufacturer's instructions, followed by further purification as follows. The final product from the extraction with the kit was added to 0.1 volumes of 3 M sodium acetate, three volumes of 100% ethanol and vortexed. Samples were frozen at -80 °C for one hour before centrifugation at 4 °C and 12000 rpm for 15 minutes. Pellets were washed in 3 volumes of 70% ethanol and centrifuged at 4 °C and 12000 rpm for 10 minutes. The pellets were air dried at 37 °C for 15 minutes before resuspension in 1 x TE buffer. Concentrations were determined using a NanoDrop spectrophotometer.

Reactions were set up as follows: 25 µl of 2 x SYBR® Green Jumpstart *Taq* Ready Mix (Sigma Aldrich), 1 µl of forward primer, 1 µl of reverse primer, 0.5 µl of reference dye, 0.7 µl of DNA diluted in 1 x TE buffer, with the final volume made up to 50 µl with water. Reactions without DNA were used as negative controls. Technical triplicates were prepared.

Reactions were carried out using a Stratagene Mx3005P set to detect fluorescence from SYBR® Green with a ROX reference dye. The cycle was set as follows:

94 °C for 2 minutes
94 °C for 15 seconds } 40 cycles
* °C for 1 minute }

* Appropriate annealing temperature as determined by previous gradient PCR.

To calculate the efficiency of the primer sets, the average ΔCt values of the replicates were plotted against the concentration of DNA standards from SC5314. Efficiency was calculated using the gradient of this line (Δx/Δy) and the following calculation taken from Applied Biosystems:

$$\text{Efficiency} = \left(\left(10^{-\frac{1}{\text{gradient}}} \right) - 1 \right) \times 100$$

3.2.6.2 Restriction Enzyme Verification

Verification of the RNA sequencing results was attempted with an allele-specific restriction enzyme digest (for the gene *VPS1*) as has been previously used for verification of AEI in maize (Zhang *et al.*, 2011). As the genes of interest were present in such low abundance in cDNA samples, a PCR step was introduced prior to digestion to amplify the signal. This was carried out as described in section 2.7 using a selection of appropriate oligonucleotides from Table 2.7. PCR products were checked using DNA gel electrophoresis (see section 2.5) and gel extracted as described in section 2.6.

PCR products amplified from genomic DNA from appropriate heterozygous knockout strains (see Table 2.1) and from the wild-type strain SC5314 were used as controls. Heterozygous strains were constructed using transformation

and homologous recombination of a cassette as detailed in sections 2.8 – 2.11. The wild-type genomic DNA should produce bands of the same intensity for both alleles, whereas the cDNA will have bands of varying intensity depending on the ratio of expression between allele 1 and allele 2.

Restriction digest reactions were prepared as follows: 8 µl of the PCR product or genomic DNA (from the wild-type strain SC5314 or from an appropriate heterozygous knockout as described in Table 2.1) was combined with 1 µl of restriction enzyme (see Table 3.4), 2 µl of appropriate buffer and 9 µl of sterile water. The reaction was incubated at 37 °C for 3 and a half hours and the entire reaction was separated using DNA gel electrophoresis (see section 2.5).

Table 3.4 Restriction enzymes used for allele-specific PCR fragment digestion

Restriction Enzyme[1]	**Buffer**	**Gene**	**Fragment Size (bp)**	
HaeIII	4	*VPS1*	Allele 1	314
				123
				172
			Allele 2	26
				288
				295

1. All restriction enzymes were purchased from New England BioLabs UK.

3.2.6.3 Western Blotting

To monitor the protein expression of individual alleles, strains were constructed with each individual allele tagged with a V5-6xHis marker (Milne *et al.*, 2011). Strain construction was achieved using transformation and homologous recombination of a cassette as detailed in sections 2.8 – 2.11. All strains used are listed in Table 2.1. Expression of each allele was detected using an Anti-V5 antibody (Invitrogen) and normalised against expression of the actin protein detected using an Anti-Actin antibody (Thermoscientific).

3.2.6.3.1 Soluble Protein Extract

To extract total soluble proteins, cells were grown to an optical density at 600 nm of 1, as described in section 3.2.6.1.1, to make conditions comparable to the

RNA sequencing. Unless otherwise stated, all further steps were carried out on ice and centrifugation was carried out at 4 °C, Cells were then pelleted at 4000 rpm for five minutes, washed in 10 ml of ice-cold sterile water and resuspended in 1 ml of ice-cold breaking buffer (100 mM Tris pH 7.5, 0.01% SDS, 1 mM DTT, 10% (v/v) glycerol, 1 M EDTA, 100 µl protease inhibitor cocktail (Melford Laboratories)). Cells were then transferred to screw top microfuge tubes, pelleted at 13000 rpm for five minutes and resuspended in 250 µl of ice-cold breaking buffer. Acid-washed 0.4 – 0.6 mm glass beads were added to the meniscus layer and cells were broken open using a FASTPREP-24 bead beater (MP) four times at 6.5 m/s for 20 seconds with one minute intervals on ice. Cell debris and glass beads were pelleted at 13000 rpm for 10 minutes. The supernatant containing the protein was transferred to a fresh tube, clarified by centrifugation at 13000 rpm for 10 minutes and stored at -20 °C. Protein concentrations were determined using a Bradford Assay (Sigma Aldrich) following the manufacturer's instructions.

3.2.6.3.2 SDS-PAGE and Protein Transfer

Approximately 50 µg of each protein sample (section 3.2.6.3.1) were combined with 1x protein sample buffer and 25 mM DTT before incubation at 70 °C for 10 minutes. Samples were run on NuPAGE gels (4 – 12% gradient) in running buffer (1.21% (w/v) Tris, 2.38 % (w/v) HEPES, 0.1% (w/v) SDS) at 110 volts for approximately one hour. PVDF membrane (Invitrogen) was activated in methanol for 10 minutes. The proteins were then transferred to the activated membrane in transfer buffer (5 % (v/v) Transfer solution (Invitrogen), 10% methanol) at 30 volts for 90 minutes using an X Cell II™ Blot Module (Invitrogen). The membrane was then washed in TBS solution (10 mM Tris pH 8.0, 137 mM sodium chloride). To ensure sufficient transfer of proteins, the membrane was stained with Ponceau S (0.1% Ponceau S, 5% acetic acid) for five minutes. Excess stain was rinsed away with deionised water and the membrane was photographed. The stain was then removed using 100 mM sodium hydroxide and the membrane was washed thoroughly in TBS solution.

3.2.6.3.3 Detection of Protein Expression

Membranes were blocked in 10 ml TBS-T + 5% BSA (TBS solution (10 mM Tris pH 8.0, 137 mM sodium chloride) + 0.1% Tween-20 + 5% BSA) overnight at 4

°C with gentle agitation. The primary antibody (Anti-V5 or Anti-Actin) was appropriately diluted in TBS-T + 5% BSA. Anti-V5 (Invitrogen) was diluted in a 1:5000 ratio and Anti-Actin (Thermoscientific) was diluted in a 1:1000 ratio. The membrane was incubated in this solution for one hour with gentle agitation and then washed three times in 10 ml TBS-T for five minutes. The appropriate secondary antibody (Anti-mouse linked to horseradish peroxidase (HRP), Molecular probes) was diluted in a 1:10000 ratio in TBS-T + 5% BSA and the membrane was incubated in this solution for one hour with gentle agitation. Again, the membrane was washed three times in 10 ml TBS-T for five minutes. The membrane was then developed using an ECL Plus kit (Pierce Antibodies) following the manufacturer's instructions and imaged using a G:Box system (SynGene) set with a chemiluminescence filter for detection of ECL substrates.

For detection with multiple primary antibodies the blot was stripped after the first detection. The blot was washed in 20 ml pre-warmed western stripping buffer (2% SDS, 100 mM β-mercaptoethanol, 50 mM Tris pH 6.8) for 30 minutes at 65 °C in a hybridisation oven (HL-2000 Hybrilinker, UVP). This was followed by washing five times in 10 ml TBS-T for five minutes with gentle agitation. The protocol was then repeated with the new antibody from the point of adding the primary antibody.

3.3 Results

3.3.1 Identification of Genes with AEI and the Structural Trends Associated with These Genes

3.3.1.1 Genes with Allelic Expression Imbalance

RNA sequencing of the *C. albicans* wild-type strain SC5314, grown in standard laboratory conditions in triplicate (YPD at 30 °C, OD at 600 nm = 1), yielded 16,166,757 reads for replicate 1, 17,369,675 for replicate 2 and 16,853,362 for replicate 3. 26,125,364 sequence tags mapped to 5,807 ORFs, producing a median count of 320 tags per ORF (average per base coverage, 117x). To identify genes with allelic expression imbalance, only reads which aligned uniquely to the diploid reference genome (Jones *et al.*, 2004) were used. Reads were normalised to produce a RPKM value (Mortazavi *et al.*, 2008) for each allele and statistically compared using a Fisher's Exact Test with a cut off of 2x fold difference in expression at a p-value of <0.0000077 (set using Bonferroni

correction). These criteria identified a total of 152 genes with significant allelic expression imbalance and a further 81 genes with monoallelic expression (Appendix I Table Ia and Ib).

3.3.1.2 Gene Ontology (GO) Analysis

Gene Ontology (GO) term analysis was carried out, using the GO Term Finder tool available from the *Candida* Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012), to identify any functional patterns present within genes with allelic expression imbalance. No significant over representation was observed in process, component or function. This suggests that AEI is present in a random selection of genes and that there is no strong selection for AEI to operate in any concerted manner in specific biological processes. Therefore reasons behind the phenomenon of AEI need to be considered on a gene-by-gene basis.

Identification of the GO terms associated with these 233 genes, using the GO Slim Mapper tool available from the *Candida* Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012), shows that some genes with AEI are however associated with virulence attributes (Table 3.5).

Table 3.5 GO Process Terms found in the set of genes with allelic expression imbalance. Identified using "CGD Gene Ontology Slim Mapper" (www.candidagenome.org)

GOID	GO term	Frequency	Background Frequency	Gene(s)
50789	"regulation of biological process"	49 out of 233 genes, 21.0%	1328 out of 6712 genes, 19.8%	*ALS1 ARP4 BCY1 CAR1 CDC6 CST5 ECM25 HGT1 IFH1 LTP1 MDN1 MSS4 NUP84 POR1 RCK2 RFG1 RPN4 RPS23A SFL2 SMI1 SNF4 SRR1 SSK2 SSY1 TAC1 TBF1 VID21 VMA7 ZCF6* orf19.1185 orf19.1196 orf19.1212 orf19.1643 orf19.1694 orf19.2309 orf19.232 orf19.2458 orf19.2743 orf19.3792 orf19.3920 orf19.3954 orf19.4295 orf19.4488 orf19.4728 orf19.48 orf19.5221 orf19.6080 orf19.643 orf19.748
8150	"biological process"	48 out of 233 genes, 20.6%	1699 out of 6712 genes, 25.3%	*BMT6 EMC9 FAV3 HIT1 IFF9 PGA45 PGA57 PHM7 WSC1* orf19.1152 orf19.1219 orf19.1246 orf19.1266 orf19.1383 orf19.1440 orf19.1637 orf19.1725 orf19.1736 orf19.1948 orf19.1953 orf19.2051 orf19.2381 orf19.246 orf19.2521 orf19.254 orf19.2724 orf19.2731 orf19.2742 orf19.310 orf19.3353 orf19.3448 orf19.3607 orf19.3644 orf19.4068 orf19.4332 orf19.4349 orf19.4398 orf19.4470 orf19.4749 orf19.4880 orf19.4952 orf19.4959 orf19.5103 orf19.5626 orf19.5648 orf19.6235 orf19.6351 orf19.6556
6810	"transport"	41 out of 233 genes, 17.6%	1001 out of 6712 genes, 14.9%	*ATP1 CHS6 DAL4 ECM1 ECM21 EXO84 FCY21 HGT1 IFC1 ITR1 MSN5 MTR10 NUP84 PIR1 PLD1 POR1 SAC3 SEO1 SNX4 SSY1 SUL2 TFP1 TPO3 VMA11 VMA7 VPS1 YDJ1* orf19.1356 orf19.1386 orf19.1536 orf19.2002 orf19.3556 orf19.4184 orf19.4337 orf19.4466 orf19.5095 orf19.5534 orf19.6020 orf19.6346 orf19.6555 orf19.748
6996	"organelle organization"	35 out of 233 genes, 15.0%	913 out of 6712 genes, 13.6%	*ARP4 ATS1 BBC1 CDL1 IFM1 MDN1 MSS4 NUP84 POR1 SAC3 SNF4 SSK2 TBF1 VID21 VPS1 YDJ1* orf19.1185 orf19.1212 orf19.1386 orf19.1646 orf19.2002 orf19.2309 orf19.2743 orf19.3161 orf19.4184 orf19.4295 orf19.4488 orf19.4728 orf19.48 orf19.5221 orf19.5534 orf19.6020 orf19.643 orf19.6555 orf19.748

16070	“RNA metabolic process”	30 out of 233 genes, 12.9%	723 out of 6712 genes, 10.8%	*ATS1 CDL1 EDC3 ERB1 EXO84 IFH1 MDN1 MPP10 POL5 PUS4 PWP1 RPS23A SAC3 SES1 SGD1 TAC1 TBF1 TIF4631* orf19.1356 orf19.1646 orf19.1938 orf19.2309 orf19.3103 orf19.3161 orf19.3792 orf19.4295 orf19.48 orf19.494 orf19.518 orf19.581
6950	“response to stress”	26 out of 233 genes, 11.2%	794 out of 6712 genes, 11.8%	*ARP4 BCY1 CHT2 ECM25 EDC3 HGT1 HSP12 NUP84 OCA1 PLD1 RCK2 RFG1 RPN4 SAC3 SGD1 SNF4 SRR1 SSK2 TAC1 VID21 VMA7* orf19.2458 orf19.3954 orf19.5221 orf19.6020 orf19.748
30447	“filamentous growth”	22 out of 233 genes, 9.4%	555 out of 6712 genes, 8.3%	*ALS1 BCY1 CHT2 CST5 ECM25 EDC3 HGT1 LMO1 PLD1 RCK2 RFG1 SFL2 SNF4 SRR1 SSY1 TAC1 VMA7 VPS1* orf19.1536 orf19.3524 orf19.3954 orf19.583
6464	“cellular protein modification process”	19 out of 233 genes, 8.2%	535 out of 6712 genes, 8%	*ARP4 LTP1 OCA1 PPT1 RCK2 SNF4 SSK2 VID21* orf19.1092 orf19.1185 orf19.1196 orf19.1557 orf19.261 orf19.2743 orf19.3524 orf19.3996 orf19.4466 orf19.4728 orf19.6020
42221	“response to chemical stimulus”	18 out of 233 genes, 7.7%	637 out of 6712 genes, 9.5%	*CST5 ERB1 FCY21 HGT1 OCA1 PLD1 RBT4 RCK2 RPN4 SMI1 SRR1 SSK2 SSY1 TAC1 TPO3* orf19.3954 orf19.4488 orf19.583
42254	“ribosome biogenesis”	17 out of 233 genes, 7.3%	281 out of 6712 genes, 4.2%	*CDL1 ECM1 ERB1 MDN1 MPP10 PWP1 RPS23A RPS7A SAC3 SGD1 TIF4631* orf19.1646 orf19.2002 orf19.3161 orf19.3797 orf19.494 orf19.6346
7010	“cytoskeleton organization”	11 out of 233 genes, 4.7%	180 out of 6712 genes, 2.7%	*ARP4 ATS1 BBC1 MSS4 SSK2 VPS1 YDJ1* orf19.1185 orf19.5221 orf19.5534 orf19.643
16192	“vesicle-mediated transport”	11 out of 233 genes, 4.7%	308 out of 6712 genes, 4.6%	*CHS6 ECM21 EXO84 PLD1 SNX4 TFP1 VPS1* orf19.1386 orf19.4184 orf19.5095 orf19.5534
7049	“cell cycle”	9 out of 233 genes, 3.9%	407 out of 6712 genes, 6.1%	*ARP4 CDC6 SAC3* orf19.1185 orf19.3411 orf19.3556 orf19.5534 orf19.6240 orf19.643

6629	“lipid metabolic process”	9 out of 233 genes, 3.9%	252 out of 6712 genes, 3.8%	*MSS4 PDX3 PLD1* orf19.1092 orf19.1212 orf19.273 orf19.3954 orf19.3996 orf19.4122
6412	“translation”	9 out of 233 genes, 3.9%	320 out of 6712 genes, 4.8%	*IFM1 RPL20B RPL24A RPS23A RPS7A SES1* orf19.3792 orf19.4751 orf19.48
9405	“pathogenesis”	9 out of 233 genes, 3.9%	215 out of 6712 genes, 3.2%	*ALS1 PLD1 RBT4 RCK2 RFG1 SFL2 SRR1 VMA7* orf19.3524
7165	“signal transduction”	8 out of 233 genes, 3.4%	189 out of 6712 genes, 2.8%	*BCY1 CST5 ECM25 MSS4 RCK2 SRR1 SSK2* orf19.1196
6259	“DNA metabolic process”	8 out of 233 genes, 3.4%	371 out of 6712 genes, 5.5%	*ARP4 CDC6 NUP84 SAC3 TBF1 VID21* orf19.4295 orf19.748
5975	“carbohydrate metabolic process”	7 out of 233 genes, 3.0%	251 out of 6712 genes, 3.7%	*CHS6 CHT2 SMI1* orf19.1092 orf19.261 orf19.3996 orf19.4488
71555	“cell wall organization”	7 out of 233 genes, 3.0%	168 out of 6712 genes, 2.5%	*ECM1 ECM21 ECM25 LMO1 PIR1 RCK2* orf19.5221
42493	“response to drug”	7 out of 233 genes, 3.0%	357 out of 6712 genes, 5.3%	*ERB1 FCY21 HGT1 RBT4 SMI1 TAC1 TPO3*
42710	“biofilm formation”	5 out of 233 genes, 2.1%	125 out of 6712 genes, 1.9%	*ALS1 CST5 IFD6 SMI1 VPS1*
30163	“protein catabolic process”	5 out of 233 genes, 2.1%	196 out of 6712 genes, 2.9%	*APE3 RPN4 RPN6 YDJ1* orf19.6630

19725	“cellular homeostasis”	5 out of 233 genes, 2.1%	152 out of 6712 genes, 2.3%	*POR1 TFP1 VMA7* orf19.1536 orf19.3920
746	“conjugation”	4 out of 233 genes, 1.7%	92 out of 6712 genes, 1.4%	*CST5 ITR1 PLD1 SSY1*
16044	“cellular membrane organization”	4 out of 233 genes, 1.7%	179 out of 6712 genes, 2.7%	*SNX4* orf19.1386 orf19.4184 orf19.6020
48468	“cell development”	4 out of 233 genes, 1.7%	108 out of 6712 genes, 1.6%	*ITR1 MSS4 PLD1* orf19.4466
910	“cytokinesis”	3 out of 233 genes, 1.3%	120 out of 6712 genes, 1.8%	*CYK3* orf19.3411 orf19.6240
45333	“cellular respiration”	3 out of 233 genes, 1.3%	98 out of 6712 genes, 1.5%	*RIB3* orf19.2309 orf19.4468
6457	“protein folding”	3 out of 233 genes, 1.3%	98 out of 6712 genes, 1.5%	*YDJ1* orf19.2828 orf19.3920
6091	“generation of precursor metabolites and energy”	3 out of 233 genes, 1.3%	138 out of 6712 genes, 2.1%	*RIB3* orf19.2309 orf19.4468
6997	“nucleus organization”	3 out of 233 genes, 1.3%	46 out of 6712 genes, 0.7%	*NUP84* orf19.2002 orf19.748
6766	“vitamin metabolic process”	3 out of 233 genes, 1.3%	37 out of 6712 genes, 0.6%	*PDX3 RIB3* orf19.3411

30448	“hyphal growth”	2 out of 233 genes, 0.9%	60 out of 6712 genes, 0.9%	*ALS1* orf19.3524
44419	“interspecies interaction between organisms”	2 out of 233 genes, 0.9%	126 out of 6712 genes, 1.9%	*ALS1 SMI1*
7155	“cell adhesion”	1 out of 233 genes, 0.4%	49 out of 6712 genes, 0.7%	*ALS1*
7114	“cell budding”	1 out of 233 genes, 0.4%	36 out of 6712 genes, 0.5%	*ATS1*
70783	“growth of unicellular organism as a thread of attached cells”	1 out of 233 genes, 0.4%	76 out of 6712 genes, 1.1%	*CST5*
7124	“pseudohyphal growth”	0 out of 233 genes, 0.0%	39 out of 6712 genes, 0.6%	
32196	“transposition”	0 out of 233 genes, 0.0%	4 out of 6712 genes, 0.1%	

Seven genes (*ERB1, FCY21, HGT1, RBT4, SMI1, TAC1* and *TPO3*) are associated with the "response to antifungal drugs". This includes *TAC1,* the transcription factor which activates the drug resistance genes *CDR1* and *CDR2.* This gene has already been identified as having functionally distinct 'normal' and 'hyperactive' alleles which can be found together in a heterozygous strain (Coste *et al.*, 2006).

There are also nine genes under the category of "pathogenesis" (*ALS1, PLD1, RBT4, RCK2, RFG1, SFL2, SRR1, VMA7* and orf19.3524). *ALS1* (an adhesin-like mannoprotein) is involved in adhesion to host epithelial cells. Other members of this gene family, for example *ALS9,* have also been found to have functionally distinct alleles (Zhao *et al.*, 2003, Zhao *et al.*, 2007). *RFG1* is a transcriptional regulator of filamentous growth (Kadosh and Johnson, 2001). Defects in virulence in mice are seen in both the heterozygous and homozygous knockout strains. Interestingly, this paper shows differences in virulence phenotypes between a heterozygous knockout and a heterozygous reintegrant strain, with suggestions of difference in allele expression or activity (Kadosh and Johnson, 2001).

22 genes are associated with "filamentous growth" (*ALS1, BCY1, CHT2, CST5, ECM25, EDC3, HGT1, LMO1, PLD1, RCK2, RFG1, SFL2, SNF4, SRR1, SSY1, TAC1, VMA7, VPS1,* orf19.1536, orf19.3524, orf19.3954 and orf19.583). The morphological switch between yeast and hyphal form has been associated with the virulence of *C. albicans* strains (Biswas *et al.*, 2007). The gene *ECM25,* which has been highlighted as a gene with AEI, has been shown to be important during this transition. Homozygous knockout strains of *ECM25* have reduced virulence in mice associated with defects in filamentous growth, cell morphology and cell growth (Zhang *et al.*, 2008). Other genes which fall under this category of filamentous growth associated include *TAC1, ALS1* and the protein kinase A regulatory subunit *BCY1.*

Five genes that display AEI are associated with "biofilm formation" (*ALS1, CST5, IFD6, SMI1, and VPS1*). Biofilms are a clinically relevant phenotype, which form on medically implanted devices such as catheters and dentures, and often show increased resistance to antifungals (Ramage *et al.*, 2004). Alongside

ALS1, inducible homozygous knockouts of the vacuolar sorting protein *VPS1* have been identified as defective in biofilm formation and filamentous growth (Bernardo *et al.*, 2008).

A further seven genes with AEI that exhibit with “cell wall organisation” (*ECM1, ECM21, ECM25, LMO1, PIR1, RCK2* and orf19.5221) including the aforementioned gene *ECM25.* The cell wall is an integral structure with involvement in adhesion to host cells, drug resistance and response to stress. Additionally, *PIR1* has a role in cell wall assembly. This gene produces two distinct alleles that differ in length. Heterozygous knockouts of either allele have shown hypersensitivity to various drugs (Martínez *et al.*, 2004).

No significant differences are observed in the percentages of genes within GO terms between the 152 genes with AEI and the 81 genes with monoallelic expression (Figure 3.2). But it should be noted that there are a much larger proportion of genes associated with “transport” and “protein catabolic” processes in monoallelic genes (24.7% monoallelic vs. 13.8% AEI). Additionally there are no monoallelic genes involved with “cytokinesis”, “cell budding”, “cell adhesion” and “hyphal growth” suggesting that *C. albicans* may require both alleles of a gene to be expressed to maintain these vital processes.

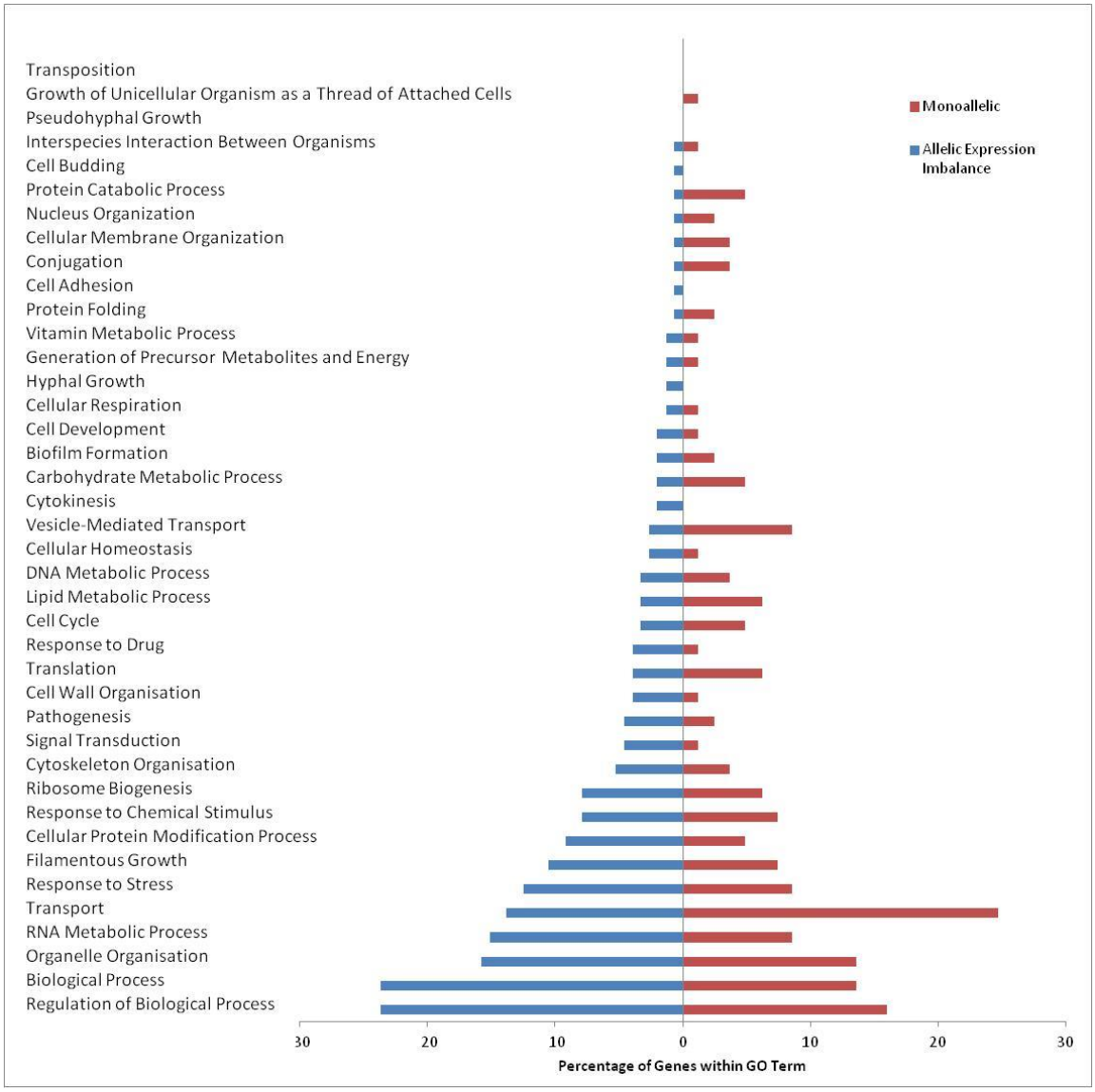


Figure 3.2 Percentage of genes with AEI (blue) and monoallelic expression (red) within each Gene Ontology term. As defined by CGD Gene Ontology Slim Mapper www.candidagenome.org.

3.3.1.3 Differences in Promoter Sequences of Genes with AEI

As promoter sequences are undefined in *C. albicans,* the 1000 bp upstream (or the region up to the neighbouring open reading frame) of an allele were compared for a gene as described in section 3.2.3 to assess if differences in promoter sequences are the driving force behind AEI. Of the 233 genes identified with AEI, the upstream DNA sequences were available for 190 genes. 75% of these genes with AEI were found to have a difference of at least one base pair in this region. However, when analysis was carried out on an equivalent data-set of genes with equal allele expression (fold difference ≈ 1) (Appendix I Table II), 78% of genes were also found to have differences in the upstream region. Therefore it cannot be concluded that differences in promoter sequences are leading to differences in allele expression levels. Based upon the observed level of heterozygosity being one SNP in every 237 bp (Jones *et al.*, 2004), the probability that 1000 bp region is homozygous can be calculated as follows:

$$\frac{236}{237}^{1000} = 0.015$$

Therefore, based on this probability, it would be expected that just 3 promoter regions (1.5%) out of the 190 would be homozygous. However, a significantly larger proportion of promoter sequences were found to be homozygous at 25% (chi-square test, d.f. = 1, $p = 4.91 \times 10^{-165}$). From this it can be inferred that differences in promoter regions are not leading to differences in allele expression, and in fact mutations within these regions, of both genes with AEI and genes with equal allele expression, are selected against, possibly to ensure that gene expression levels are not altered. It should be noted that this probability has been calculated for the simplest model where 1000 bp upstream were compared for each in gene. In reality, upstream regions varied in length dependent upon the distance to the adjacent open reading frame which may impact upon this statistical measure.

3.3.1.4 Genes with AEI Show Significantly Lower Percentage Protein Identity

To examine the hypothesis that genes with allelic expression imbalance have alleles with differing functions, the difference in the amino acid sequences of alleles with AEI was calculated with the assumption that a larger difference in protein sequence could indicate a functional difference. Percentage protein identity comparisons were carried out using the Needleman-Wunsch algorithm for global alignment (Rose and Eisenmenger, 1991). Each gene was assigned a percentage indicating the similarity of proteins produced by each allele. The genes with AEI were found to have significantly lower percentage protein identities when compared to a cohort of genes with equal allele expression (fold difference ≈ 1) (Appendix I Table II) (two-sample t-test, $p < 0.001$; Figure 3.3). A total of 34 genes have a protein identity of less than 50% (Table 3.6). The majority of these genes have unknown functions, however the list includes the transcription factor *TAC1* which has previously been identified as having functionally distinct alleles (Coste *et al.*, 2006). This provides support for the idea that functional differences in protein sequences could be the driving force behind allelic expression imbalance.

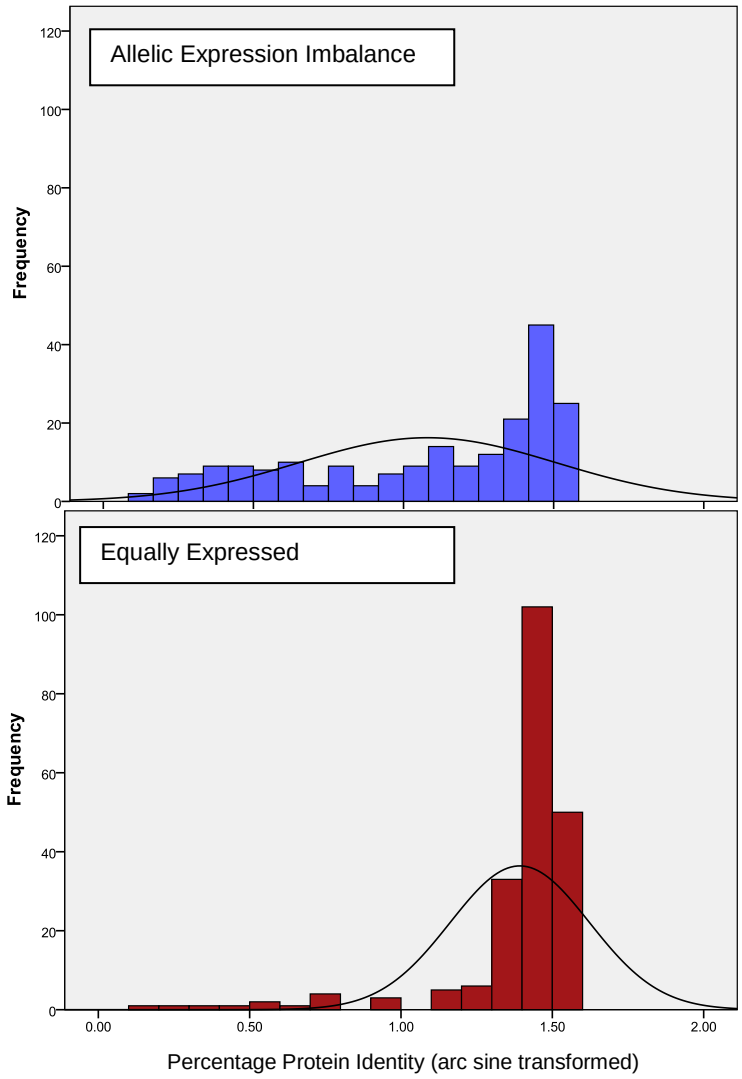


Figure 3.3. The distribution of percentage protein identities after arc sine transformation of genes with AEI (blue) and equally expressed alleles (red). The percentage protein identities of genes with AEI are significantly lower than that of equally expressed alleles (two sample t-test, $p < 0.001$). Black line shows fitted distribution assuming normal distribution.

Table 3.6 Genes with AEI and less than 50% protein identity between alleles

Allele with Lowest Expression	Allele with Highest Expression	Gene Name	Percentage Protein Identity (2 d.p.)
orf19.2053	orf19.9599		11.05
orf19.8215	orf19.583		15.13
orf19.3188	orf19.10698	*TAC1*	18.55
orf19.1694	orf19.9261		18.62
orf19.4466	orf19.11946	*ERP1*	19.41
orf19.4697	orf19.12167	*MDN1*	23.52
orf19.4880	orf19.12344	*YFW5*	23.90
orf19.10309	orf19.2791	*BBC1*	24.40
orf19.6894	orf19.14182		27.09
orf19.2665	orf19.10182	*MSN5*	29.36
orf19.5624	orf19.13069		29.84
orf19.4770	orf19.12233		30.10
orf19.2127	orf19.9674	*CST5*	30.21
orf19.11605	orf19.4122		30.89
orf19.1383	orf19.8963		31.32
orf19.9015	orf19.1440		33.33
orf19.9130	orf19.1557		35.22
orf19.4332	orf19.11806		35.64
orf19.10590	orf19.3077	*VID21*	36.82
orf19.5150	orf19.12615		37.34
orf19.4887	orf19.12352	*ECM21*	37.54
orf19.5615	orf19.13060	*AYR2*	38.25
orf19.6389	orf19.13747		38.44
orf19.10346	orf19.2828		38.52
orf19.11191	orf19.3706		43.07
orf19.8671	orf19.1069	*RPN4*	43.20
orf19.5145	orf19.12610	*SSP96*	43.26
orf19.7862	orf19.232		43.27
orf19.1230	orf19.8815		43.58
orf19.13020	orf19.5574		43.86
orf19.1151	orf19.8744		45.11
orf19.11826	orf19.4349		46.60
orf19.11082	orf19.3599	*TIF4631*	47.27
orf19.3524	orf19.11006		48.46

3.3.2 The Contribution of Structural Factors to AEI

The impact of various structural factors upon AEI was investigated to determine any potential associations. These structural factors include chromosomal location, GC content, gene length and codon usage, and are influenced both by gene sequence and gene location. From the set of 233 genes with allelic expression imbalance, 22 were removed due to lack of information or deletion from the more recent *C. albicans* genome Assembly 21 (Van Het Hoog *et al.*, 2007) leaving 210 genes for analysis.

3.3.2.1 Chromosomal Location

An initial question to be addressed when looking at the consequence of structural factors upon allelic expression imbalance is gene location. Are there any significant patterns seen in the location or clustering of genes with AEI? The distribution of genes with AEI across chromosomes was analysed to see if there are any over-representations on certain chromosomes (Figure 3.4). The overall distribution of genes with allelic expression imbalance on each chromosome differs significantly from the expected percentage based upon that seen throughout the entire genome (Chi-square test, 7 d.f., p = 0.013; Table 3.7). This significant difference is likely to be due to Chromosome R which has significantly fewer genes with AEI than expected (Chi-square test, 1 d.f., p = 0.043; Table 3.7). Although Chromosome 7 actually has a smaller percentage of genes with AEI than Chromosome R, this did not result in a statistically significant difference to the distribution across the entire genome due to the smaller overall number of genes on Chromosome 7. However, the p value was very close to being significantly different (Chi-square test, 1 d.f., p = 0.073; Table 3.7).

Table 3.7 Percentage of features across the entire *C. albicans* genome and within each chromosome that have AEI

Chromosome	Genome	1	2	3	4	5	6	7	R
Percentage of features with AEI	3.10	3.16	3.33	2.25	4.17	4.36	4.56	1.61	2.04[1]

1. Chromosome R (highlighted red) has significantly less genes with AEI than expected (Chi-squared, 1 d.f., p = 0.043).

In Figure 3.4, clusters which are statistically unlikely to occur by chance are highlighted with arrows (Poisson distribution, $p < 0.05$; Figure 3.4). No patterns are obvious, with only a few clusters appearing, all of which are at different places across the chromosomes with no association to either centromeric or telomeric regions. However on chromosome 3, genes with AEI are restricted to just one small section. To see if this is significant clustering of genes with allelic expression imbalance, the location of all polymorphic genes on each chromosome was calculated and overlaid with the position of genes with AEI (Figure 3.5). The distribution of polymorphic genes are similar to those recorded during the sequencing of the diploid genome (Jones *et al.*, 2004). In accordance with previous observations chromosome 3 is highly homozygous with significantly less polymorphic genes than expected when compared to the entire genome (Chi-square test, 1 d.f., $p < 0.001$; Table 3.8). These polymorphic genes are at the same locations as the genes with AEI suggesting that there is no significant clustering and that allelic expression imbalance is not influenced by the location of the gene.

During determination of the location of all polymorphic genes, it was noted that chromosomes 1, 7 and R also have significantly less polymorphic genes than expected when compared to distribution across the entire genome (Chi-square test, 1 d.f. $p = 0.002$, $p < 0.001$ and $p < 0.001$ respectively; Table 3.8), whilst chromosome 6 has significantly more polymorphic genes than expected (Chi-square test, 1 d.f., $p < 0.001$; Table 3.8). This may be a possible explanation as to why chromosome R has a significantly smaller percentage of genes with AEI. Clusters of polymorphic genes that are unlikely to happen by chance are highlighted in Figure 3.5 (Poisson distribution, $p < 0.05$). Surprisingly, as well as chromosomes 3, 7 and R having less polymorphisms than expected, the majority of these polymorphic genes are on one side of the centromere. This suggests that one arm is almost entirely homozygous. The advantage of this to *C. albicans* is not clear and needs further investigation. These findings support previous observations by (Odds *et al.*, 2004) that heterozygosity within the *C. albicans* genome is unevenly distributed. Chromosome 5 and 6 were stated as highly homozygous, which is echoed within these results (Odds *et al.*, 2004).

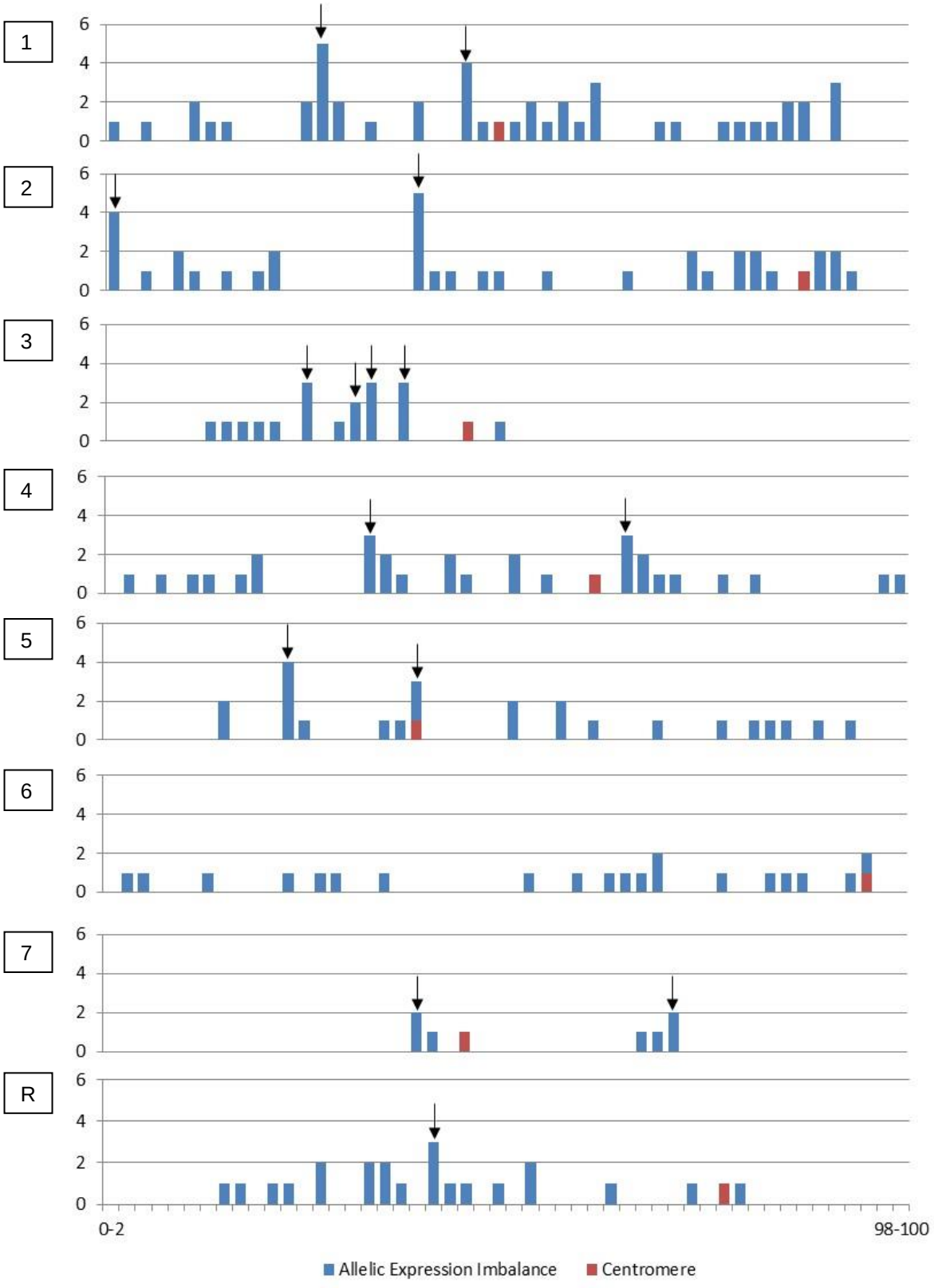


Figure 3.4 Distribution of genes with AEI across each chromosome. Genes with AEI are blue. Centromeres are red. The x-axis shows start coordinates as a percentage of the total length of the chromosome. The y-axis shows the number of differentially expressed polymorphic alleles. Clusters which are unlikely to happen by chance are highlighted with an arrow (Poisson distribution, $p < 0.05$).

Figure 3.5 Distribution of polymorphic genes (green) and genes with AEI (blue) across each chromosome. Centromeres are shown in red. The x-axis shows start coordinates as a percentage of the total length of the chromosome. The y-axis shows the number of polymorphic alleles. Clusters which are unlikely to happen by chance are highlighted with an arrow (Poisson distribution, $p < 0.05$).

Table 3.8 Percentage of polymorphic genes present across the entire *C. albicans* genome and on each chromosome

Chromosome	Genome	1	2	3	4	5	6	7	R
Percentage of polymorphic genes	25.65	22.17[1]	23.87	11.89[1]	25.69	27.09	40.13[2]	7.59[1]	16.00[1]

1. Highlighted blue as have fewer polymorphic genes than expected (Chi-square test, 1 d.f., Chromosome 1 - p = 0.002, Chromosomes 3, 7, R – p < 0.001).
2. Highlighted red as have more polymorphic genes than expected (Chi-square test, 1 d.f., Chromosome 6 – p < 0.001).

3.3.2.2 Overlapping Genes

Transcript expression levels have been shown to be altered when the coding sequences of genes overlap, with results generally showing the occurrence of transcriptional repression (Nagalakshmi *et al.*, 2008, Gagneur *et al.*, 2009). From this it is possible to hypothesize that the patterns of AEI that we have identified in *C. albicans* could be due to uneven overlapping of alleles; where one allele overlaps, and therefore undergoes transcriptional repression, whereas the other allele does not overlap, and therefore expression is unaffected. This is seen in mice for the imprinted *Mest* locus (MacIsaac *et al.*, 2011) as described in section 3.1.2.2.

To address this question, firstly the numbers of overlapping genes across each chromosome were identified and compared to the numbers of genes with AEI across each chromosome as discussed in section 3.3.2.1. In total, 5.89% of all features were found to overlap with their neighbouring open reading frame. Of these overlapping features, 20.15% were not ORFs. This figure is similar to that found in the human genome, where 1316 pairs of overlapping genes were identified, which is approximately 7% of all genes (Veeramachaneni *et al.*, 2004). The frequency of overlapping features across each chromosome appears to be random, as seen in Table 3.8. Statistical analysis shows that chromosome 1 has significantly less overlapping features than expected based on overall distribution across the genome (Chi-square test, 1 d.f., p < 0.001;

Table 3.9). Whereas chromosomes 3 and 7 and the mitochondrial genome have significantly more overlapping features than expected (Chi-square test, 1 d.f., p = 0.001, p = 0.02, p < 0.001 respectively; Table 3.9). Importantly, there are no similarities in the frequency of overlapping features and the frequency of genes with allelic expression imbalance on each chromosome (see Table 3.7) suggesting that overlapping is unlikely to be causally linked to allelic expression imbalance.

Table 3.9 Percentage of features in the entire *C. albicans* genome and in each chromosome which overlap

Chromosome	Genome	1	2	3	4	5	6	7	R	M
Percentage of Features that Overlap	5.89	3.71[1]	5.74	8.64[1]	6.39	5.45	4.99	8.74[1]	5.55	25.00[1]

1. Frequencies that differ significantly from expected are highlighted in red (chi-square test, 1 d.f., p < 0.001, p = 0.001, p = 0.02 and p < 0.001 respectively).

To further support the lack of a relationship between overlap and allelic expression imbalance, the genes with differentially expressed alleles that overlap with neighbouring features were identified. Only four genes (orf19.2127, orf19.246, orf19.3607 and orf19.4332) were found to satisfy both of these criteria out of a total of 391 overlapping genes (Table 3.10), which was deemed as a statistically significant low amount (hypergeometric distribution, p = 0.012). Of these four genes, two have both alleles overlapping with the adjacent ORF, suggesting that transcriptional interference due to overlap of only one allele is not the cause of the allelic expression imbalance. Not only does this result firmly reject the hypothesis that allelic expression imbalance is due to overlap of just one allele, it also produces a further question – why do such an unexpectedly small number of genes with AEI overlap with neighbouring ORFs? Is there a disadvantage to overlapping? If open reading frames do overlap, is the expression level of one gene or both genes affected as seen in other organisms (Nagalakshmi *et al.*, 2008, Gagneur *et al.*, 2009)?

Table 3.10 Percentage of genes with AEI and without AEI that do and do not overlap with the neighbouring feature.

	Overlapping Feature	**Non-Overlapping Feature**
Allelic Expression Imbalance	0.06 %	3.45 %
No Allelic Expression Imbalance	5.83 %	90.66 %

To investigate this, the RPKM values obtained via RNA sequencing for a pair of overlapping genes were compared, via a Fisher Exact test, to assess the effect of overlapping on gene expression. A subset of pairs were found to have significantly different RPKM values, and therefore expression levels (Fisher exact test, $p < 0.000568$; Table 3.11). The remaining pairs were deemed to have similar expression levels. The strand which genes were on was also considered, with the hypothesis that overlapping genes on the same strand were likely to have more divergent expression levels due to transcriptional interference, whereas the expression levels of overlapping genes in opposite orientation, and therefore on opposite strands, were predicted to be unaffected by each other. This has previously been shown to be the case for several genes in *C. albicans*, *CCT8* and *TRP1* have convergent overlap with each other and unaffected expression levels (Gerads and Ernst, 1998), two homologs of the human protein erbA overlap and are based on opposite strands, with the expression level of one unaffected by the other (Miyajima *et al.*, 1989), this has also been found to be the case for the human genes *TCP1* and *ACAT2* (Shintani *et al.*, 1999). The results suggest that expression levels of a pair of genes are generally unaffected when they overlap, and the strand which the genes are placed has no affect upon this difference (Chi-square test, 1 d.f., $p = 0.054$; Table 3.11). This result may be because such a small number of pairs of overlapping genes were identified on the same strand, and therefore not enough were present to statistically represent a difference in expression. However, the results do support the alternative hypothesis, with a higher number of overlapping genes on the same strand with similar expression levels (Table 3.11). Surprisingly, these findings contradict previous studies in yeast where gene overlap has been linked to expression differences (Nagalakshmi *et*

al., 2008, Gagneur *et al.*, 2009), though these investigations were carried out in *S. cerevisiae* which could explain this contradiction.

Table 3.11 Number of pairs of overlapping genes that have significantly similar or different RPKM values (expression levels) and the link to strand identity

		Expression Level	
		Same	**Different**
Strand	**Same**	10	3
	Opposite	36	39

3.3.2.3 Relationship Between Structural Factors and Gene Expression

Overall gene expression levels were determined by calculating RPKM values using the alignment against the haploid *C. albicans* reference genome (Van Het Hoog *et al.*, 2007). GC content, ORF length, the codon usage values, CAI and GCB, and the usage of the CUG codon were then compared against the RPKM values to identify the relationship between these structural factors and expression.

A very weak, but significant, negative correlation is seen between GC content and RPKM (Figure 3.6a), and between ORF length and RPKM (Figure 3.6b) suggesting that expression levels decrease as GC content and ORF length increases (Spearman's correlation for GC content, $\rho = -0.033$, $p = 0.02$; Spearman's correlation for ORF length, $\rho = -0.083$, $p < 0.001$). The trend seen with ORF length conforms to the generally accepted hypothesis that shorter genes are more transcriptionally efficient and therefore have higher expression levels (Coghlan and Wolfe, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003). The GC content results support previous studies which suggest that higher GC content in some cases causes lower expression levels (Goncalves *et al.*, 2000). The exact mechanism behind this relationship is still unclear. It is possible that higher GC content leads to a higher melting temperature, which in turn slows DNA unwinding and therefore transcription. Other elements, such as chromatin organisation could also influence transcription rates. It has been demonstrated in *S. cerevisiae* that low GC content tracts of dA:dT are rigid and therefore unable to bend around and bind nucleosomes. These nucleosome unbound regions are then associated with increased accessibility for the transcriptional machinery and therefore higher gene expression (Mavrich *et al.*, 2008).

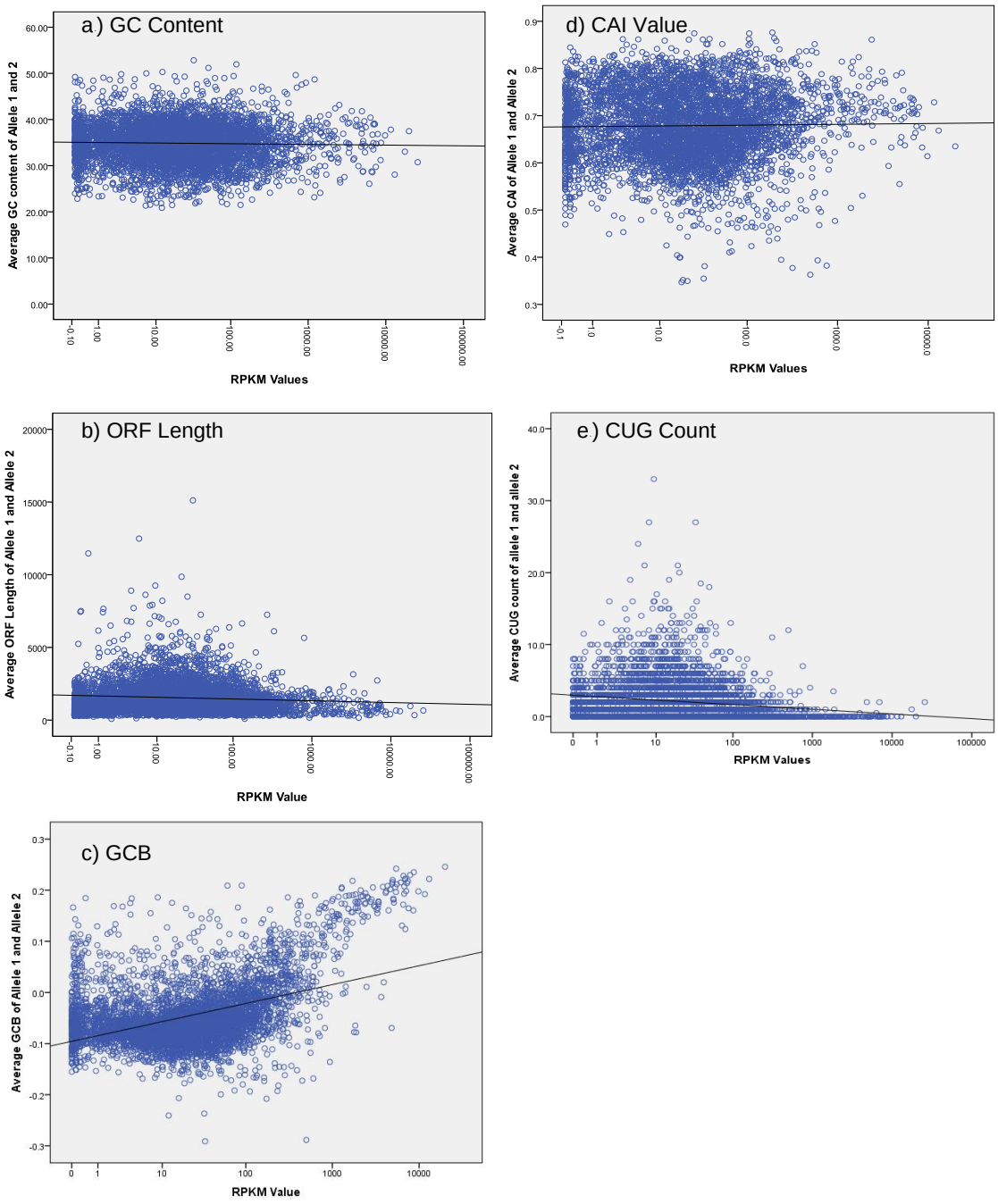


Figure 3.6 Correlation between a) average GC content of allele 1 and 2; b) average ORF length of allele 1 and 2; c) average GCB of allele 1 and 2; d) average CAI of allele 1 and 2; e) average CUG count of allele 1 and allele 2 and RPKM (reads per kilobase per million mapped reads) on a logarithmic scale (Spearman's correlation; a) ρ = -0.033, p = 0.02; b) ρ = -0.083, p < 0.001; c) ρ = 0.334, p < 0.001; d) ρ = 0.011, p = 0.439; e) ρ = -0.237, p < 0.001). Linear trend lines are shown by the black lines.

No significant correlation was found when comparing average CAI with RPKM values calculated from the haploid alignment (Spearman's correlation, $\rho = 0.011$, $p = 0.439$ – Figure 3.6d). This suggests that CAI does not reflect expression levels and contradicts previous studies in yeast (Coghlan and Wolfe, 2000) and the original paper in which CAI was developed (Sharp and Li, 1987). However, GCB was found to have a significant positive correlation with RPKM (Spearman's correlation, $\rho = 0.334$, $p < 0.001$; Figure 3.6c). In this study, a higher positive GCB value is associated with genes with high expression levels, whereas a low or negative GCB value, is associated with genes with low expression levels. This trend is what is expected as based on the original paper (Merkl, 2003) and suggests, that based on codon usage, genes with higher expression levels favour protein sequences which use more abundant tRNAs. This trend is particularly noticeable in the most highly abundant genes with RPKM values greater than 1000 (on a logarithmic scale) as seen in Figure 3.6c.

Unusually, *C. albicans* and other related *Candida* species translate the CUG codon as serine instead of leucine (Ohama *et al.*, 1993). This change occurred due to codon reassignment approximately 170 million years ago (Massey *et al.*, 2003). This has resulted in reduced usage of the CUG codon throughout these species (Butler *et al.*, 2009). Here, this theory is supported by our transcriptional data which shows a significant negative correlation between the number of CUG codons per ORF and RPKM (Spearman's correlation, $\rho = -0.237$, $p < 0.001$; Figure 3.6e). This demonstrates that genes with high expression have fewer CUG codons.

3.3.2.4 The Contribution of GC Content, Gene Length and Codon Usage to AEI

Following up on the observed correlations of structural factors with overall gene expression levels, the relationship between GC content, ORF length, codon usage and allelic expression imbalance was assessed. Although CAI values did not correlate significantly with overall gene expression, the relationship with AEI was still assessed to ensure that no significant results were overlooked. For each factor, the values obtained from the allele with the lowest and highest expression were compared. In addition, structural factors were also assessed in a cohort of genes with equally expressed alleles (fold difference ≈1) (Appendix I

Table II). This ensures that any trends observed were specific to genes with allelic expression imbalance.

GC content, GCB, CAI and CUG usage did not differ significantly between the alleles with lower and higher allele expression (two sample t-test, p = 0.76 – Figure 3.7a; two sample t-test, p = 0.83 – Figure 3.7c; two sample t-test, p = 0.20 – Figure 3.7d; two sample t-test, p = 0.32 – Figure 3.7e). The case was the same for alleles with equal expression (two sample t-test, p = 0.94 – Figure 3.8a; two sample t-test, 0.97 – Figure 3.8c; two sample t-test, p = 0.73 – Figure 3.8d; two sample t-test, p = 0.94 – Figure 3.8e). These results indicate that despite the overall correlation with gene expression, GC content and codon usage are unlikely to have a role in regulation of AEI. Following on from the overall gene expression results, CAI is also ruled out from directly influencing allelic expression imbalance. This is supportive of the results found by Muzzey *et al.* (2014) who show that CAI does not have a role in allelic expression imbalance at the translational level in *C. albicans*.

Although CAI and CUG usage did not differ significantly between the alleles, the mean difference in CAI and CUG usage (0.015 ± 0.023, 0.705 ± 1.70) is significantly larger in genes with AEI than genes with equal expression (0.005 ± 0.012, 0.124 ± 0.558) (two-sample t-test, p < 0.001 – Figure 3.9d and 3.9e). This suggests that although CAI and CUG usage in differentially expressed alleles do not follow the overall expression trend, it may still have an important role.

Surprisingly analysis of genes with AEI indicates that the ORF length of the allele with lowest expression is significantly shorter than the allele with higher expression (two sample t-test, p = 0.008; Figure 3.7b). This directly contradicts the results found when correlating ORF length with overall gene expression and the results found in previous studies in the yeast *Saccharomyces cerevisiae* (Coghlan and Wolfe, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003). This relationship is not seen in genes with equal allele expression, where no significant difference is observed between the ORF length of alleles (two sample t-test, p = 0.81 – Figure 3.8b). This suggests that the result is specific to the gene set with allelic expression imbalance. A possible explanation for this

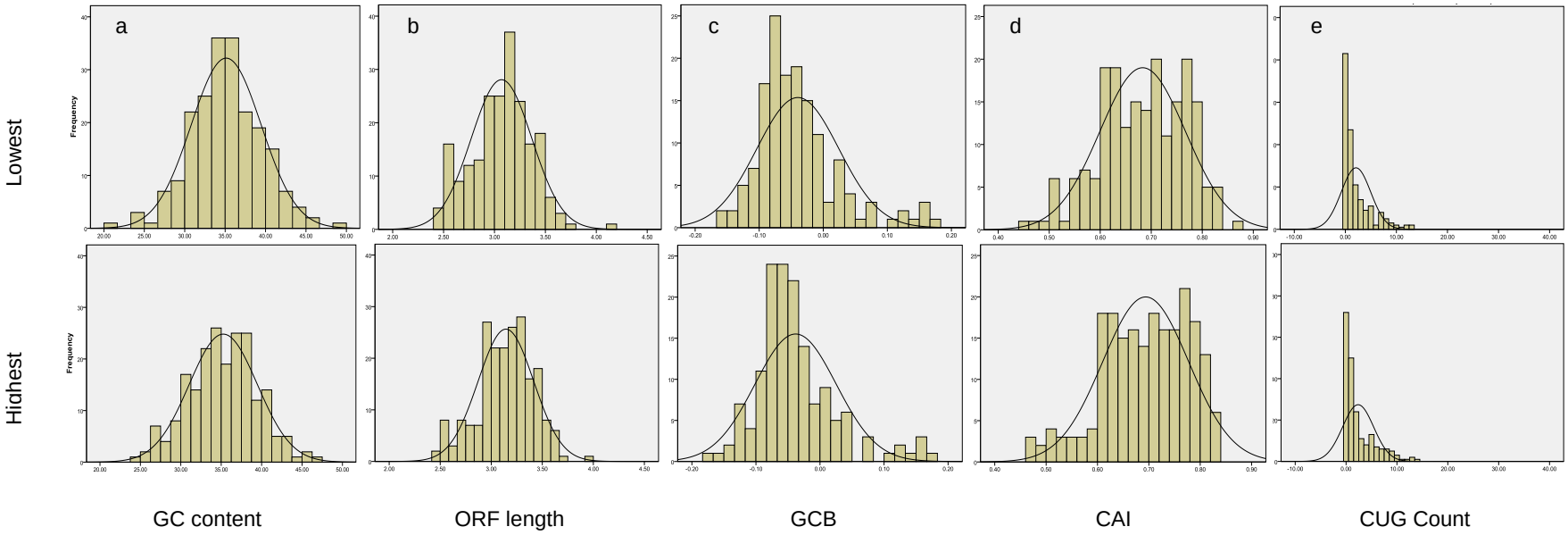


Figure 3.7. Histograms showing the distribution of a) GC content; b) allele length (after logarithmic transformation); c) GCB value; d) CAI value and e) CUG count of alleles with lower and higher expression from the genes with allelic expression imbalance. No significant differences were found for GC content (two sample t-test, p = 0.76), GCB values (two sample t-test, p = 0.83), CAI values (two sample t-test, p = 0.20) nor CUG count (two sample t-test, p = 0.32). The allele with lowest expression was found to have a significantly shorter length than the allele with higher expression (two sample t-test, p = 0.008). Black lines represent fitted distribution assuming normal distribution.

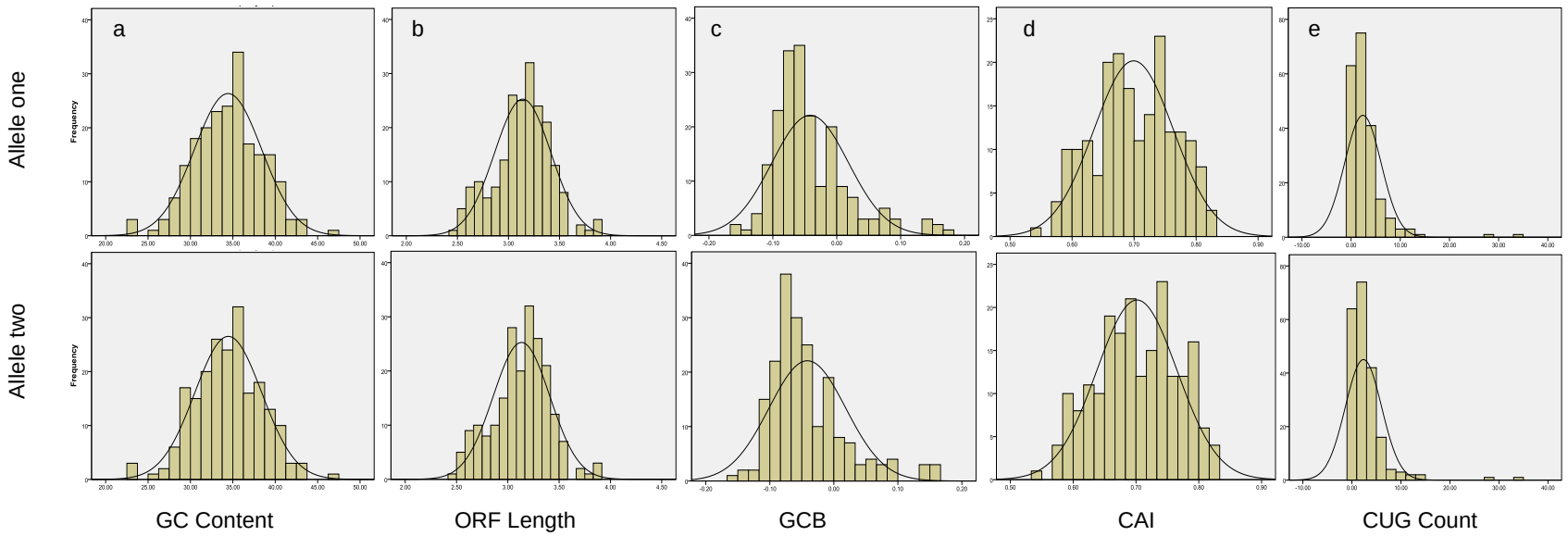


Figure 3.8. Histograms showing the distribution of a) GC content; b) allele length (after logarithmic transformation); c) the GCB value; d) the CAI value and e) the CUG count of allele one and allele two from the gene set of alleles with equal expression. No significant differences were found for GC content (two sample t-test, p = 0.94), ORF length (two sample t-test, p = 0.81), GCB values (two sample t-test, p = 0.97), CAI values (two sample t-test, p = 0.73) nor CUG count (two sample t-test, p = 0.94). Black lines represent fitted distribution based assuming a normal distribution.

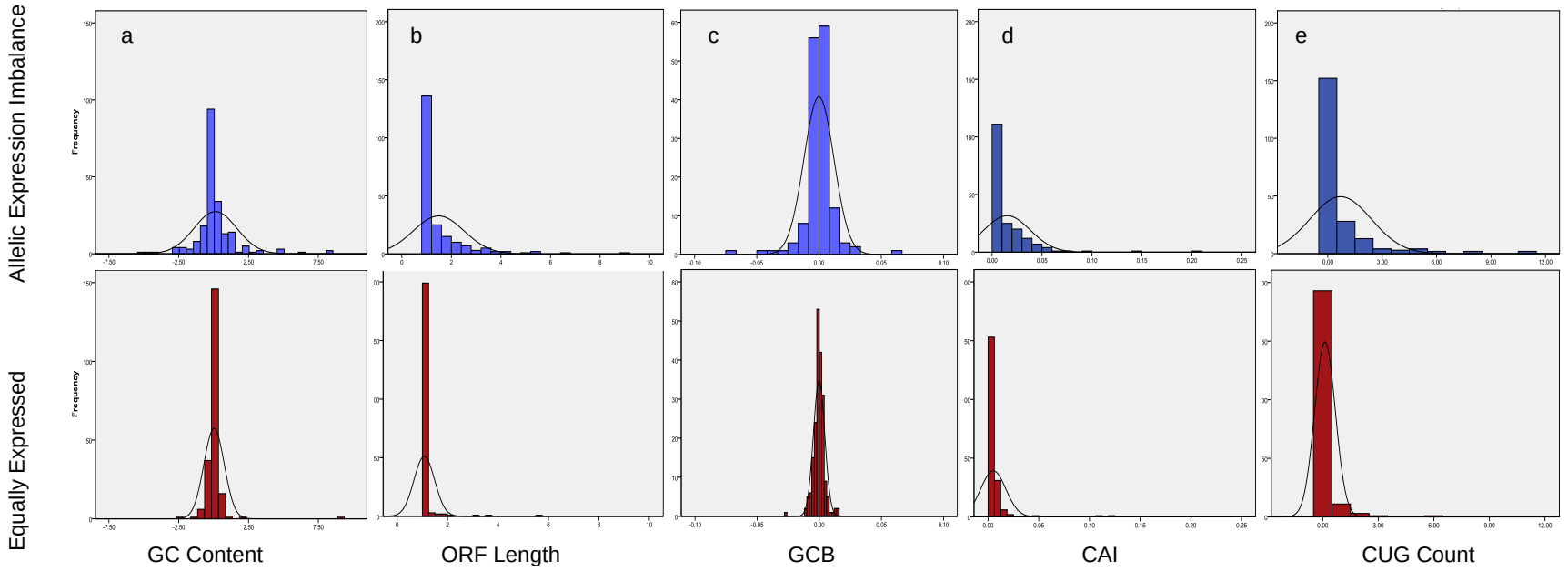


Figure 3.9. Histograms showing the distribution of a) differences in allele GC content; b) fold difference in allele length; c) differences in GCB; d) differences in CAI and e) differences in CUG count between genes with AEI (blue) and equally expressed alleles (red). A significantly larger variance in the data-set is observed for genes with AEI with each measure (F-test, $p < 0.001$). The difference in CAI and CUG count is significantly larger in genes with AEI (two sample t-test, $p < 0.001$). Black lines represent the fit to a normal distribution.

observation is that longer genes have been shown to be more “sequenceable” as more fragments are produced and therefore sequenced. This gives the impression that longer genes (or alleles) have higher expression levels, and it has been found that this difference in expression is not corrected for sufficiently by using RPKM (Bullard *et al.*, 2010a).

Another possible explanation may be found when looking at the relationship between differences in these structural factors and the percentage protein identities calculated in section 3.3.1.4. For all five factors, the difference in GC content, GCB, CAI and CUG usage and the fold difference in ORF length, there is a significantly larger variance in differentially expressed alleles when compared to equally expressed alleles (F-test, $p < 0.001$ in each case; Figure 3.9). In the case of GC content, ORF length, CAI and CUG usage, a strong negative correlation is observed between this difference and percentage protein identity (Spearman’s correlation, $\rho = -0.630$, $p < 0.001$; Spearman’s correlation, $p = -0.814$, $p < 0.001$; Spearman’s correlation, $\rho = -0.533$, $p < 0.001$; Spearman’s correlation, $\rho = -0.527$, $p < 0.001$; Figure 3.10). This suggests that as the difference in GC content, length and codon usage increases, the percentage protein identity decreases, and therefore the alleles may become more functionally distinct. A positive correlation is also observed between the fold difference in length and the difference in GC content, CAI and in CUG usage (Spearman’s correlation, $\rho = 0.560$, $p < 0.001$; Spearman’s correlation, $\rho = 0.481$, $p < 0.001$; Spearman’s correlation, $\rho = 0.456$; Figure 3.11). It can be inferred from this that alleles with large size differences have larger changes in GC content and codon usage. This directly results in more non-synonymous substitutions. Additionally, shorter alleles could be missing functional domains entirely. It now remains to be investigated whether the imbalance in the allele expression levels is due to functional divergence of the alleles caused by these structural differences. If so, it can be elucidated that although the difference in allele length contributes towards allele-specific functions, it is not the sole causal factor or control element behind allelic expression imbalance.

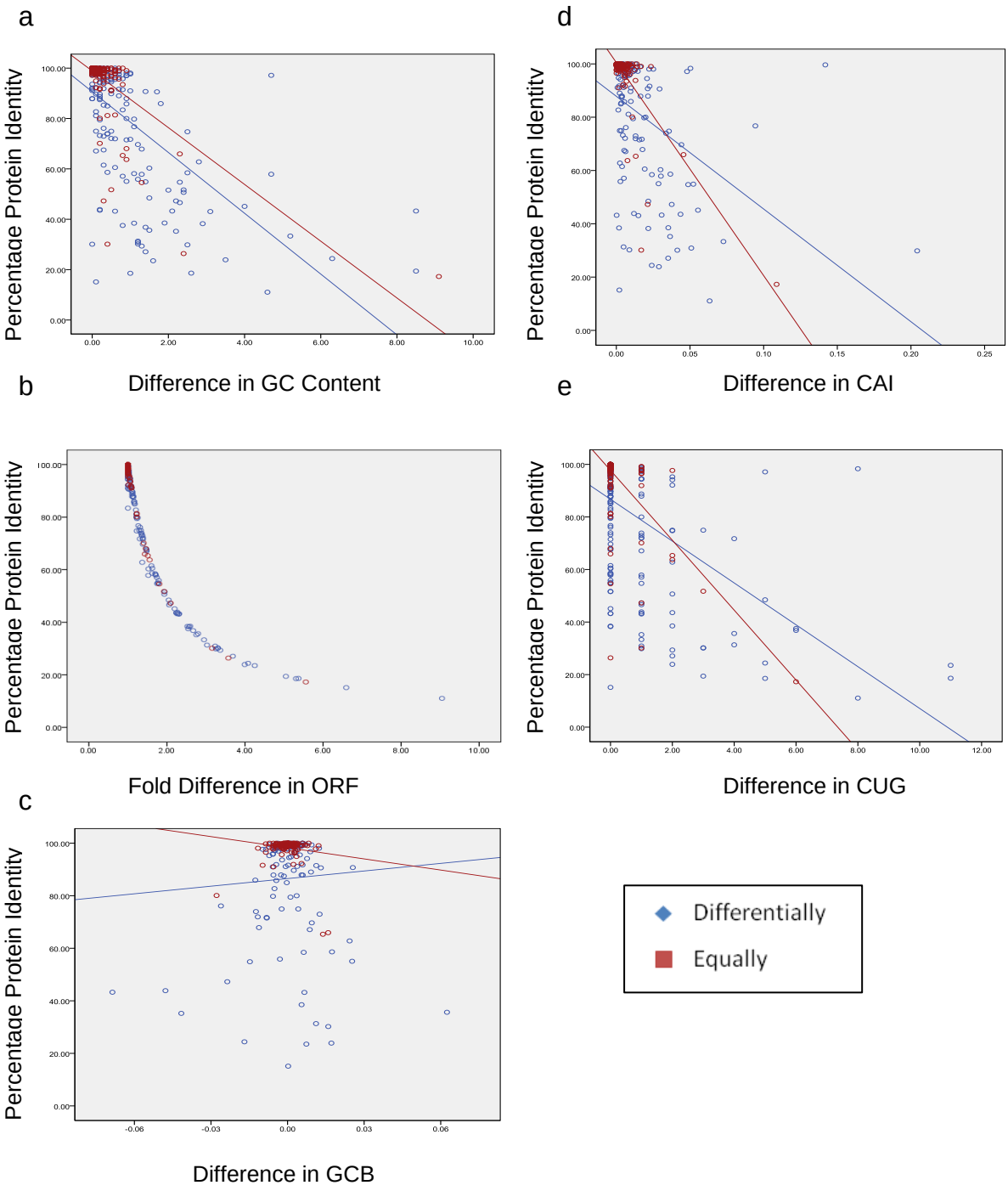


Figure 3.10. Correlation between a) difference in GC content; b) fold difference in allele length; c) difference in GCB; d) difference in CAI and e) difference in CUG count with percentage protein identity of alleles which are differentially (blue) and equally (red) expressed. (Correlation values for all genes [Spearman's correlation; a) $\rho = -0.630$, $p < 0.001$; b) $p = -0.814$, $p < 0.001$; c) $\rho = -0.004$, $p = 0.942$; d) $\rho = -0.533$, $p < 0.001$; e) $\rho = -0.527$, $p < 0.001$]). Fitted lines represent the linear trend line, except for b) where no appropriate trend lines were available to represent the data.

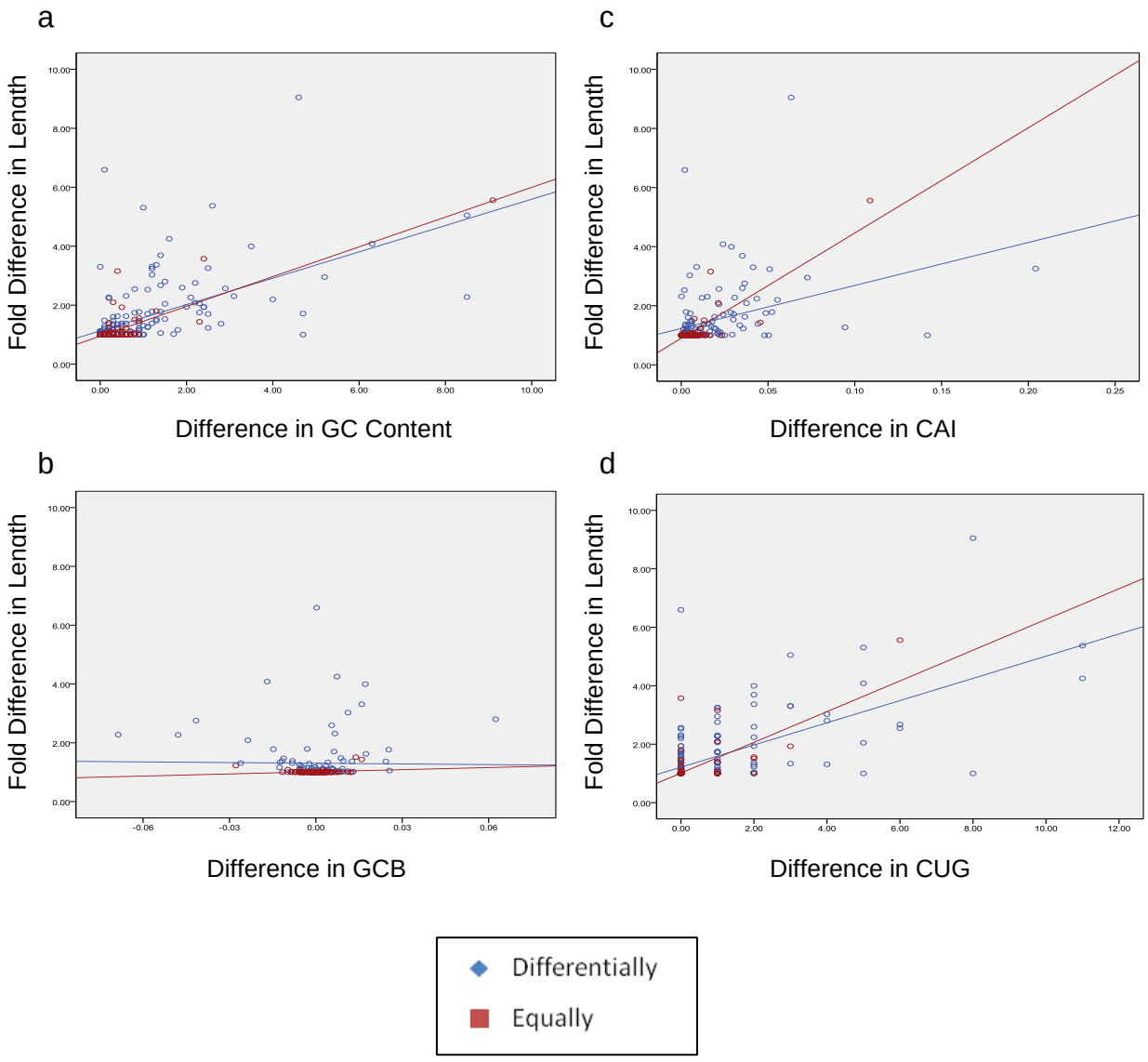


Figure 3.11. Correlation between a) difference in GC content; b) difference in GCB; c) difference in CAI and d) difference in CUG count with fold difference in ORF length in genes with alleles which are differentially (blue) and equally (red) expressed. (Correlation values for all genes [Spearman's correlation; a) $\rho = 0.560$, $p < 0.001$; b) $p = 0.498$, $p = 0.037$; c) $\rho = 0.456$, $p < 0.001$; d) $\rho = 0.481$, $p < 0.001$]). Fitted lines show linear trend lines.

3.3.3 Attempts at Expression Validation using qPCR, Restriction Enzyme Digests and Western Blotting

Validation of expression levels determined through RNA sequencing has previously been achieved using various methods including microarrays (Marioni *et al.*, 2008, Bloom *et al.*, 2009, Esteve-Codina *et al.*, 2011, Guida *et al.*, 2011) and qPCR (Marioni *et al.*, 2008, Nagalakshmi *et al.*, 2008, Bloom *et al.*, 2009, Bruno *et al.*, 2010). However, validating allelic expression imbalance has added complications with finding a method with the required high levels of both specificity and sensitivity. Therefore, four methods were adopted here in an attempt to verify the results produced from RNA sequencing, two types of allele-specific qPCR, allele-specific restriction enzyme digestion of cDNA, and western blotting using strains with individually tagged alleles.

3.3.3.1 Validation using Allele-Specific qPCR and TaqMan Probes

Allele-specific qPCR has previously been used to both identify AEI and validate RNA sequencing results using TaqMan Genotyping Assays-By-Design (Applied Biosystems) (Harries *et al.*, 2006, Tuch *et al.*, 2010a). In this system, a region of DNA around 150 – 200 bp in length is amplified. Within this region, two probes are designed to bind over a SNP-containing region; a probe specific to “allele one” and a probe specific to “allele two”. Each probe is conjugated to a fluorescent probe, FAM or VIC, which is released and quantified when the probe binds to the DNA.

In this instance two genes with significant AEI were chosen for validation using allele-specific qPCR, *VPS1* and *CDC6.* In total, three assays were tested for *VPS1* and two assays for *CDC6* (see Table 3.1). Initially specificity and efficiency of the assays were tested using a 10x fold serial dilution of genomic DNA, in triplicate, from the wild-type strain SC5314, where both alleles are expected to be detected in equal quantities, and also from heterozygous knockout strains, where expression of just one allele is expected to be detected. For each concentration of DNA, the average Ct value of each probe was calculated, with increasing concentrations of DNA predicted to have lower Ct values. Figure 3.12 demonstrates the set of graphs that would be produced when plotting DNA concentration against average Ct value if the specificity and efficiency of the probes is optimal. Of the five assays tested, none showed

sufficient specificity of the probes with the wrong probe being detected in at least one heterozygous knockout strain in each assay. Figure 3.13 shows the results of testing one assay as an example.

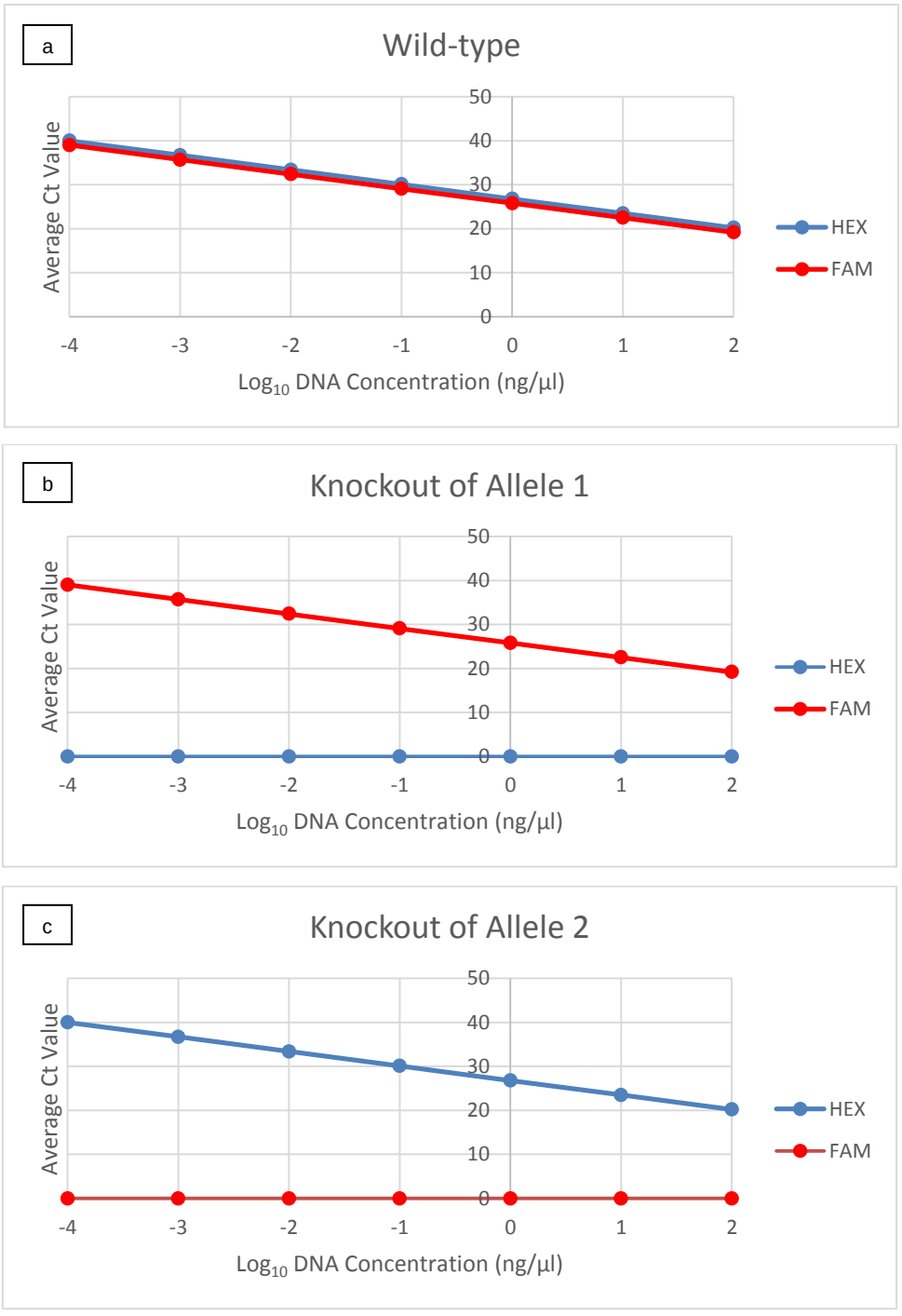


Figure 3.12 Predicted plots from a TaqMan genotyping assay with optimal efficiency and specificity. The average Ct value of the HEX probe (blue) and FAM probe (red) is plotted against the Log_{10} DNA concentration. The HEX probe is designed to bind to allele one and the FAM probe is designed to bind to allele two. Plots demonstrate the results for genomic DNA isolated from the a) wild-type strain, b) a heterozygous knockout of allele one and c) a heterozygous knockout of allele two.

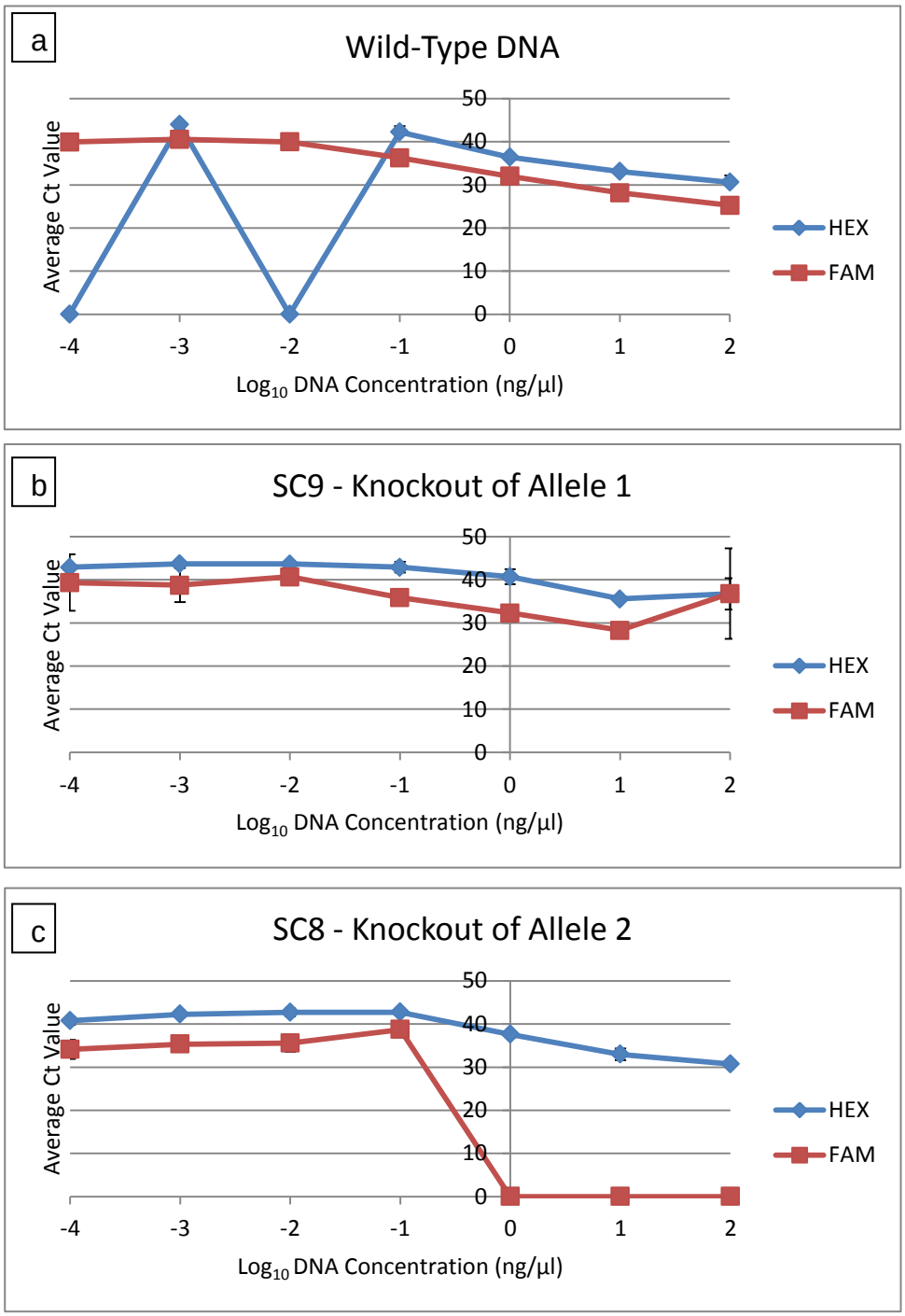


Figure 3.13 Specificity of TaqMan genotyping assays. The average Ct value of the HEX probe (blue) and FAM probe (red) is plotted against the Log_{10} DNA concentration. The HEX probe is designed to bind to allele one and the FAM probe is designed to bind to allele two. Errors bars represent ± one standard deviation. The results shown here are for *CDC6* assay 1 but are representative of all five assays tested. a) Wild-type SC5314 DNA shows approximately equal detection and efficiency of the probes when DNA concentrations are above 0.1 ng/µl. b) SC9, the knockout of allele one, shows approximately equal detection of both probes where only FAM is expected. c) SC8, the knockout of allele two, shows probe specificity, with only HEX detected at concentrations above 0.1 ng/µl.

Efficiency of each probe was calculated, as described in section 3.2.6.1.5, using the range of DNA concentrations which have a linear relationship with the average Ct values. This range varied from probe to probe. To calculate efficiency, the linear trend line was plotted, and the gradient (Δx/Δy) calculated. Ideally, gradients will be about -3.3, equating to an efficiency of the probe of around 100%. Table 3.12 shows that on average, efficiencies were either below what was expected, or were "too efficient" i.e. average Ct values did not decrease sufficiently as DNA concentrations increased. Figure 3.14 graphically represents the range of DNA concentrations, and subsequent linear equations, used to calculate the efficiencies for one assay. This was repeated for all five assays. Therefore, due to the lack of specificity and poor efficiencies of the probes tested it was decided that these assays would not be taken forward for use with cDNA to validate the levels of allelic expression imbalance.

Table 3.12 Efficiencies of TaqMan genotyping assays

Assay	**Probe**	**Gradient**[1]	**Efficiency**
CDC6 Assay 1	FAM	-3.695	86.48%
	VIC	-3.8267	82.52%
CDC6 Assay 2	FAM	-4.028	77.12%
	VIC	-2.785	128.59%
VPS1 Assay 1	FAM	-3.8633	81.49%
	VIC	-3.677	87.05%
VPS1 Assay 2	FAM	-3.627	88.67%
	VIC	-3.7677	84.25%
VPS1 Assay 3	FAM	-3.5513	91.24%
	VIC	-3.827	82.52%

1. Calculated as (Δx/Δy).

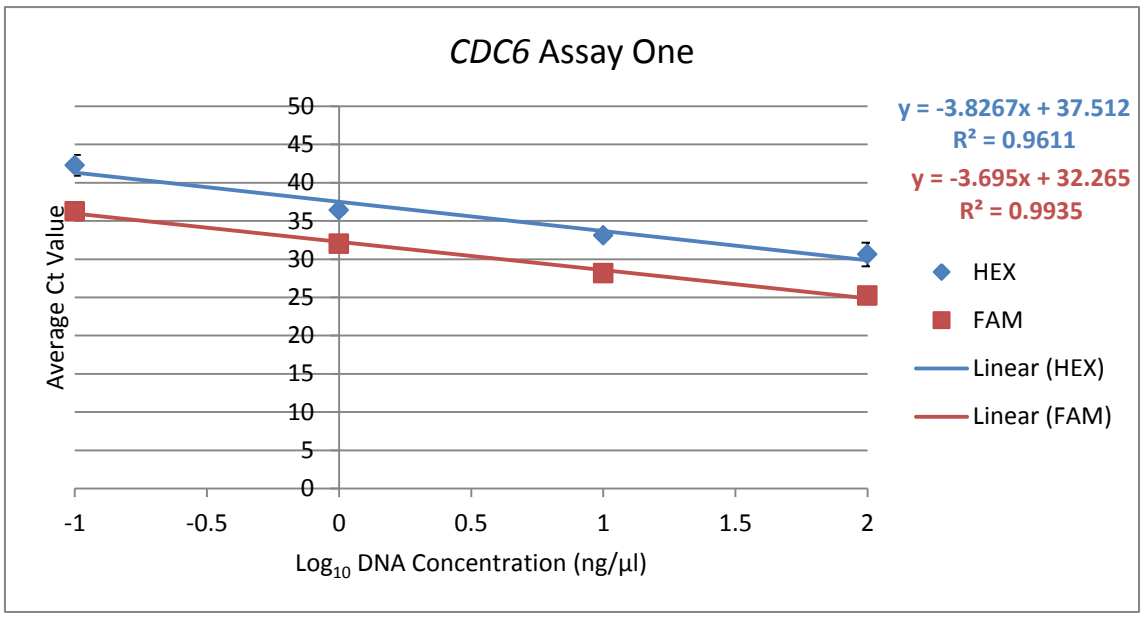


Figure 3.14 Graphical representation of efficiency calculations for TaqMan genotyping assays. The average Ct value of the HEX probe (blue) and FAM probe (red) is plotted against the Log_{10} DNA concentrations which are in a linear range. Errors bars represent ± one standard deviation. Linear trend lines are plotted and gradient calculations shown for the *CDC6* assay 1. This was repeated for all five assays. The FAM probe (red) has an efficiency of 86.48% and HEX probe (blue) has an efficiency of 82.52%.

3.3.3.2 Validation using Allele-Specific qPCR and SYBR® Green

As validation proved unsuccessful using a TaqMan genotyping assay (see previous section), validation using allele-specific primers and a traditional qPCR system with SYBR® green was attempted. Here two sets of primers were designed, each set specific to a certain allele. During qPCR, SYBR® green binds to the double stranded DNA product produced from amplification of the cDNA from the specific primers, and the fluorescence is quantified. The threshold value (Ct) can then be compared for each allele and the difference in expression level inferred. This system does, however, rely on the efficiency of the primer sets being equal (Bustin *et al.*, 2009).

Allele-specific primers were designed for three genes identified by RNA sequencing as having significantly divergent levels in allele expression: *CDC6, VPS1* and *RBT4.* Primers were designed to bind over SNP locations to ensure specificity, as demonstrated in Figure 3.15 (see Table 3.2 for a full list of

oligonucleotides tested). Computational analysis of specificity was carried out using BLASTn as described in section 3.2.6.1.6. Amplification of genomic DNA from the wild-type strain, SC5314, and appropriate heterozygous knockout mutants, was used to test the specificity of primer sets. Gradient PCR was used to optimise the annealing temperature for specificity. Figure 3.16 shows that the combination of CDC6-F7-1A with CDC6-R7-1A is specific for allele one of *CDC6*, as amplification only occurs with DNA from the wild-type and the knockout of allele two. CDC6-F7-2A with CDC6-R7-2A is specific for allele two of *CDC6*, as amplification only occurs with DNA from the wild-type and the knockout of allele one. This specificity for *CDC6* occurs under three different annealing temperatures: 56.6 °C, 58.3 °C and 60.1 °C. Therefore these primer sets were taken forward, and tested for efficiency, with the median annealing temperature of 58.3 °C.

Figure 3.17 shows that combining RBT4-F4-1A with RBT4-R4-1A, and RBT4-F4-2A with RBT4-R4-2A shows specificity for *RBT4* allele one and two respectively. This specificity for *RBT4* only occurs when the annealing temperature is 56.6 °C. However, as the intensity of the bands is observably different, it can be concluded that the amplification efficiency of the primer pairs is not equal. Therefore use of these oligonucleotides for validation was not taken any further.

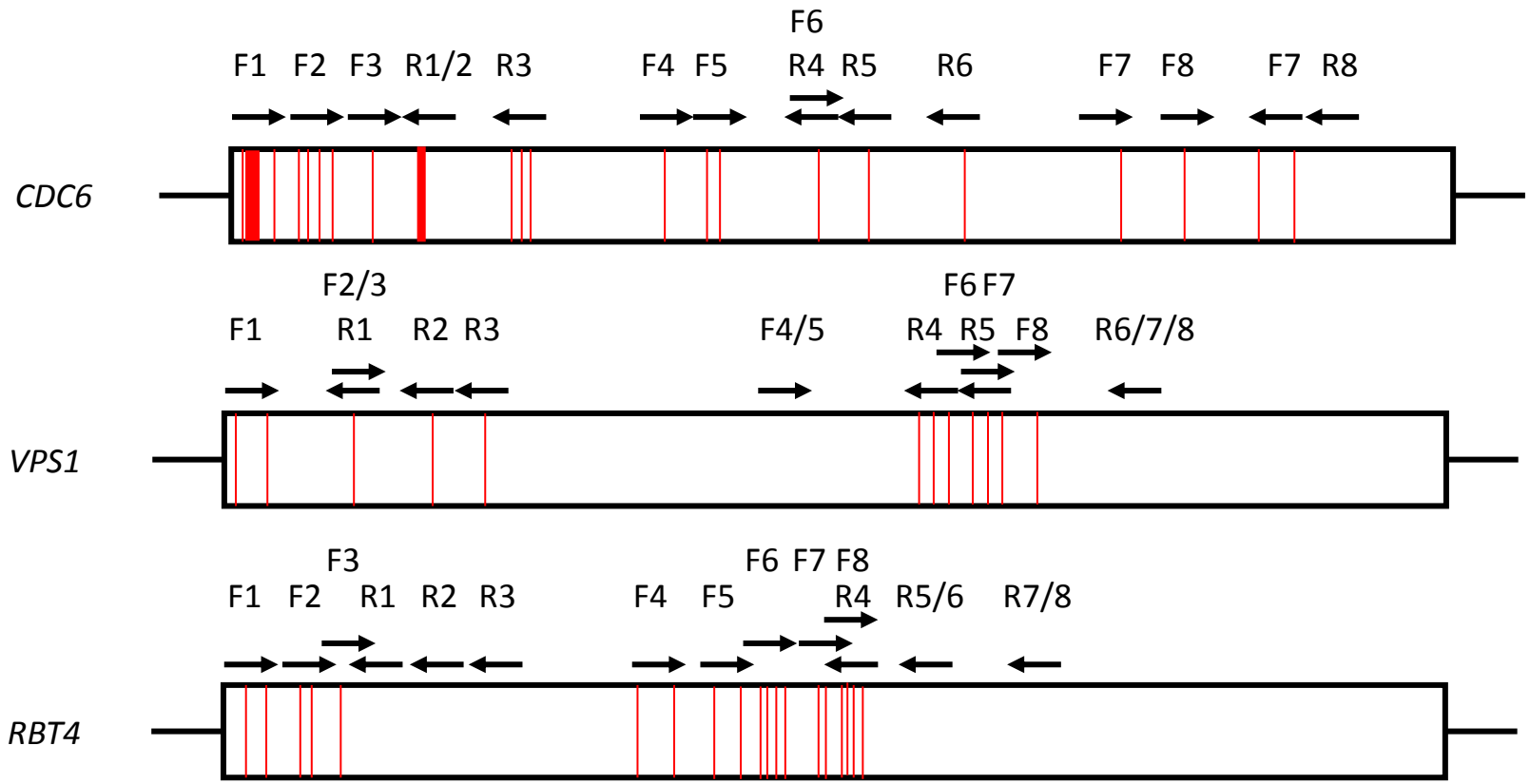


Figure 3.15 Allele specific primer locations for *CDC6, VPS1* and *RBT4.* SNP locations are indicated by red lines. Note, not to scale.

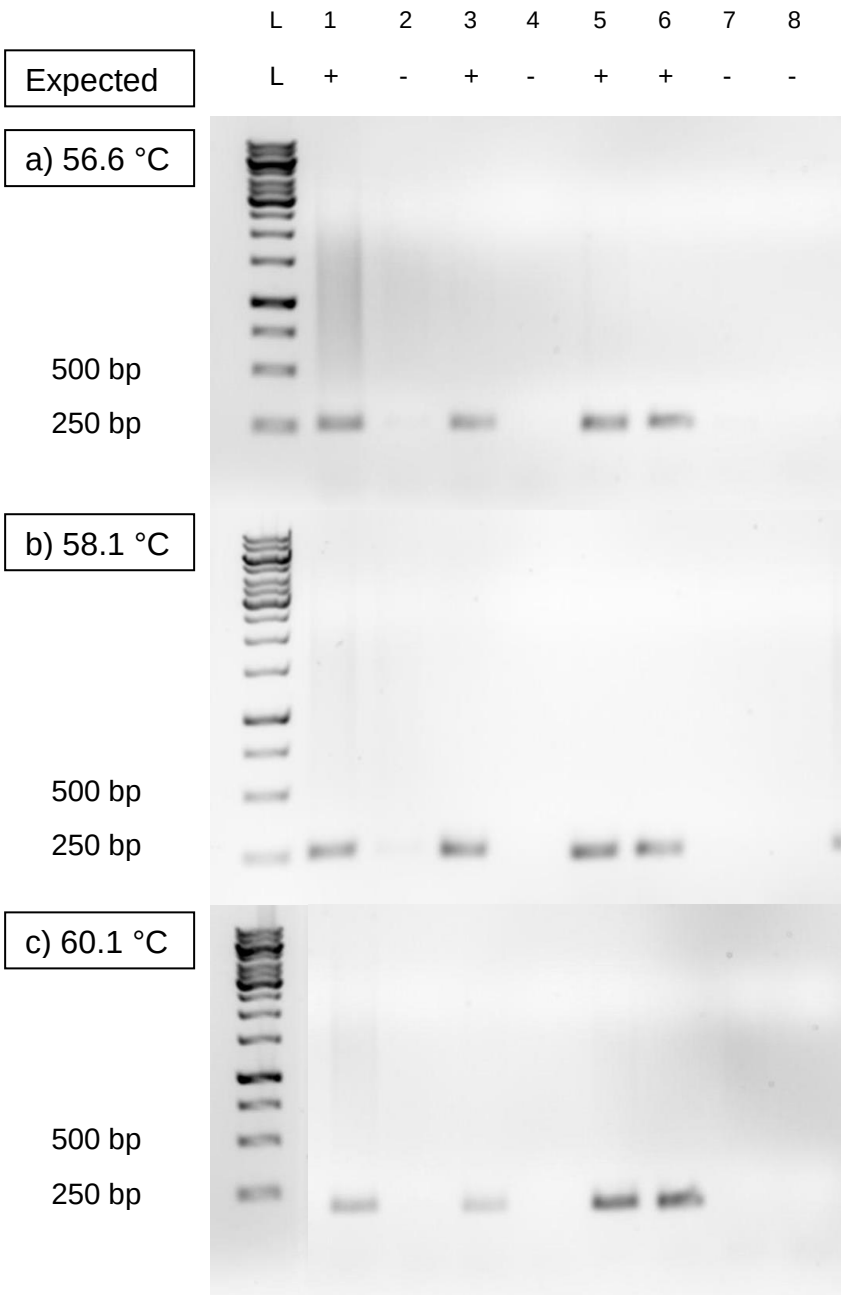


Figure 3.16 Allele-specific oligonucleotide primer combinations for *CDC6.* Lanes 1 – 4 contain oligonucleotide primers specific to allele 1. Lanes 5 – 8 contain oligonucleotide primers specific to allele 2. Expected band size 264 bp. Lanes 1 and 5 show amplification of wild-type DNA. Lanes 2 and 6 contain DNA from the heterozygous knockout of allele 1, SC9, therefore amplification only occurs with the allele 2 primers. Lanes 3 and 7 contain DNA from the heterozygous knockout of allele 2, SC8, therefore amplification only occurs with the allele 1 primers. Lanes 4 and 8 are negative controls with no template DNA. Annealing temperatures are a) 56.6 °C, b) 58.3 °C and c) 60.1 °C. L denotes a 1 kb DNA ladder (Fermentas, UK).

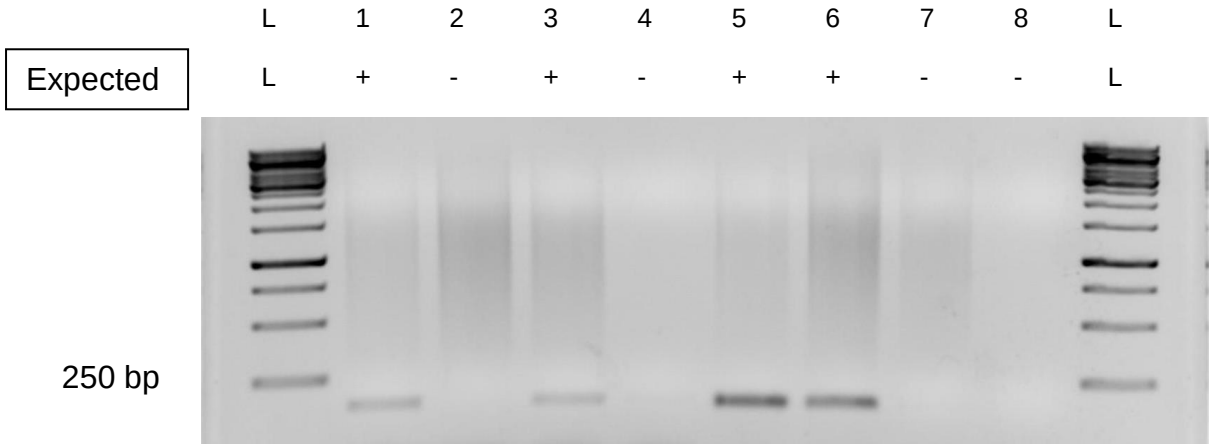


Figure 3.17 Allele-specific oligonucleotide primer combinations for *RBT4.* Lanes 1 – 4 contain oligonucleotide primers specific to allele 1. Lanes 5 – 8 contain oligonucleotide primers specific to allele 2. Expected band size 169 bp. Lanes 1 and 5 show amplification of wild-type DNA. Lanes 2 and 6 contain DNA from the heterozygous knockout of allele 1, SC3, therefore amplification only occurs with the allele 2 primers. Lanes 3 and 7 contain DNA from the heterozygous knockout of allele 2, SC4, therefore amplification only occurs with the allele 1 primers. Lanes 4 and 8 are negative controls with no template DNA. Annealing temperature is 56.6 °C. L denotes a 1 kb DNA ladder (Fermentas, UK).

No combinations of the oligonucleotide primers tested were found to be specific for the *VPS1* gene.

Efficiency of the specific oligonucleotide primer combinations for *CDC6* was tested using qPCR with SYBR® green and a range concentrations of genomic DNA (in triplicate) from the wild-type strain SC5314. Efficiency of the primer pairs was calculated using the average threshold value (Ct) for each DNA concentration as described in section 3.2.5.1.6. Ideal Ct values fall in the range of 10 – 30, and ideal efficiencies lie around 100%. Figure 3.18 shows the graphical representation of the efficiency calculation for the primers specific to allele one. Fluorescence, and therefore amplification, was detected at concentrations of genomic DNA over 1 ng/µl. However, the efficiency (86.7%) was too low for use. Figure 3.19 shows the graphical representation of the efficiency calculation for the primers specific to allele two. Ct values were high, suggesting little detection of fluorescence, and inconsistent across all concentrations of DNA. The efficiency of the oligonucleotides was also calculated at 118%. Therefore, due to the unequal and poor efficiency of the

oligonucleotide pairs, validation using allele-specific primers and SYBR® green was taken no further.

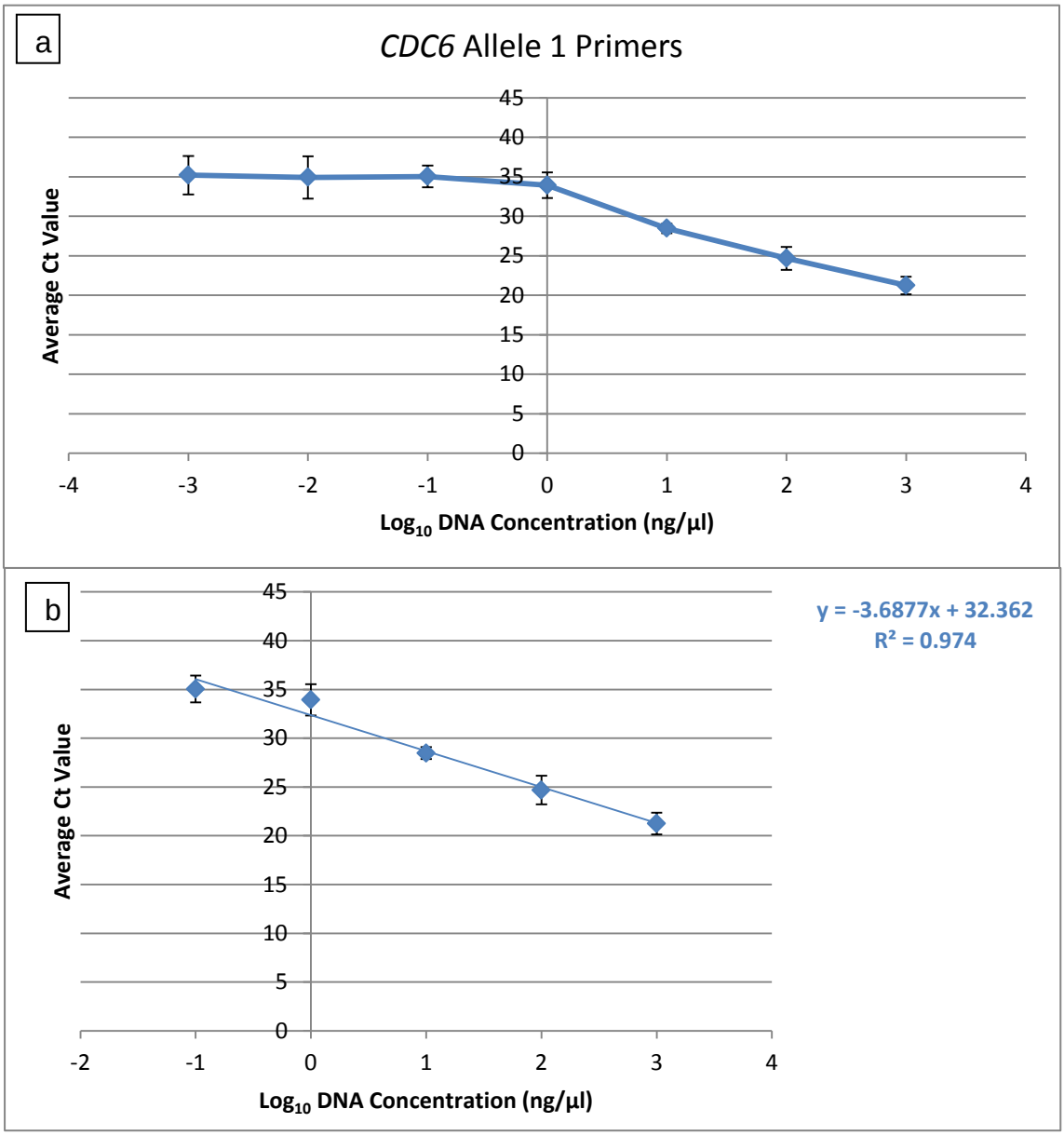


Figure 3.18 Graphical representation of efficiency calculations for allele-specific primers for *CDC6* allele one. The average Ct value is plotted against the Log_{10} DNA concentrations for; a) all DNA concentrations. It can be seen here that amplification only occurs at DNA concentrations above 1 ng/µl; and for b) DNA concentrations which are in a linear range. The gradient of the line has been calculated as used to calculate the primer efficiencies at 86.7%. Errors bars represent ± one standard deviation.

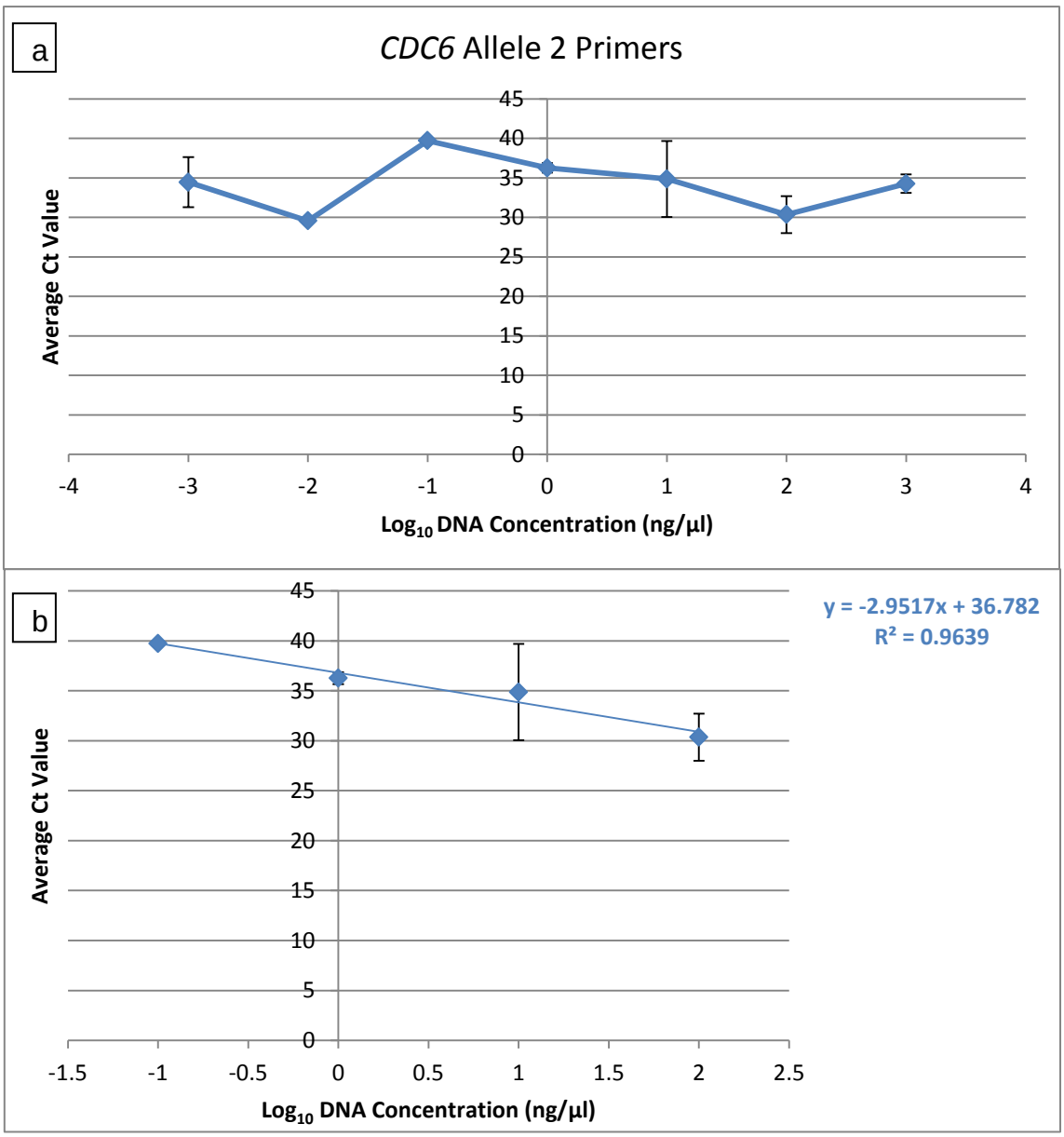


Figure 3.19 Graphical representation of efficiency calculations for allele-specific primers for *CDC6* allele two. The average Ct value is plotted against the Log_{10} DNA concentrations for; a) all DNA concentrations. It can be seen here that amplification is poor for all concentrations; and for b) DNA concentrations which are in a linear range. The gradient of the line has been calculated as used to calculate the primer efficiencies at 118.0%. Errors bars represent ± one standard deviation.

3.3.3.3 Validation using Allele-Specific Restriction Enzyme Digests

Allele-specific restriction enzyme digests of PCR products amplified from cDNA were designed to verify the levels of AEI. This has previously been used to verify AEI identified by RNA sequencing in maize (Zhang *et al.*, 2011). Allele-specific restriction digests produce different sized fragments for each allele. The expression levels of each allele can then be inferred from the intensity of the bands from each allele through use of DNA gel electrophoresis.

Figure 3.21 shows the results when verification was attempted for the gene *VPS1.* PCR amplification of the gene from genomic DNA or cDNA was followed by digestion with the enzyme HaeIII and separation of the fragments using DNA gel electrophoresis. Here it can be observed that digestions of PCR products amplified from genomic DNA of heterozygous knockouts of each allele produce distinct differences in band sizes (for expected sizes see Figure 3.20 and Table 3.4). The strain containing only allele one, SC7, (lane 6) has three bands differing in length, 314 bp, 172 bp and 123 bp, whereas the strain containing only allele two, SC6, (lane 5) has a single band of around 290 bp in length. This is likely to be a combination of the bands at 288 bp and 295 bp in length, with the 26 bp band not being visible. Digestion of PCR products amplified from genomic DNA of the wild-type strain SC5314 (lane 4) produces a combination of these band sizes, although the three bands relating to allele one are clearer. Although the results from digestion of PCR products amplified from cDNA are faint (lanes 1 – 3), the pattern of bands closely resembles that of the allele two knockout strain, SC7. Therefore it can be suggested that allele one has a more intense band, and therefore higher expression, than allele two. This is especially evident in lane 3. RNA sequencing results indicate that *VPS1* has significantly higher expression levels of allele one (see section 4.3.1) which is loosely supported by these results. However, the pattern produced from the wild-type DNA, where allele one and allele two are present in equal quantities, is indistinguishable to that produced from the heterozygous knockout strain containing only allele one, SC6. This is suggestive of an unequal amplification efficiency of the alleles in the PCR step, leading to a bias of results towards allele one. Further attempt at validation using restriction digests was therefore stopped.

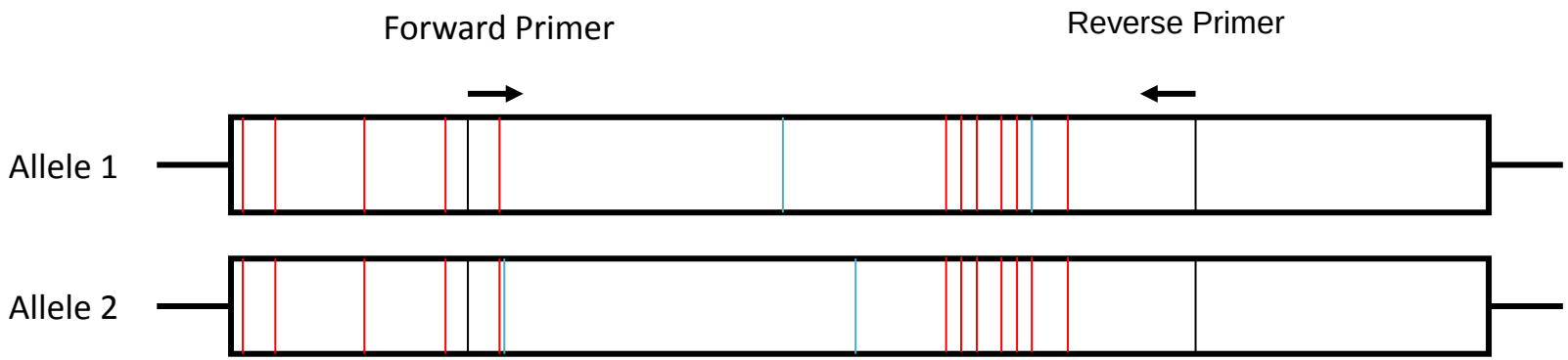


Figure 3.20 Allele specific restriction enzyme digest locations for *VPS1* PCR products. Arrows and black lines indicate the region amplified by PCR prior to digestion. Blue lines show restriction digestion sites. SNP locations are indicated by red lines. Allele 1 produces three fragments; 314 bp, 128 bp and 172 bp in size. Allele 2 produces three fragments; 26 bp, 288 bp and 295 bp in size. Note, not to scale.

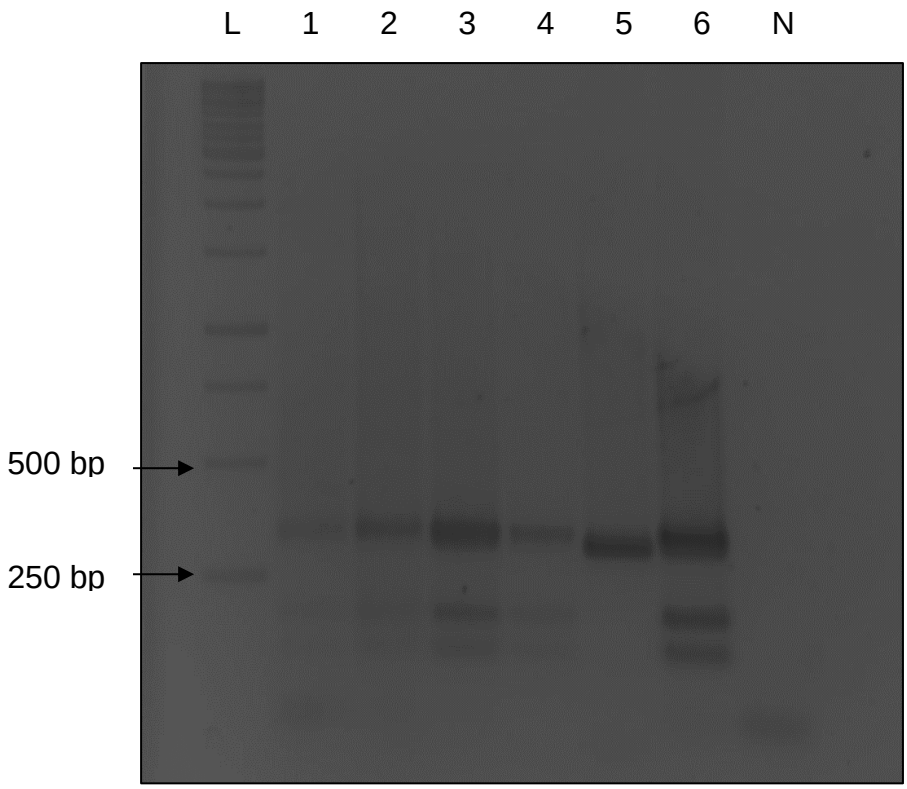


Figure 3.21 Allele-specific restriction enzyme digest of *VPS1* PCR products. Lanes 1 to 3 show digestion of PCR products amplified from three biological replicates of cDNA. Although faint, three distinct bands can be seen at 314 bp, 172 bp and 123 bp. Lane 4 shows digestion of PCR products amplified from genomic DNA extracted from the wild-type strain SC5314. Although faint, three distinct bands can be seen at 314 bp, 172 bp and 123 bp. Lane 5 shows digestion of PCR products amplified from genomic DNA extracted from a heterozygous knockout strain of allele 1, SC6. One clear band can be observed at approximately 290 bp. This is a combination of bands 288 bp and 295 bp in length. The 26 bp band is not visible. Lane 6 shows digestion of PCR products amplified from genomic DNA extracted from a heterozygous knockout strain of allele 2, SC7. Three distinct bands can be seen at 314 bp, 172 bp and 123 bp. Lane N shows a negative control digestion, containing all components except the PCR product. L denotes a 1 kb DNA ladder (Fermentas, UK).

3.3.3.4 Validation using Western Blotting

As an alternative to the use of qPCR and cDNA to validate the AEI identified by RNA sequencing, validation using protein expression was attempted. To achieve this, separate strains were constructed that had each allele tagged with a V5-6xHIS protein tag. This enabled western blotting to be used with an Anti-V5 antibody to compare and quantify the protein expression from each allele. Quantification of the actin protein was used as a loading control to normalise expression levels between samples. This validation was attempted for two genes, *CDC6* and *VPS1.* All strains used can be found in Table 2.1.

Growth rates of the tagged strains were assayed at 30 °C, as described in 2.13.1, to ensure that the presence of the protein tag did not inhibit growth. Appendix I Figure 1 and Appendix I Table III show that this was the case, and growth was not significantly affected by the protein tag in any of the constructs made.

To replicate the results from the RNA sequencing experiment as closely as possible, proteins were extracted when growth in YPD at 30 °C reached an OD at 600 nm equal to 1. Protein concentrations were quantified using a Bradford Assay before use in a Western blot. Figure 3.22 shows that this approach is not successful for validation of AEI. The assay was not sensitive enough to detect expression of *CDC6* alleles with no bands seen in Figure 3.22a. Expression was detected for the *VPS1* alleles, however the differences between the alleles were not strong enough to quantify the difference in protein expression levels and support the RNA sequencing data. Difficulties in detecting actin expression from the loading control with the *VPS1* strains also complicate the interpretation of results.

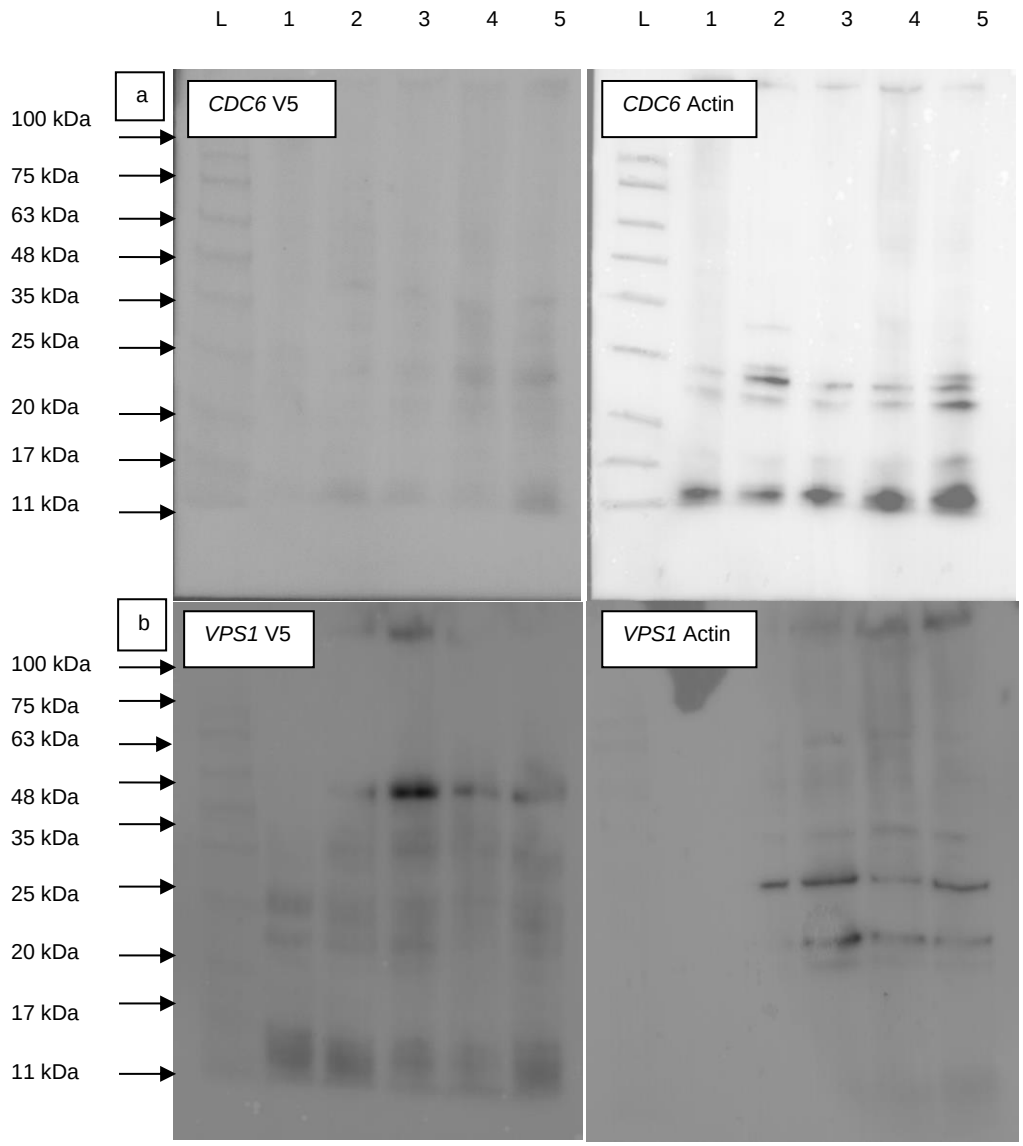


Figure 3.22 Western blots showing protein expression of V5 tagged alleles. In all cases, Lane 1 represents protein extracted from the wild-type strain SC5314, which contains no V5-6xHis tag. a) Expression of *CDC6* alleles; Lanes 1 and 2 show protein from the allele one tagged strains SC35 and SC36; Lanes 3 and 4 show protein from the allele two tagged strains SC61 and SC62. 1) Expression of V5 is undetectable with only faint bands present whereas 2) expression of actin is detectable across all strains. b) Expression of *VPS1* alleles; Lanes 1 and 2 show protein from the allele one tagged strains SC13 and SC14; Lanes 3 and 4 show protein from the allele two tagged strains SC27 and SC28. 1) Expression of V5 is detected, with allele one appearing to have higher expression with strain SC14 and lower expression with strain 13 however 2) shows expression of actin is also higher in strain SC14 and lower in strain 13 suggesting no differences in allele expression. L = BLUeye prestained protein ladder (Geneflow).

3.4 Discussion

3.4.1 The Biological Consequence of AEI

As Gene Ontology analysis of the genes identified with AEI revealed no significant patterns in predicted functionality, it can be concluded that AEI is not operating within a set of genes with a specific biological function or within a specific network. However, individual genes were identified that have virulence and pathogenesis related properties such as *ERB1, RBT4, SMI1, TAC1, TPO3 ALS1, RBT4, RCK2* and *RFG1*, amongst others. Therefore AEI may have a role in genes which are important during the infection process.

A possible hypothesis is that these alleles which differ in expression level may also differ in function due to the high level of heterozygosity seen across the *C. albicans* genome. Examples of divergence in allele function are highlighted in section 1.4.1, and show that differences in functionality of alleles occurs in a wide variety of genes that all operate in distinct biological processes. To further support our hypothesis that AEI may be linked to differences in allele function, the percentage protein identity between alleles was calculated. This was found to be on average significantly lower in genes with AEI than in genes with equal allele expression, suggesting that the protein sequences of alleles with AEI are significantly different. This question now needs to be taken into the laboratory to decipher if these differences in protein sequence occur within functional domains, and therefore if they have biological consequences. Investigating this question will be addressed at length in the next chapter.

It should be noted that difference in DNA sequence, which impact upon GC content and length, may produce bias in measurement of expression level by the sequencing technology. This matter is discusses further in the next section. Therefore the identification of AEI in a set of genes with large differences in DNA, and therefore protein, sequence may in fact be a false positive result due to bias in the sequencing technology, and that AEI may not be linked to functional differences in alleles at all.

3.4.2 Using RNA Sequencing to Identify Allelic Expression Imbalance

In previous studies of AEI in human T cells (Heap *et al.*, 2010), human cancer cells (Tuch *et al.*, 2010a), pig species (Esteve-Codina *et al.*, 2011), and yellow

baboons (Tung *et al.*, 2011) a haploid reference genome alongside variant calling has been used to identify AEI. Use of a haploid reference has, however, been associated with bias towards the reference genome (Stevenson *et al.* 2013). RNA sequencing technologies have been used here to identify significant levels of allelic expression imbalance in the yeast *Candida albicans* using a diploid reference genome with quantification of reads aligning uniquely to each allele. Across the genome, a total of 233 genes were identified as having AEI when cells were grown in YPD at 30 °C, of which 81 genes were identified with monoallelic expression.

Little evidence had been previously obtained for the presence of AEI in *C. albicans.* However, in 2013, Muzzey *et al.* developed a method to calculate expression levels of individual alleles in *C. albicans*. This study developed an improved diploid reference genome with phasing information and this information was used to identify AEI in a similar manner to our approach. In a follow up investigation, AEI at both the transcriptional and translational level was investigated (Muzzey *et al.*, 2014). However, as both studies lacked a measure of significance to the levels of allelic expression imbalance, there is no definitive list of genes with AEI from these studies that we can directly compare with our results. Therefore, the research carried out here presents a unique list of genes with significant levels of allelic expression imbalance which can be taken forward for various investigations, which has not previously been reported for *C. albicans.* However, it should be noted that the AEI in these genes is yet to be confirmed using an alternative method to RNA sequencing.

RNA sequencing has distinct advantages over other existing technologies, like microarrays and qPCR, for identification of AEI. For example, there is no need for prior assumptions of coding sequences, and, now with the advancement of *de novo* transcriptome assembly algorithms, there is no need for a reference genome, making RNA sequencing accessible to a wider range of organisms. Sensitivity has also been seen to be greater when using RNA sequencing allowing for identification of transcripts with lower expression levels (Wang *et al.*, 2009, Tuch *et al.*, 2010a).

However, RNA sequencing is not a perfect method and errors occur which could impact upon the identification of AEI. Library preparation introduces areas of bias in coverage. A 3' bias in coverage of reads is readily reported in cDNA fragmented libraries, like the ones used here (Wilhelm *et al.*, 2008), which can be improved by using RNA fragmentation (Mortazavi *et al.*, 2008). However, RNA fragmentation is depleted in reads at both transcript ends, making exact start and end points of transcripts unclear (Wang *et al.*, 2009). As our method relies on reads which align uniquely to SNP positions, if the SNPs are localised to a region of deeper coverage, the level of allele expression may be skewed in comparison to a gene where the SNPs lie in an area of low coverage. Heap *et al.* (2010) also found that reads had a bias in aligning towards alleles in a forward or reverse direction and that INDELs close to SNPs contribute to bias in measurement of allele-specific expression. To overcome this, all SNPs within 45 bp of an INDEL were removed in the 2010 study.

As well as bias in the mapping of reads, other issues have also been identified when using RNA sequencing to identify AEI. Variations in the "sequenceability" of alleles could impact upon expression levels measured. GC-rich sequences tend to be preferentially sequenced under the Illumina platform and a study by (Bullard *et al.*, 2010b) showed that a 5% differences in GC content conferred a 10% difference in expression level. Therefore alleles which differ vastly in GC content could be false positive results in our list of genes with AEI. The results of our analysis of structural factors indicate that alleles with divergent allele expression do have a significantly larger variance in differences in GC content than alleles with equal expression level. Therefore could the differences in expression that we have observed actually be a by-product of bias within the sequencing technology? A recent study has developed a method of measuring AEI using RNA sequencing which corrects for differences in the GC content of alleles to account for this difference in "sequenceability" (Skelly *et al.*, 2011), which could be used to overcome this problem in future investigations of AEI.

Problems also occur when reads map with equal efficiency to paralogous genes and when reads containing mutations are mapped incorrectly (Mortazavi *et al.*, 2008). It has been documented that some RNA sequencing aligners such as TopHat and MapSplice produce both false negative and false positive results

due to this problem of reads which can align to multiple regions in the genome (Zhang *et al.*, 2013). These factors could lead to reads being misaligned to the wrong allele or gene, and impact upon the levels of AEI identified.

RPKM was used here to normalise the expression levels of the alleles across sequencing replicates which differ in overall numbers of reads, and across alleles which differ in length. However, a meta-analysis of a number of RNA sequencing data-sets showed that normalisation using RPKM is biased by numerous factors including GC content as well as gene length and dinucleotide frequencies (Zheng *et al.*, 2011). A similar study also confirmed that RPKM is biased for gene length, producing more reliable results with longer genes (Bullard *et al.*, 2010a). Therefore, alleles which differ significantly in length and GC content are at risk of further bias in expression levels from the normalisation method we have used here.

Systematic errors have been readily identified in all sequencing technologies. Particular motifs and genome locations are more prone to base-call errors, which occur at the initial steps of sequencing, where bases are identified using imaging software. Positions proceeded by GG or GGC, or sequences at the end of reads have been identified as particularly problematic (Nakamura *et al.*, 2011). These errors have been recorded to occur as frequently as once in a 1000 base pairs and have been highlighted as problematic for studies investigating heterozygosity and allele-specific expression using RNA sequencing (Meacham *et al.*, 2011). Although the study investigating systematic errors developed a software package, SysCall, to identify likely error prone motifs, this was designed for use with human genome sequences and it is outside the scope of this project to modify the software for our use.

The method applied in the work presented here could still be improved by identifying the strand of origin for transcripts as it is currently unknown if reads have originated from the sense or anti-sense strand. As cases of imprinting, such as the example of the *Mest* locus (MacIsaac *et al.*, 2011), show relationships between sense and anti-sense transcription, identification of strand origin could reveal important patterns of expression linked to AEI. Several methods, all with varying degrees of success, are available for strand-

specific RNA sequencing, which fall under two classes, methods that involve ligation of strand-specific adaptors and methods that mark one strand of RNA or cDNA such as bisulfite modification (Levin *et al.*, 2010). In *C. albicans,* strand-specific RNA sequencing identified that 50% of all transcripts overlapped with their antisense transcript by at least one base pair (Tuch *et al.*, 2010b), however it was found that some but not all antisense transcripts repress the expression of the sense transcript, suggesting that overlap does not necessarily result in repression. This further suggests that identifying strand specificity could increase our knowledge of the mechanisms behind AEI in *C. albicans*.

3.4.3 The Impact of Structural Factors on AEI

Here structural factors related to genes with AEI were investigated. These factors included chromosomal location, overlap with neighbouring genes, GC content, length and codon usage, and the results of these findings are discussed below. However, there are structural factors that still remain unanalysed in terms of AEI in *C. albicans.* Some of these factors are discussed below to indicate other factors which may be important to investigate in the future.

3.4.3.1 Chromosomal Location

Although evidence has previously been seen suggesting that there is a relationship between AEI and chromosomal location, reports are often contradictory. In maize, imprinted genes have been seen to cluster in location when compared to the frequency of genes across the entire genome (Zhang *et al.*, 2011). In mice, position dependent gene silencing was also observed for the *HoxD* gene due to differences in chromosomal conformation of the alleles (Lonfat *et al.*, 2013). However studies looking at both autosomal monoallelic genes and genes with AEI in humans have shown that although some clustering of genes occurs, most genes are randomly distributed (Lo *et al.*, 2003, Savova *et al.*, 2013). These differences in findings may be due to inconsistencies in the measure of a “gene cluster” or differences between species.

Here, similar findings were found to those by Lo *et al.* (2003) and Savova *et al.* (2013). Although some individual clusters of genes with AEI were identified, the

patterns observed were random and infrequent. In the most part, the location of genes with AEI resembled the location of heterozygous genes across the genome. This does however support the findings of the Gene Ontology analysis; that AEI occurs across a random set of genes in *C. albicans,* with no underlying concerted biological mechanism.

3.4.3.2 Overlapping Genes

The extent of gene overlap in genes with allelic expression imbalance has previously been investigated in the yeast *S. cerevisiae* by Gagneur *et al.* (2009), where a total of 36 genes out of 371 with AEI were found to overlap with their neighbouring open reading frame, with varying patterns of expression. However, Gagneur found no clear mechanistic relationship between gene overlap and AEI. Here, only 4 genes with AEI were found to overlap suggesting that overlap of genes may be actively selected against in terms of AEI in *C. albicans*. However, our results also indicate that gene overlap has little impact on gene expression in *C. albicans* with a large proportion of genes which overlap on the same strand having similar expression levels. Therefore, it could be hypothesised that gene overlap is unrelated to AEI, and that such a small number of genes with AEI overlapping is purely coincidental. However it cannot be ruled out that gene overlap does contribute towards AEI at these few loci.

It was also observed that the mitochondrial genome has a particularly high percentage of overlapping features at 25%. It has been suggested that the purpose of overlapping genes is to conserve space in small genomes (Iwabe and Miyata, 2001, Johnson and Chisholm, 2004), and as the mitochondrial genome originated from a small prokaryotic ancestor it could have far higher percentage of overlapping features to increase efficiency. This theory needs further investigating as the purpose of overlapping genes is often disputed, with the suggestion that in bacterial genomes, overlapping genes are used for regulation of expression (Scherbakov and Garber, 2000). Further confusion is added by studies showing that mitochondrial genomes in a few organisms, such as molluscs and nematodes, have large non-coding regions, questioning their efficiency (Boore, 1999).

3.4.3.3 GC Content, Gene Length and Codon Usage

Despite evidence here and in previous studies indicating that GC content and codon usage correlate significantly with overall gene expression levels (Coghlan and Wolfe, 2000, Marín *et al.*, 2003), these factors were not found to directly relate to levels of allelic expression imbalance. When comparisons were made between the structural factors of the allele with the lowest expression and the allele with the highest expression, no clear patterns were observed relating to expression. This rules out these structural factors as the direct causal mechanism behind the divergence in allele expression.

However, when analysing allele length, it was found that the shorter allele had a significantly lower level of expression. It remains unclear as to whether this difference in length is functionally important to levels of AEI, or if it is in fact a bias of sequencing technologies as mentioned previously. Shorter alleles produce shorter fragments and are therefore sequenced less often, with this difference not being sufficiently accounted for through RPKM normalisation (Bullard *et al.* 2010b).

Alleles of genes with AEI were found to have significantly larger variance in differences of structural factors than alleles with equal expression. These differences, especially the fold difference in allele length, had a positive and significant correlation with the difference in protein identity seen between alleles. Consequently, following on from the hypothesis that divergent allele expression is linked to differences in allele function, it can be suggested that although structural factors are not mechanistically causing the differences in allele expression, they are leading to differences in DNA sequences. These differences in DNA sequence are, in turn, leading to differences in function, and this is driving the difference in expression level.

Again, it should be noted from previous discussions that these differences in structural factors being larger in genes with AEI may actually be a consequence of bias in the sequencing technologies. The identification of these genes as having significant AEI may be false positive results as differences in structural factors such as length and GC content can lead to differences in

sequenceability (Bullard *et al.*, 2010b) and biases in normalisation (Zheng *et al.*, 2011), causing an over representation of the differences in expression.

3.4.3.4 Structural Factors Still to Be Investigated

Although this study has investigated a number of structural factors, there remain many more which could influence levels of allelic expression imbalance which have not been explored here.

Methylation patterns have been associated with allelic expression imbalance, especially imprinting, as described in section 1.6.1. It has been seen in humans that all but one imprinted gene exhibit allele-specific methylation (Brannan and Bartolomei, 1999). Genome-wide analysis of methylation patterns in *Candida albicans* revealed that methylation is centred upon 150 genes, with a general association with environmentally cued pathways (Mishra *et al.*, 2011). However, this study looked at *C. albicans* in a haploid sense and did not identify allele-specific methylation patterns. Unfortunately, identifying allele-specific methylation through the use of whole genome bisulfite sequencing was unpractical for use in this study and is associated with difficulties in pairing SNPs with methylcytosines as described in section 1.6.1, but is an area for further investigation. However, when investigating the 150 methylated genes identified by Mishra *et al.* (2011) it was observed that only six of these genes were also identified as having allelic expression imbalance in this study. This suggests that allele-specific methylation patterns are unlikely to be the sole causal factor in differential allele expression. This reflects upon the results found by a study of methylation in maize where it was shown that not all imprinted genes have methylation differences between alleles (Zhang *et al.*, 2011). Nevertheless, Mishra *et al.* (2011) also showed that methylation patterns varied significantly between distinct morphological forms. Condition specific methylation differences could account for such few methylated genes being identified with AEI here, and therefore the influence of this mechanism upon AEI still requires further examination.

Aspects of DNA packaging are also reported to contribute towards gene expression levels. Nucleosomes consists of a stretch of 147 bp of DNA helix which wrap and bend around a histone protein, with 10 – 50 bp of unwrapped

DNA separating individual nucleosomes (Richmond and Davey, 2003). Generally unbound, and therefore unoccupied, regions of DNA have greater access for DNA-binding proteins, and are associated with higher expression levels, for example centromeres are known to have low levels of gene expression and have a higher occupancy of nucleosomes than other stretches of DNA (Wyrick *et al.*, 1999). The nucleosome-DNA interaction model has been constructed in *S. cerevisiae* using a combination of isolated nucleosome-bound DNA fragments and computational modelling methods (Segal *et al.*, 2006). This study stated that around 50% of *in vivo* nucleosome organisation can be explained solely by sequence preferences. Different DNA sequences have different propensities to bend, and therefore wrap around histones. For instance, dinucleotide repeats of AA/TT/AT bend easily and bind to nucleosomes more readily. Therefore, it could be proposed that alleles which differ in sequence could have differing propensities towards nucleosomes, and therefore differing expression levels. To support this idea, it has recently been demonstrated in mice that alleles of an imprinted gene differ in their 3D chromosomal conformation; a compact structure relates to repression of the *HoxD* gene in the maternal line (Lonfat *et al.*, 2013). In this study, insertion of genes into different chromosomal locations near the *HoxD* locus resulted in varying levels of imprinting, demonstrating that these differing chromosomal conformations result in position dependent gene silencing. However, this theory of nucleosome occupancy and gene expression being dictated by DNA sequence and position produces interesting questions, for example, how do gene expression levels change under different conditions where DNA sequences will always remain the same?

The impact of differences in the 5'-UTR sequence upon protein expression levels has been previously investigated in yeast. Using a constructed library of over 2000 mutants which differ only in the 10 bp upstream of the start codon and all translate to the same YFP-tagged gene, the contribution of specific motifs in UTRs upon protein levels were identified (Dvir *et al.*, 2013). Motifs with the highest impact included the nucleotide at positions -3 to -1, with a purine at -3 leading to an increase in protein levels; mRNA secondary structure with stable secondary structures leading to a decrease in protein levels; and out-of-frame upstream AUGs which again decrease protein levels. It should be noted that

this study used only a single gene with a short 5'-UTR and the observations made may not be applicable across all genes, but results do imply that differences in UTR sequence effect expression, and therefore could impact upon AEI. Interestingly this study demonstrates that differences in up-stream motifs had a much larger impact upon protein levels (4.5-fold) than mRNA levels (2.9-fold). Therefore this may not have been detected within our results. When investigating AEI at both transcriptional and translational levels in *C. albicans,* a weak but significant correlation was seen between SNPs in the 60 bp around the start codon and the disparity in allelic expression, however this relationship was not deemed significant enough to fully account for the levels of AEI (Muzzey *et al.*, 2014). Longer UTRs have also been associated with highly regulated genes both in *C. albicans* (Tuch *et al.*, 2010b) and in *C. parapsilosis* (Guida *et al.*, 2011)*,* therefore differences in UTR lengths between alleles could influence allelic expression imbalance. Both of these studies identified UTR lengths from RNA sequencing data, and so this is an area that can be returned to for further investigation with our data. However, there is also contradictory evidence for the relationship between UTR length and expression in *C. albicans.* Since Sellam *et al.* (2010) demonstrated more stable transcripts have shorter 5'-UTRs (Sellam *et al.*, 2010), and therefore any analysis of UTR length and AEI would need to be mindful of this.

Differing RNA decay rates of alleles could also lead to differences in RNA abundance, giving the impression of differences in expression level. As RNA decay rates have been proposed to be regulated by sequence motifs, such as AU-Rich elements which modulate poly(A)-shortening rates (Vasudevan and Peltz, 2001), it is possible to infer that transcripts produced from different alleles could differ in decay rate, especially in the case of alleles which differ vastly in sequence. Therefore, decay rates could account for the AEI observed. It is possible to quantitatively measure RNA decay rates in yeast using various methods including in rich media with 1,10 phenanthroline and dot blotting (Santiago *et al.*, 1986), and with the use of strains carrying a temperature-sensitive mutation in RNA polymerase II (Wang *et al.*, 2002). However, using these methods to identify allele expression levels requires a system, such as PCR, specific to each allele. As seen in the development of a validation system,

obtaining specificity is challenging, and would need to be optimised before decay rates were observed.

Interestingly, RNA decay rates have also been linked to transcript lengths. In a study of 15 genes in *S. cerevisiae* it was shown that length has an inverse relationship with mRNA stability with longer mRNAs decaying faster (Santiago *et al.*, 1986). An increase in targets for endonucleolytic cuts in longer transcripts was suggested as a possible mechanism for this relationship. The results presented here showing that gene length has a negative correlation with overall expression support this hypothesis. Although alleles with lower expression levels were were in fact found to be significantly shorter than alleles with high expression, this change in length and therefore possible change in RNA stability could still impact upon AEI. However, a more recent genome-wide study, using DNA microarrays and a strain with a temperature-sensitive mutation in RNA polymerase II, found that there was no correlation between decay rates and transcript lengths. Nor was there a relationship seen between decay rates and codon bias, ribosome density or abundance (Wang *et al.*, 2002). Instead it was proposed that decay rates could be dictated by function, with groups of proteins seen in the same complexes or pathways having very similar decay rates. This included complexes such as histones, the 20S proteasome, ribosomal proteins and the trehalose phosphate synthase complex and pathways including energy metabolism enzymes and mating pheromone signal transduction pathway (Wang *et al.*, 2002). Therefore if the levels of AEI seen are attributed to differences in allele decay rates, this could be supportive of a functional difference between the alleles.

An idea moving away from structure of alleles is the age of the allele. In yeast, “younger” recently duplicated genes are thought to be less essential than older more conserved genes. These older genes have higher expression levels and appear to have more severe phenotypes when removed (Chen *et al.*, 2012). This leads to the idea that the allele with higher expression level is the older functional copy. This is supported by the evidence that duplicated genes have a higher probability of being functionally compensated for, with less lethal effects upon removal. Through use of microarrays, it was also revealed that knocking out the duplicate with higher expression levels had a more severe phenotypic

effect than loss of the duplicate with lower expression (Gu *et al.*, 2003). However, here alleles are being directly compared to duplicate genes, with the assumption that they have arisen by the same mechanism.

3.4.4 Validation of AEI

To validate the results produced from RNA sequencing, numerous methods were explored. In previous studies, allele-specific qPCR with TaqMan probes (Harries *et al.*, 2006, Tuch *et al.*, 2010a) and allele-specific restriction enzyme digestion (Zhang *et al.*, 2011) have been used to both identify and validate levels of AEI. Use of allele-specific qPCR with SYBR® green and use of protein tagged alleles were also explored for their efficacy in validation. However, all approaches attempted encountered similar problems with specificity against each allele and with sufficient efficiency to achieve reproducible results. Validation of the levels of AEI identified by RNA sequencing, therefore, remained unsuccessful.

An alternative method for validation of the results produced from RNA sequencing which was not explored is the use of the NanoString nCounter gene expression system (Geiss *et al.*, 2008). This method captures and counts individual mRNA transcripts through the use of capture and reporter probes, which complementarily bind to a transcript, and a colour-coded tag which produces the signal. High concordance was seen between the results from NanoString, microarrays and TaqMan qPCR. However, there is currently no evidence for use of this method to identify or validate AEI, and it is unclear if the probes are subject to cross-hybridisation of highly similar sequences such as alleles.

Measuring allele-specific protein expression was attempted using western blots and allele specific tags, but proved unsuccessful. Other protein expression measurement techniques is another area which could have been explored further. Recently, work has been published which has monitored allele-specific protein expression using liquid chromatography mass spectrometry. The study used a hybrid diploid strain of *S. cerevisiae* and *S. bayanus,* which was labelled with a heavy isotope, and two parental homozygous strains labelled with a light isotope. Comparisons were made between the ratios of variant peptides to

shared peptides in these strains to identify the protein allele-specific expression level (Khan *et al.*, 2012). This study found that on a proteome-wide average, both alleles are expressed at equal expression levels. However 589 proteins in replicate one and 426 in replicate two were identified with protein allele-specific expression. A modest correlation was also found between protein allele-specific methods and mRNA allele-specific measurements made in the same hybrid species, however, the cells were grown in different conditions for both of these experiments, questioning the comparability. Currently, this novel method is unsuitable for use with *Candida albicans* as homozygous parental strains are not available for comparison against a heterozygous offspring.

Measuring protein levels to validate RNA sequencing data relies upon the assumption that mRNA levels have a proportional relationship with protein levels. However evidence to contradict this assumption has been found. In *E. coli,* the protein and mRNA expression levels have been quantified simultaneously using an YFP protein fusion library, single molecule fluorescence microscopy and fluorescence *in situ* hybridisation. It was found that on a single cell level, protein and mRNA expression levels did not significantly correlate with each other, although this was only determined for 137 genes with high protein expression and not across the entire genome (Taniguchi *et al.*, 2010). Ribosome profiling has been recently developed to interrogate levels of protein abundance. This method involves isolation and sequencing of ribosome bound mRNA fragments (Ingolia *et al.*, 2009). Recently this technique has been used to assess the relationship between allele specific levels of transcription and translation, for example showing that small divergences in allele expression levels are buffered for and not present at the level of translation in F1 hybrids of *S. cerevisiae* and *S. paradoxus* (Artieri and Fraser, 2013, McManus *et al.*, 2014). However, assessment of allelic expression imbalance at the translational level in *C. albicans* has shown that allelic bias at the transcriptional and translational level showed similar magnitudes of bias but on a gene-by-gene basis, levels of AEI did not always predict levels of translational allelic bias. For example the gene *CHO2* was shown to have equal levels of allele expression at the transcriptional level but was seen to favour allele two at the level of translation (Muzzey *et al.*, 2014). Alternatively, other genes were seen to have a compensatory relationship

between transcriptional and translational bias with uneven levels of mRNA being corrected for by opposing uneven levels on translation (Muzzey *et al.*, 2014).

3.4.5 Conclusion

To conclude this chapter, genes with significant levels of AEI have been identified in *C. albicans* wild-type strain SC5314 using RNA sequencing. These genes were not themselves associated with specific biological functions. However, analysis of the structural factors associated with these alleles indicate significant differences in DNA sequence, which could lead to functional differences in proteins. Deciphering the true extent of the link between AEI and functional divergence will be the main focus of the following chapter.

Chapter 4: Investigating the Phenotypic Contribution of Allelic Expression Imbalance

4.1 Introduction

This chapter examines the functional consequences of allelic expression imbalance in the wild-type *Candida albicans* strain SC5314. The previous chapter used RNA sequencing to identify genes with significant divergence in allele expression level. Percentage protein identity comparisons indicated that the protein sequences of many of the differentially expressed alleles are also highly divergent. This has led to the hypothesis that alleles which differ in expression level may also have functionally distinct roles due to differences in protein sequence. To investigate this theory, heterozygous knockout mutants of each allele of a number of genes identified as having significant levels of AEI were constructed and phenotypically compared under a range of conditions.

4.1.1 Methods of Knockout Construction

Investigations of the phenotypic contributions of genes in *C. albicans* was initially problematic due to its diploid nature and lack of a conventional sexual cycle. The construction of an auxotrophic knockout strain in 1993 by Fonzi and Irwin opened up the possibilities of genetic manipulation in *Candida albicans.* Through targeted mutagenesis and homologous recombination, the strain CAI4 was constructed, which lacks both alleles of the *URA3* gene (Fonzi and Irwin, 1993). This strain, and subsequent other auxotrophic marker deletion strains, have led to the development of various methods to “knockout” genes of interest via homologous recombination, constructing both heterozygous and homozygous strains. These methods involve targeted replacement of the gene of interest with an auxotrophic marker, allowing for selection of successful transformants. The Ura-blaster method (Figure 4.1a), as developed in 1993, uses a targeted cassette containing the *URA3* gene and

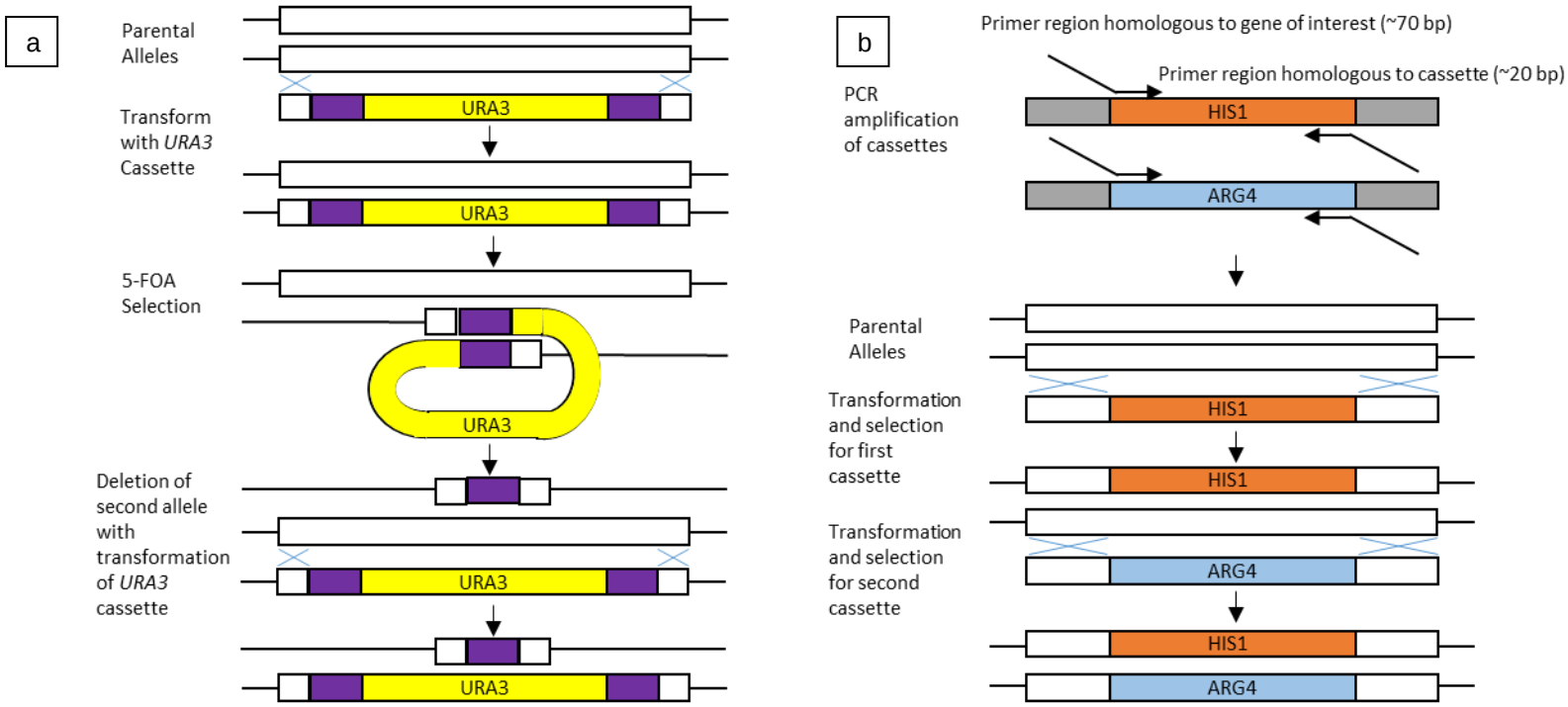


Figure 4.1 Methods of constructing gene knockout strains. a) The Ura-blaster method involves use of a *URA3* cassette (yellow) flanked by repeats (purple). Transformants that are URA+ are used to select isolates that carry the Ura-blaster cassette. Isolates that have lost the *URA3* sequences through recombination between the repeats are counter-selected for using 5-FOA selection. This followed by a second transformation with the *URA3* cassette to remove the second allele. b) PCR-mediated transformation use a cassette containing a selectable marker flanked by 70 nucleotides of homologous sequence. Amplification of the cassette uses 5' and 3' primers that include short regions (~20 nucleotides) of sequence homologous to the marker vector and ~70 nucleotides of sequence homologous to the sites of insertion. The cassette is then used to transform *Candida albicans* strains. Sequential transformation with amplified fragments that carry the same flanking sequences and two different selectable markers is required to generate a homozygous deletion strain. Figure adapted from (Berman and Sudbery, 2002)

5-fluoro-orotic acid (5-FOA) selection to recycle and reuse the cassette to produce homozygous mutants (Fonzi and Irwin, 1993). Later PCR amplification based methods (Figure 4.1b) involve amplification of a cassettes targeted to the gene of interest and sequential gene knockouts with different auxotrophic markers such as *HIS1* and *ARG4* (Noble and Johnson, 2005). Other knockout approaches which are available for use in *C. albicans* include the flipper cassette system and the Cre-*loxP* system. The flipper system uses a cassette containing a selectable marker and a *C. albicans* adapted *FLP* gene. This allows for sequential deletion of the target alleles, in a similar way to the PCR amplification methods, but by the same cassette. After the removal of the first allele, the cassette is "flipped out" by the *FLP* gene and then reused for removal of the second allele. This system also allows for easy reintegration of sequences for complementation at the native locus. Flipper cassettes have been developed with various selectable markers including auxotrophic markers (Morschhäuser *et al.*, 1999) and the antibiotic resistance cassette nourseothricin (discussed in more detail below) (Reuß *et al.*, 2004). The Cre-*loxP* system also allows for recycling of the selection cassette. Here the selection cassette is flanked by *loxP* elements which allows for removal via site-specific recombination catalysed by the Cre recombinase enzyme (Dennison *et al.*, 2005). An alternative approach to knocking out genes is knocking down genes, and RNAi is an approach that has been developed to achieve this. Expression of the target gene is reduced using small RNAs complementary to the target gene mRNA. Double stranded RNA then forms via complementation and is targeted for degradation via small interfering RNA sequences and the RNA-induced silencing complex (Jinek and Doudna, 2009). This system has been used in *C. albicans* to knock down the *EFG1* gene involved in the morphological switch to hyphae (Moazeni *et al.*, 2012).

Phenotypic screening of knockout mutants allows for elucidation of gene function. Large-scale screening of gene function using several hundred homozygous and heterozygous mutants in *C. albicans* has been undertaken, identifying genes involved in infection related processes such as antifungal drug resistance (Xu *et al.*, 2007) and morphological switching (Noble *et al.*, 2010).

As detailed in section 1.4.1, phenotypic screening of heterozygous knockout strains has indicated that certain genes have divergence in allele function, lending support to our hypothesis that alleles with AEI may differ in function. Examples include, not exclusively, the adenine gene *ADE2* which has one functional and one non-functional allele (Tsang *et al.*, 1999), the histidine gene *HIS4* where a single SNP renders allele 1 inactive (Gómez-Raja *et al.*, 2008), the *ALS9* gene which has highly divergent allele sequences with suggestions of functional differences (Zhao *et al.*, 2007), the *TAC1* transcription factor which have alleles that differ in function, impacting upon drug resistance (Coste *et al.*, 2006), and the *CDR* genes which have been shown to have alleles with differing pumping capabilities, again effecting susceptibility to antifungal drug treatments (Holmes *et al.*, 2006). Of the genes identified with AEI in chapter three, Table 4.1 shows that 31 have already been shown to have a phenotype of the heterozygous knockout mutant (in any strain, not exclusively the wild-type strain SC5314; information obtained from the *Candida* genome database http://candidagenome.org/ (Inglis *et al.*, 2012)), although these studies often do not compare the functionality of the two alleles.

Traditionally, auxotrophic markers such as the *URA3* gene described above have been used for marker selection during the transformation process. However, evidence has shown that altered expression of these genes, especially *URA3,* can cause phenotypic impacts themselves, particularly with regards to pathogenicity (Kirsch and Whitney, 1991, Brand *et al.*, 2004). This, therefore, has put phenotypic predictions obtained from these mutants into question. To avoid this problem, an alternative method has been adopted here using the antibiotic resistance cassette nourseothricin. This marker was developed for use in *Candida albicans* transformations and has been shown to not effect growth in the yeast or hyphal form, unlike use of the *URA3* marker (Shen *et al.*, 2005). Other selectable markers available for use in *Candida albicans* which avoid auxotrophic markers include a gene conferring resistance to the antibiotic hygromycin B (Basso *et al.*, 2010) and a strain lacking the *MET15* gene which grows as brown colonies on media containing lead, allowing for replacement of a target gene with a *MET15* cassette causing white colony growth and simple colour selection for transformants (Viaene *et al.*, 2000).

Table 4.1 Genes with AEI shown to have a heterozygous null phenotype. Data taken from the *Candida* genome database http://candidagenome.org/.

ORF name	Gene Name	Strain	Phenotype
orf19.1047	*ERB1*	SC5314	Resistance to 5-fluorouracil, tubercidin and flucytosine decreased
		CAI-4	Virulence decreased (mouse model)
orf19.1246		BWP17	Resistance to clotrimazole decreased
orf19.1357	*FCY21*	SC5314	Resistance to 5-fluorouracil and flucytosine increased
orf19.1440		BWP17	Invasive growth decreased
orf19.1557		BWP17	Invasive growth decreased
orf19.1736		BWP17	Invasive growth decreased
orf19.1949	*VPS1*	CAI-8	Resistance to fluconazole decreased. Filamentous growth decreased
orf19.2014	*BCY1*	CAI-4	Filamentous growth abnormal
orf19.220	*PIR1*		Resistance to Congo red and calcofluor white decreased. Vegetative growth decreased.
orf19.2268	*RCK2*	BWP17	Invasive growth decreased
orf19.2521		BWP17	Invasive growth decreased
orf19.2555	*URA5*	CAI-4	Virulence absent (mouse model)
orf19.2743		BWP17	Invasive growth decreased
orf19.2823	*RFG1*	CAI-4	Filamentous growth increased. Virulence decreased (mouse model).
orf19.3077	*VID21*	BWP17	Invasive growth decreased
orf19.3161		CAI-4	Virulence absent (mouse model)
orf19.3188	*TAC1*		Resistance to fluconazole and terbinafine increased
orf19.3353		BWP17	Invasive growth decreased
orf19.3526	*ITR1*	BWP17	Invasive growth decreased
orf19.3969	*SLF2*	SC5314	Invasive growth decreased. Colony shape abnormal. Germ tube formation decreased.
orf19.4332		BWP17	Invasive growth decreased
orf19.4737	*TPO3*	SC5314	Resistance to 5-fluorouracil, tubercidin and flucytosine decreased
		BWP17	Invasive growth decreased
orf19.5104	*LTP1*	BWP17	Invasive growth decreased
orf19.5574		BWP17	Invasive growth decreased
orf19.5623	*ARP4*	SC5314	Resistance to virgineone decreased
orf19.5741	*ALS1*	CAI-4	Hyphal growth decreased
orf19.5768	*SNF4*	CAI-4	Filamentous growth decreased
orf19.6202	*RBT4*	SC5314	Resistance to ergosterol analogs decreased
		CAI-4	Virulence decreased (mouse model)
orf19.6630		BWP17	Invasive growth decreased
orf19.6854	*ATP1*	BWP17	Invasive growth decreased
orf19.801	*TBF1*	CAI-4	Virulence absent (mouse model)

4.1.2 Aims of this Chapter

This chapter has three main aims to investigate the functional consequence of AEI:

1. Identification of a set of target genes for heterozygous knockout construction.
2. Construction and validation of heterozygous knockout strains.
3. Phenotypic screening of heterozygous knockout strains under a range of general and gene specific assays.

4.2 Materials and Methods

4.2.1 Identification of Target Genes

Genes to be used for heterozygous knockout construction were selected based upon three criteria: greater than 2x fold difference in RPKM value of alleles, more than 20 counts for the allele with the lowest expression level (except for the cases of monoallelic expression), and association with GO terms that are related to virulence and are easily screened to assess phenotypic impact – pathogenesis, morphology, cell cycle, metabolism and drug resistance. GO terms for each gene were identified using the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012).

4.2.2 Heterozygous Knockout Mutant Construction

For a full list of all heterozygous knockout mutants constructed and used in this study, please refer to Table 2.1. Heterozygous deletions were carried out as described in sections 2.8 to 2.11.

4.2.3 Phenotypic Screening

For methodology of general phenotypic assays used both here and in chapter five, see section 2.14. Below lists the phenotypic screens used in just this chapter.

4.2.3.1 Biofilm Production

The ability of a strain to form a biofilm was measured using a 96-well plate assay. Cells were grown overnight in 10 ml of YNB + 50 mM glucose (0.67% (w/v) yeast nitrogen base, 0.9% (w/v) glucose). This media has previously been shown to promote growth of biofilms (Hawser and Douglas, 1994). Cells were

pelleted at 4000 rpm for five minutes, washed twice in 20 ml of PBS, and resuspended at a concentration of 1 x 10^7 cells/ml using a haemocytometer. 100 µl of cells were added to a well of the 96-well plate. Individual strains were inoculated in a minimum of technical triplicate. Cells were left to adhere to the plate at 37 °C for 90 minutes. Non-adherent cells were washed off twice with 150 µl of PBS. 100 µl of YNB + 50 mM glucose was then added and plates were incubated at 37 °C for 24 or 48 hours.

Biofilm formation was measured using an XTT assay, where the solution turns increasingly red with increasing metabolic activity of mitochondrial dehydrogenase (Chandra *et al.*, 2012). Fresh XTT/menadione solutions were prepared on the day of the assay (10 ml of 0.5 mg/ml XTT and 1 µl of 10 mM menadione). Spent media was removed from the wells leaving the biofilm behind. Unbound cells were removed by washing twice with 150 µl of PBS. 100 µl of the XTT/menadione solution was added to each well, followed by incubation in the dark at 37 °C for two hours. The optical density at 490 nm was then measured using a spectrophotometer (Bio-Rad iMark Microplate Reader). Measurements were statistically compared using an ANOVA followed by a Dunnett's *post-hoc* test.

4.2.3.2 Antifungal Resistance during Biofilm Formation

The protocol for measuring antifungal resistance of biofilms was based on a method by Nett *et al.* (2011). To summarise, cells were grown overnight in 5 ml of YPD at 30 °C, 180 rpm. Cell concentrations were estimated using a haemocytometer and diluted to a concentration of 1 x 10^6 cells/ml in RPMI 1640 medium (1.04% (w/v) RPMI 1640, 3.45% (w/v) MOPS, 1.8% (w/v) glucose, pH 7.0). 100 µl of cells were then added to the well of a 96-well plate and incubated at 37 °C for six hours. Wells were then washed twice with 150 µl of PBS to remove unbound cells. 100 µl of fresh RPMI 1640 containing the antifungal drug of interest was added to the well and allowed to incubate at 37 °C for 24 hours. Spent media was then removed and 100 µl of fresh RPMI containing the antifungal drug of interest was applied. The plate was then incubated at 37 °C for a further 24 hours and the XTT assay described in section 4.2.3.1 was used to measure metabolic activity.

Antifungal drugs and concentrations were based on those described in section 2.14.3.

4.2.3.3 Cell Cycle Analysis

Flow cytometry was used to assess if a heterozygous knockout strain has an altered cell cycle distribution. For cell cycle analysis to be carried out, all cells need to be in the same stage of their cell cycle. To achieve this synchronisation, a technique involving starvation was used. This method was chosen over alternatives, as other forms of cell-cycle block have been associated with a switch to hyphae-like cell growth (Berman, 2006). Cells were grown overnight in 10 ml of YPD at 30 °C, 180 rpm. 200 µl of cells were transferred to 5 ml of 1% (w/v) yeast extract and 2% (w/v) bacto-peptone and incubated overnight at 30 °C, 180 rpm. To release all cells simultaneously, cells were spun at 2000 rpm for five minutes and resuspended in 5 ml of YPD. This was followed by incubation at 30 °C, 180 rpm. 500 µl of cells were then taken every 15 minutes for a total of two hours.

Cell cycle distribution of the wild-type strain SC5314 was checked using flow cytometry prior to analysis of heterozygous knockout mutants (see Appendix I Figure II). This was achieved using an adapted method designed for use with fission yeast (Sabatinos and Forsburg, 2009) as follows: at each 15 minute sampling, cells were resuspended in 1 ml of ice-cold 70% ethanol whilst vortexing, and cooled for one hour at 4 °C. From this suspension, 300 µl was removed and added to 3 ml of 50 mM sodium citrate. The solution was mixed and cells were pelleted at 2000 rpm for five minutes. To remove contaminating RNA, which may affect analysis of DNA content, the cells were resuspended in 500 µl of 50 mM sodium citrate with 0.1 mg/ml RNase A and incubated at 37 °C for two hours. Finally to stain the DNA, 500 µl of 50 mM sodium citrate containing 8 µg/ml propidium iodide was added and mixed thoroughly. Samples were stored in the dark at 4 °C until analysis. Samples were analysed using a BD FACSARIA II (Becton Dickinson, San Jose, USA) set with detector FSC E00 at Gain 3 and detector FL2-A Voltage at 890 and Gain 2.

4.2.3.4 Vacuole Staining with FUN-1 Solution

To monitor vacuole size and formation a FUN-1 stain (Molecular Probes, Invitrogen) was used. This assay is typically used to assay whether cells are alive or dead, but as a lipophilic stain it can also be used to visualise vacuoles. The appropriate strains were grown overnight in 5 ml of YPD at 30 °C, 180 rpm. 200 µl were removed, pelleted at 13000 rpm for five minutes and resuspended in 1 ml of GH solution (2% (w/v) glucose, 10 mM sodium-HEPES, pH 7.2). 100 µl of cells were combined with 100 µl of 50 mM FUN-1 reagent in GH solution. The mixture was incubated for 30 minutes in the dark at 30 °C before visualisation of cells with a fluorescence microscope (Leica DFC300 Fx). FUN-1 reagent is excited at around 470 nm and emission is detected at between 500 and 700 nm.

4.2.3.5 Lipase Secretion

To measure the amount of lipase secreted from cells, a colony halo protocol was used as previously described by Fu *et al.* (1997). Cells were grown overnight in 5 ml of YPD at 30 °C, 180 rpm. Cells concentrations were adjusted to 1 x 10^7 cells/ml using a haemocytometer. 5 µl of cells were spotted onto Egg Yolk media plates (6.5% (w/v) Sabouraud dextrose agar, 5.85% (w/v) sodium chloride, 0.06% (w/v) calcium chloride, 10% (v/v) egg yolk (spun at 500 rpm for 10 minutes, supernatant used)) and incubated at 30 °C. As lipase is secreted, the cloudy media becomes clear and produces a halo around the colony. The diameter of this halo was measured every 24 hours for a total of 5 days and statistically compared using an ANOVA followed by a Dunnett's *post-hoc* test.

4.2.4 Cloning of Genes

Cloning of *RCK2* to determine allele sequences was carried out as described in section 2.8.

4.3 Results

4.3.1 Identification of Target Genes for Phenotypic Analysis and Construction of Heterozygous Knockout Mutants

From the 233 genes identified as having significant levels of AEI in chapter 3, genes to be used for heterozygous knockout construction were selected based upon three criteria: greater than 2x fold difference in RPKM value of alleles,

more than 20 counts for the allele with the lowest expression level (except for the cases of monoallelic expression), and association with GO terms that are related to virulence and are easily screened to assess phenotypic impact – pathogenesis, morphology, cell cycle, metabolism and drug resistance. These strict criteria produced too small a list of genes to work with at just two genes, *CDC6* and *RBT4*. Therefore genes matching with 2 out of 3 criteria were also included.

This filter identified 11 possible gene targets for use in heterozygous knockout construction. Knocking out the gene *RFG1* would result in interference with the neighbouring open reading frame, and therefore, this gene was removed from the list of gene targets, leaving 10 genes in total (Table 4.2). Fold differences in the RPKM levels of each gene are graphically represented in Figure 4.2. Of the 10 genes selected, heterozygous knockout mutants were successfully constructed for both alleles of six genes, *CDC6, ERB1, RBT4, RCK2, SMI1* and *VPS1.* Validation of insertion at the correct genomic location was carried out using PCR as described in sections 2.9 and 2.10 (Figure 4.3). Southern blotting was used to further validate insertion in the correct position and ensure that only one copy of the nourseothricin cassette had been inserted, as described in section 2.15 (Figure 4.4). Southern blotting was carried out for all strains except for *RCK2* knockouts, as sequencing only identified knockouts of allele one. Due to the inability to produce knockouts of both alleles, there are also no phenotypic screens for *RCK2.* This is further discussed in section 4.3.3.

To show that the nourseothricin cassette itself was not causing phenotypic differences, a control strain containing nourseothricin at the *RPS1* locus was constructed as described in section 2.14.1. Herein, this strain is named SC12. This location was chosen based on evidence that replacement has no phenotypic contributions (Murad *et al.*, 2000). Validation of the strain was carried out using PCR (Figure 4.5).

Table 4.2 Target Genes for Heterozygous Knockout Construction

Gene Name	Allele 1	Allele 2	Allele 1 Tags	Allele 2 Tags	Allele 1 RPKM (3 d.p.)	Allele 2 RPKM (3 d.p.)	Fold Difference in RPKM (3 d.p.)	P Value
	orf19.2051	orf19.9599	737	1908	15.840	46.483	2.935	3.55×10^{-12}
	orf19.3556	orf19.11040	94	0	5.863	0.000	-	7.97×10^{-6}
	orf19.4516	orf19.11991	175	19	21.020	2.477	8.485	3.38×10^{-12}
	orf19.5095	orf19.12561	468	1	21.456	0.044	491.344	1.79×10^{-19}
CDC6	orf19.5242	orf19.12707	49	199	9.450	26.730	2.830	1.92×10^{-7}
ERB1	orf19.1047	orf19.8649	80	2	6.177	0.148	41.739	4.01×10^{-6}
RBT4	orf19.6202	orf19.13583	23	312	3.713	50.963	13.727	2.60×10^{-34}
RCK2	orf19.2268	orf19.9808	139	1192	14.091	121.551	8.626	4.13×10^{-68}
SMI1	orf19.5058	orf19.12525	216	0	22.524	0.000	-	1.18×10^{-20}
VPS1	orf19.1949	orf19.9505	1723	633	147.921	55.058	2.687	1.39×10^{-28}

- Indicates that expression levels for one allele were below the detectable limit and therefore the fold difference could not be calculated.

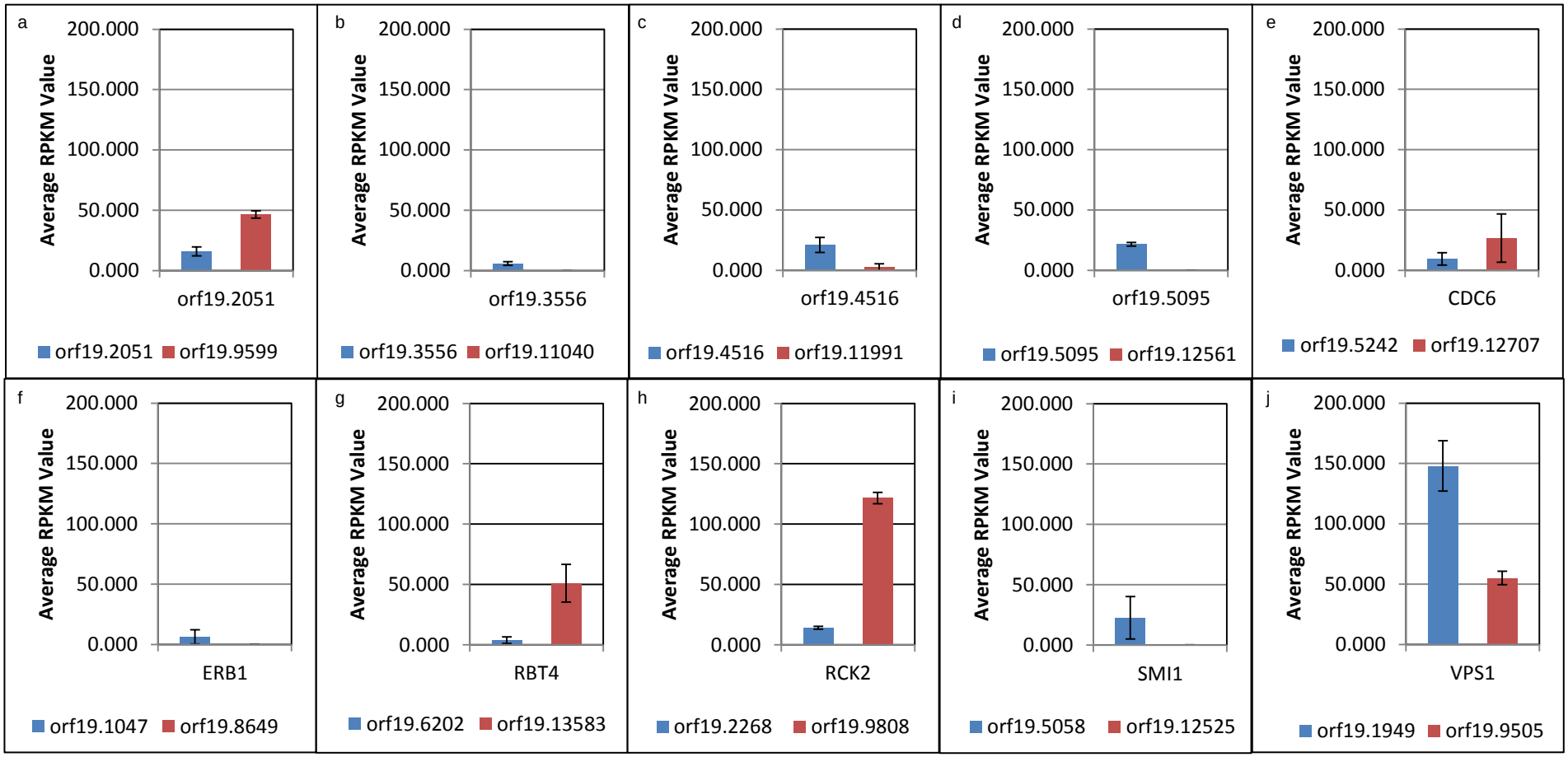


Figure 4.2 Average differences in allele expression of target genes. Errors bars ± one standard deviation. Blue shows “allele one” and red shows “allele two”. a) orf19.2051, b) orf19.3556, c) orf19.4516, d) orf19.5095, e) *CDC6,* f) *ERB1,* g) *RBT4,* h) *RCK2,* i) *SMI1* and j) *VPS1.*

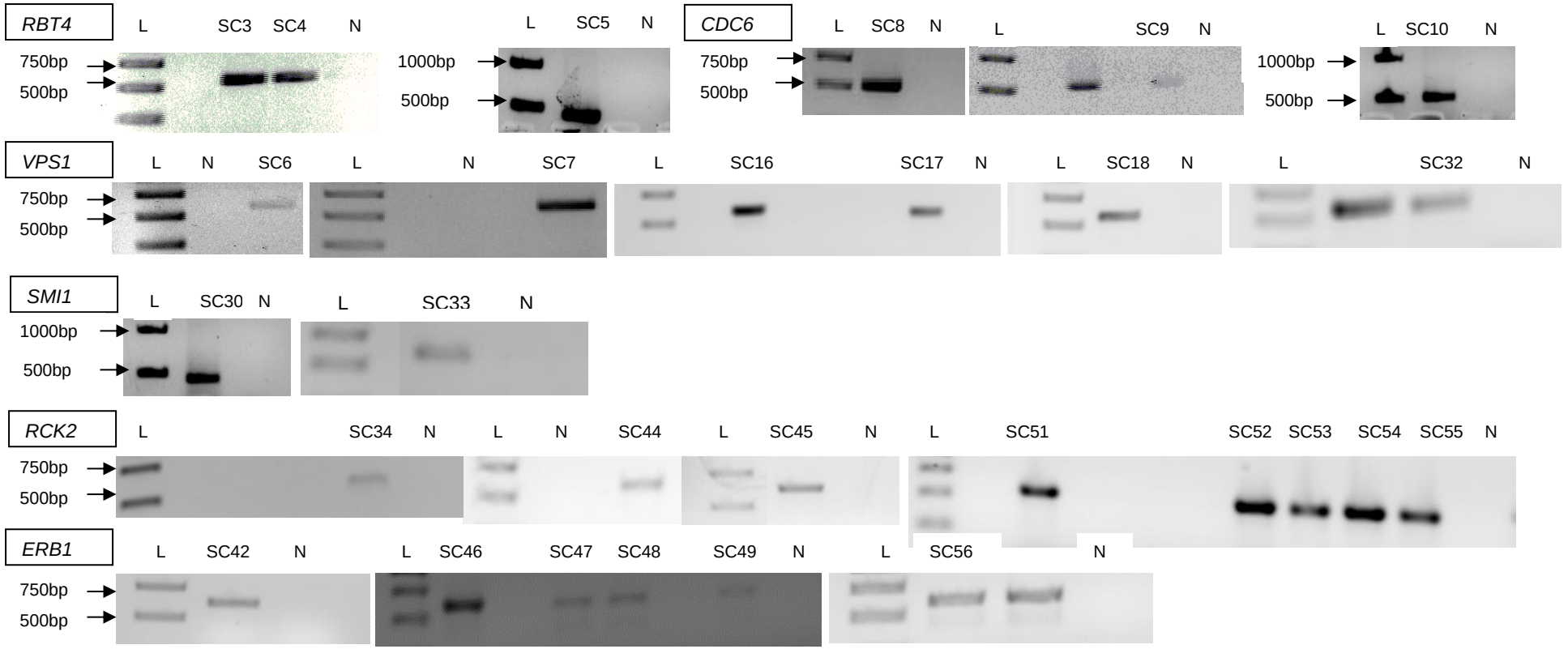


Figure 4.3 PCR validations of heterozygous knockout mutants. L = 1 kb DNA ladder (Fermentas, UK). N = negative control containing no DNA. SC3, SC4 and SC5 = *RBT4* knockouts with a band expected at 557 bp. SC6, SC7, SC16, SC17, SC18 and SC32 = *VPS1* knockouts with a band expected at 590 bp. SC8, SC9 and SC10 = *CDC6* knockouts with a band expected at 525 bp. SC30 and SC33 = *SMI1* knockouts with a band expected at 607 bp. SC34, SC44, SC45, SC51, SC52, SC55, SC54 and SC55 = *RCK2* knockouts with a band expected at 581 bp. SC42, SC46, SC47, SC48, SC49 and SC56 = *ERB1* knockouts with a band expected at 627 bp.

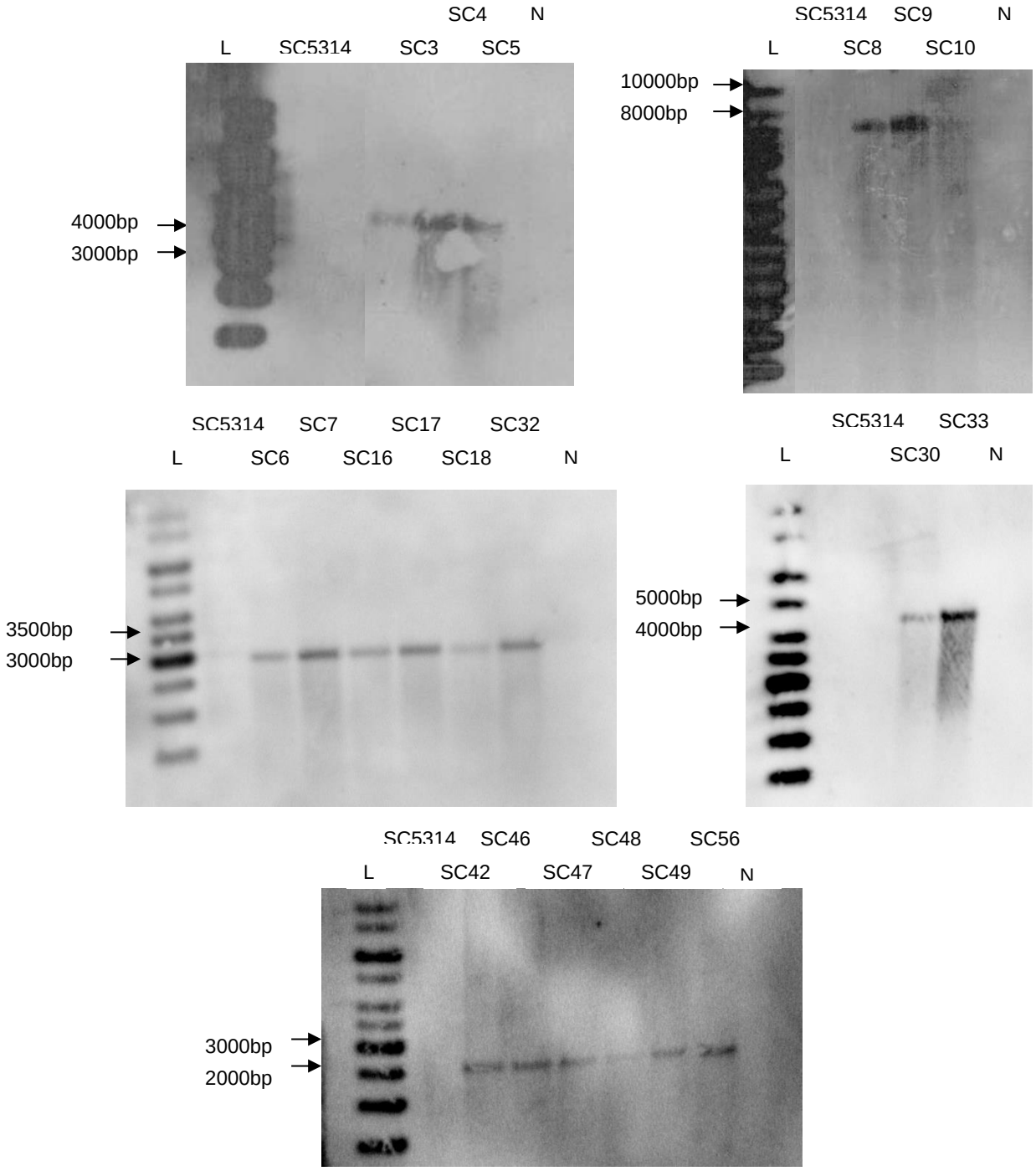


Figure 4.4 Southern blotting validations of heterozygous knockout mutants using a hybridisation probe against the *NAT1* sequence. L = 1 kb DNA ladder (Fermentas, UK). SC5314 = negative control containing untransformed wild-type DNA. N = negative control containing all reaction components except DNA. SC3, SC4 and SC5 = *RBT4* heterozygous knockouts showing a single band at 3635 bp. SC8, SC9 and SC10 = *CDC6* heterozygous knockouts showing a single band around 7000 bp, this is higher than expected but may be due to errors in the reference sequence used to design the digests. SC6, SC7, SC16, SC17, SC18 and SC32 = *VPS1* heterozygous knockouts with a single band at 2951 bp. SC30 and SC33 = *SMI1* heterozygous knockouts with a single band at 4588 bp. SC42, SC46, SC47, SC48, SC49 and SC56 = *ERB1* heterozygous knockouts showing a single band at 2573 bp.

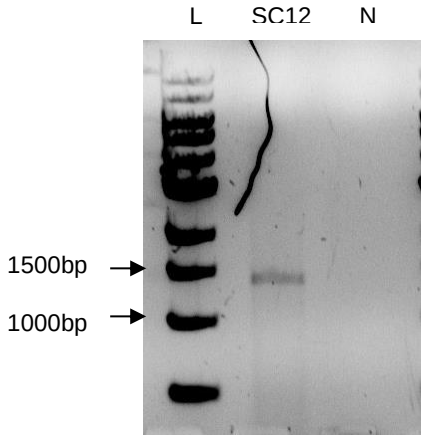


Figure 4.5 PCR validation of control strain SC12 with *NAT1* cassette integrated at *RPS1*. L = 1 kb DNA ladder (Fermentas, UK). N = negative control containing no DNA. SC12 shows a band at approximately 1500 bp as expected.

4.3.2 Phenotypic Screening of Heterozygous Knockout Mutants

Five standard phenotypic tests, which are related to the infection process, were undertaken on the heterozygous knockout mutants of all genes to elucidate any general defects in function. Phenotypic assays included growth rate at 30 °C and 37 °C (for methods see section 2.14.2), induction of hyphae with foetal calf serum (for methods see section 2.14.5), an adhesion assay using buccal epithelial cells (for methods see section 2.14.6), a virulence assay using a *Galleria mellonella* infection model (for methods see section 2.14.7), and resistance to the antifungal compounds fluconazole, 5-flucytosine and amphotericin B (for methods see section 2.14.3). These assays were accompanied by phenotypic tests specific for the gene in question based upon the findings of previous published work.

4.3.2.1 *CDC6* Phenotypic Screening

CDC6 forms part of the DNA pre-replication subunit and is expressed during the M/G1 phase of the cell cycle (Cote *et al.*, 2009). Work in *S. cerevisiae* has shown that *CDC6* is essential for DNA replication, with a role in origin firing and establishment of the pre-replication complex (Cocker *et al.*, 1996). Deletion leads to an accumulation of cells in S phase (Yu *et al.*, 2006) due to a lack of initiation of replication. Over-expression has demonstrated a block in M phase suggesting that *CDC6* also has a checkpoint role to ensure completion of the S phase before progression in the cell cycle (Bueno and Russell, 1992). Therefore, based on these previous studies, the heterozygous knockout

mutants of *CDC6* were analysed for any cell cycle malfunctions using flow cytometry as described in section 4.2.3.3.

Figure 4.6k shows that all three mutants, two knockouts of allele 1 and one knockout of allele 2, have a shift in their profile when compared to the wild-type strain SC5314 and the *NAT1* control strain SC12. There are a number of interesting points that can be made from this observation. All strains have two peaks, one representing a DNA content of 2n and one representing a DNA content of 4n, indicative of no block in the cell cycle at the S phase as is observed with *S. cerevisiae CDC6* knockout. However the shift in profiles suggests that all three *CDC6* heterozygous knockout mutants may be tetraploid, and in fact the peaks represent a DNA content of 4n and 8n, with cell separation defects being unlikely as appearance under the microscope is normal. Comparative genome hybridisation is an assay that could use in the future to confirm the ploidy of all chromosomes. Despite this phenotypic variation, the alleles themselves do not indicate a difference in function.

Mutants of *CDC6* in *S. cerevisiae* have also shown abnormal cellular morphology (Hartwell *et al.*, 1973) and an abnormal growth rate at 37 °C (Detweiler and Li, 1997). These phenotypes, however, have been shown to be unaffected in the *C. albicans CDC6* heterozygous knockout mutants with growth rates comparable to the wild-type strain SC5314 at both 30 °C and 37 °C (see Table 4.3 and Figures 4.6a and 4.6b), with only a minimal statistically significant increase in end-point optical density for SC9 at 30 °C, one isolate lacking allele one. However, this difference is also observed for the control strain SC12 containing *NAT1* at the *RPS1* locus, suggesting that the difference is not due to the loss of the allele.

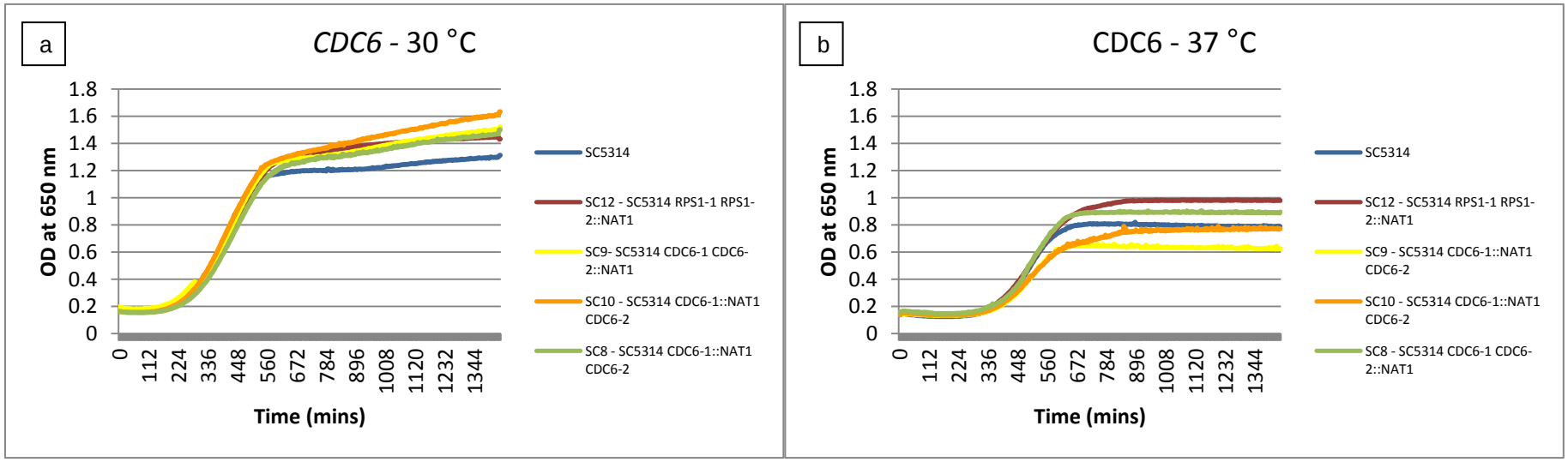
Other phenotypic assays also indicated no differences between the heterozygous knockout strains and the wild-type strain SC5314. All strains adhered to buccal epithelial cells (two sample t-test, d.f. = 2, $p > 0.05$; Figure 4.6c-e), all strains were able to switch to the hyphal growth form (Figure 4.6f), and all strains had comparable virulence using a *Galleria mellonella* infection model (Kaplan-Meier test, d.f. = 4, $p > 0.05$),

Table 4.3 Average generation times, times to maximum inflection and end-point optical densities of *CDC6* heterozygous knockout mutants at 30 °C and 37 °C (± one standard deviation)

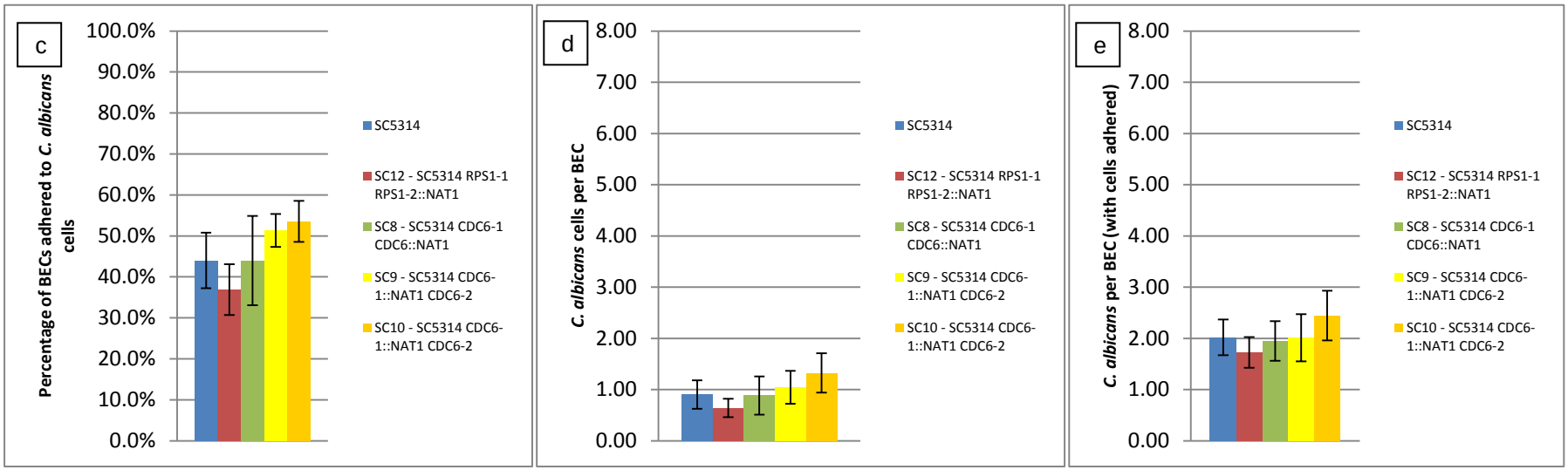
Growth Curve (from Figure 4.6)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	117.23 ± 36.54	306.69 ± 28.29	1.22 ± 0.17
	SC12	124.92 ± 59.84	326.20 ± 34.27	1.39* ± 0.20
Allele 1 knockout	SC9	106.63 ± 14.98	310.19 ± 31.17	1.45* ± 0.15
	SC10	114.53 ± 20.35	332.50 ± 29.75	1.34 ± 0.20
Allele 2 knockout	SC8	112.31 ± 15.55	321.30 ± 37.09	1.37 ± 0.17
b) 37 ° C	SC5314	120.53 ± 29.04	448.58 ± 56.94	0.80 ± 0.23
	SC12	135.60 ± 28.20	449.75 ± 91.35	0.98 ± 0.19
Allele 1 knockout	SC9	127.23 ± 23.74	465.85 ± 79.12	0.76 ± 0.46
	SC10	144.61 ± 36.91	455.70 ± 61.44	0.64 ± 0.30
Allele 2 knockout	SC8	141.81 ± 35.88	429.92 ± 51.80	0.89 ± 0.23

* Significantly different measurements from SC5314, identified by ANOVA followed by *post-hoc* analysis using a Dunnett's test, at $p < 0.05$, are annotated with an asterisk.

Figure 4.6 Phenotypic assays of *CDC6* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC9 and SC10 = knockout of “allele one” (yellow) and SC8 = knockout of “allele two” (green).

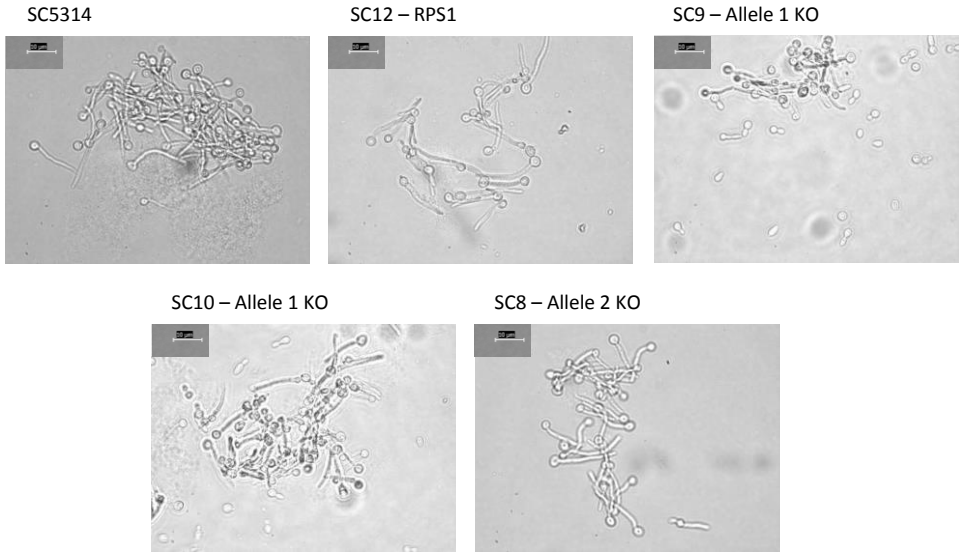


a) Growth rate at 30 °C. b) Growth rate at 37 °C.



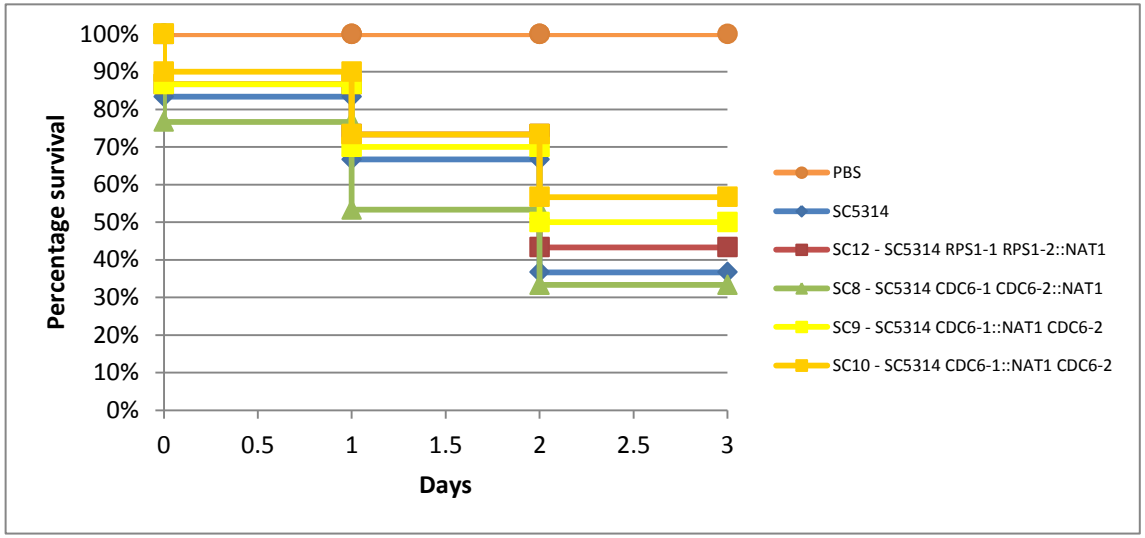
c) Percentage of BECs adhered to *C. albicans* cells. d) Number of *C. albicans* cells per BEC. e) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

f



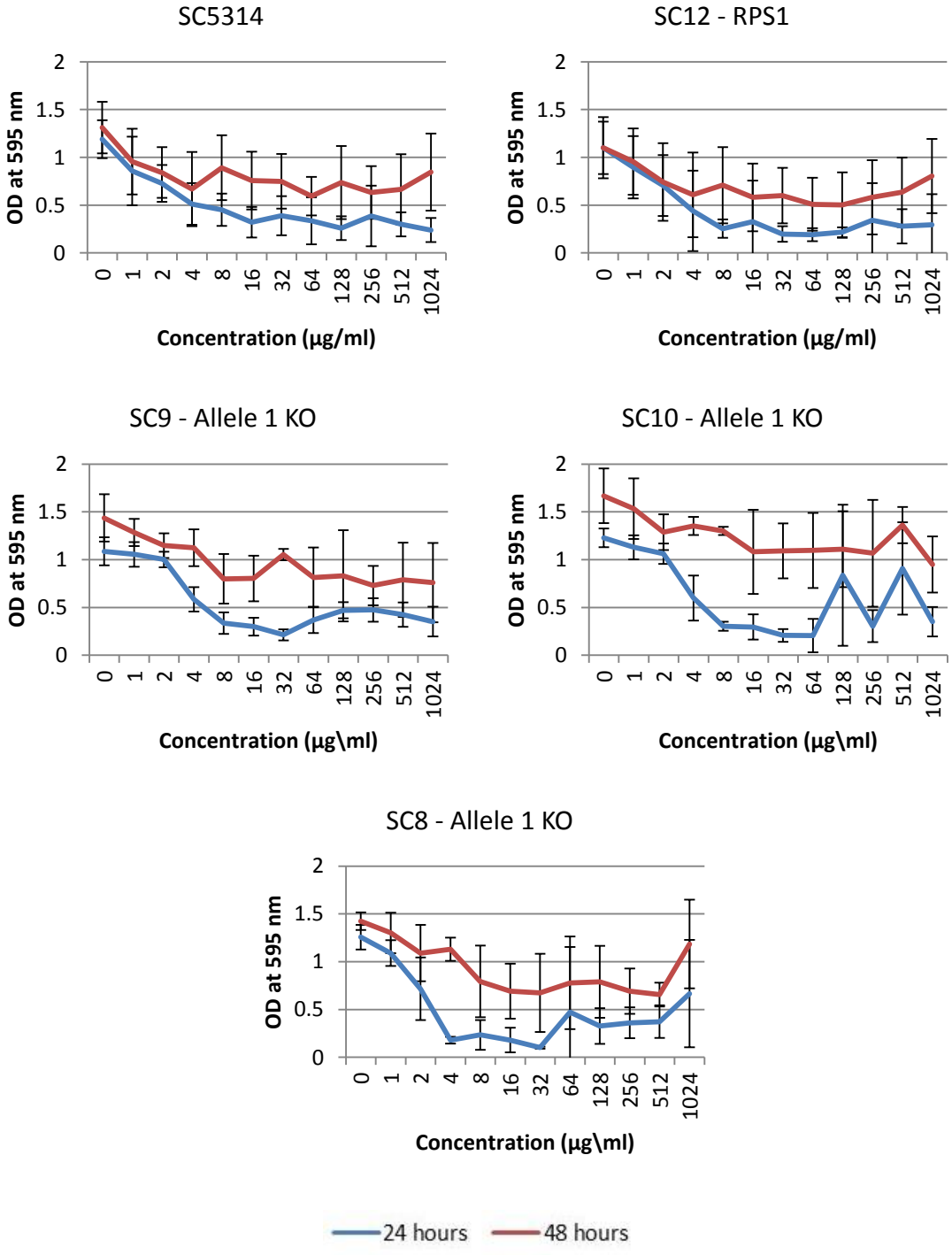
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

g

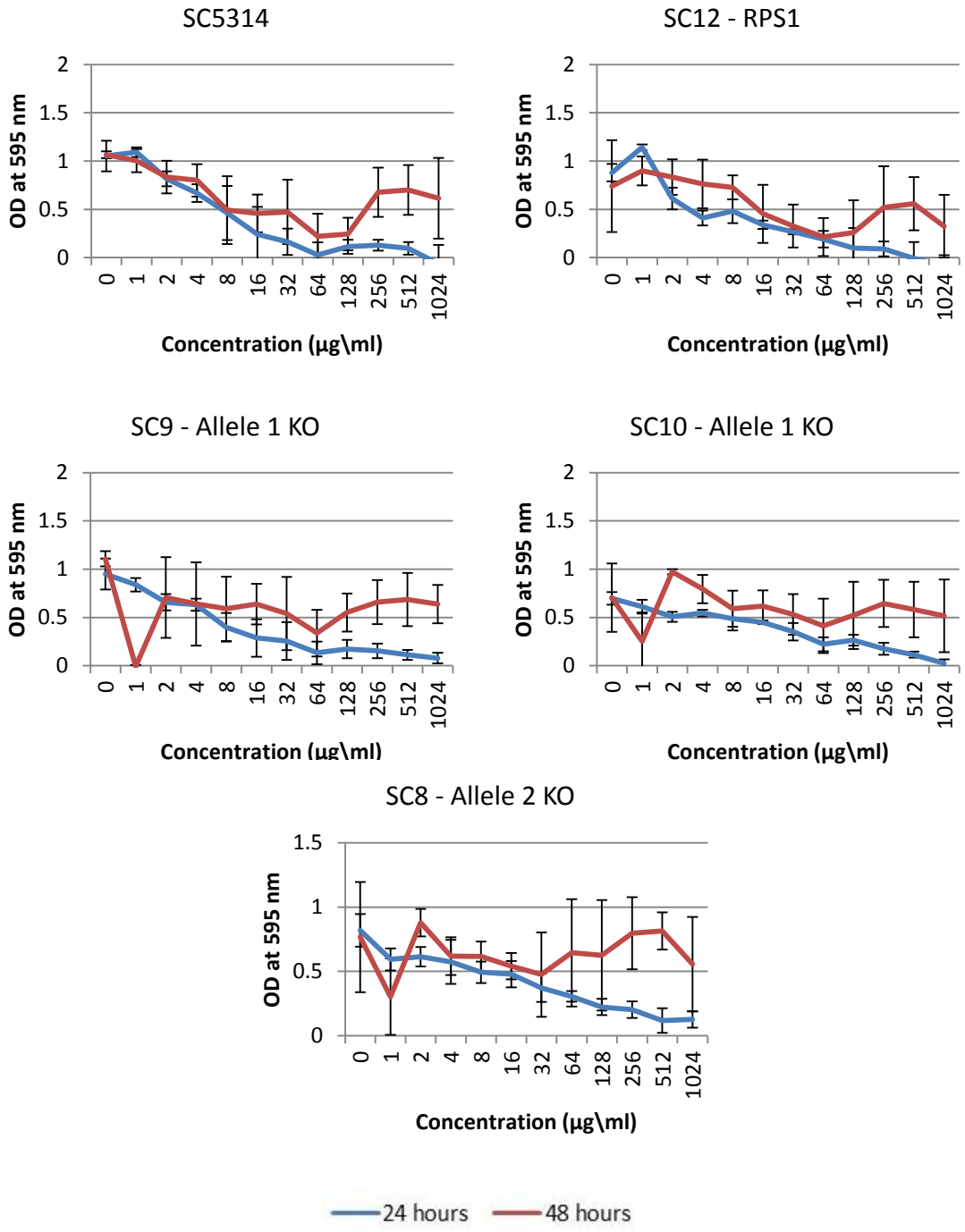


g) *Galleria mellonella* virulence assay at a cell concentration of 2×10^7 cells/ml.

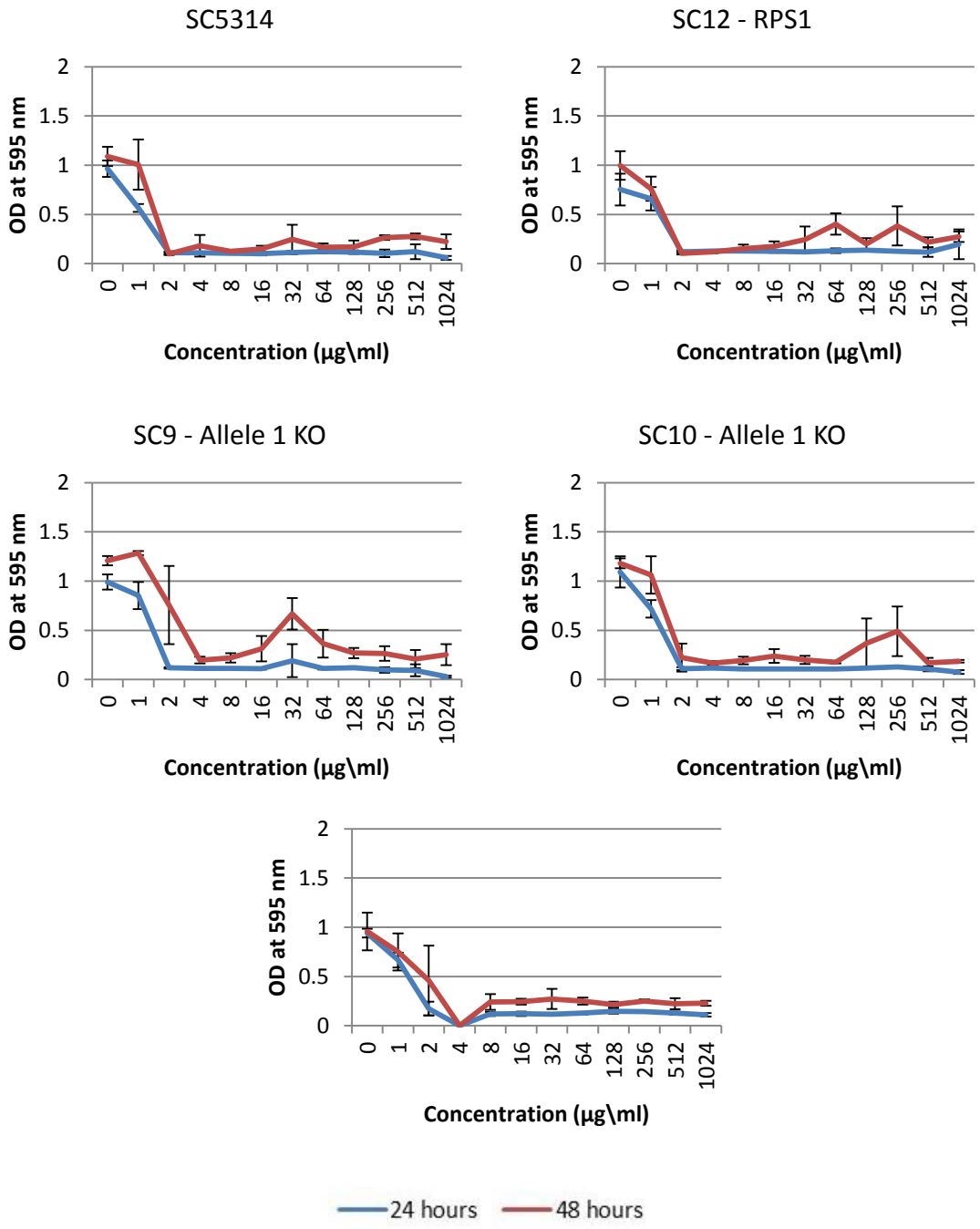
h) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



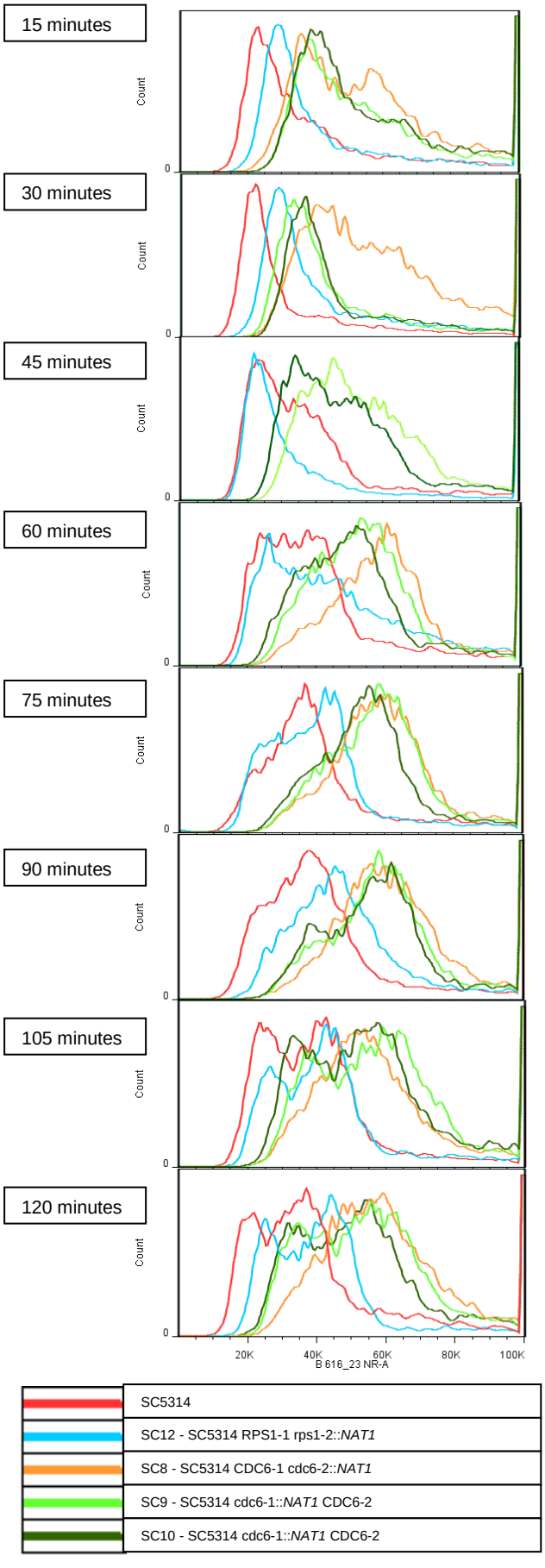
i) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



j) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



k



k) Cell cycle analysis using flow cytometry. Samples taken every 15 minutes.

4.3.2.2 *ERB1* Phenotypic Screening

The function of *C. albicans* gene *ERB1* is yet to be fully characterised, however, based upon orthology to the *S. cerevisiae* gene *ERB1,* predictions have been made for its involvement in synthesis of ribosomal subunits (Pestov *et al.*, 2001). Observations of tagged protein levels have been shown that *ERB1* levels are reduced during the transition to hyphae (Lee *et al.*, 2005). However, our results show that heterozygous knockout mutants of this gene have normal morphology both in the yeast and hyphal forms (Figure 4.7f). Large scale genome-wide studies have shown that heterozygous null mutants of *ERB1* have reduced resistance to a range of antifungal drugs including 5-flucytosine, 5-fluorouracil and tubercidin (Xu *et al.*, 2007). Conversely, our results have shown a normal response of heterozygous knockout mutants to 5-flucytosine (Figure 4.7i) and the other antifungal compound amphotericin B (Figure 4.7j). However, in growth in fluconazole, all knockout mutants demonstrated a slight increased sensitivity at 48 hours (Figure 4.7h). A TET-down *ERB1* repressive strain has shown attenuation of virulence in a murine infection model (Becker *et al.*, 2010), but virulence of heterozygous knockout mutants, using a *Galleria mellonella* model, has been shown to be largely unaffected (Figure 4.7g). Some significant differences for strains SC48 and SC56, both knockouts of allele 2, were observed at 2×10^7 cells/ml (Kaplan-Meier, d.f. = 1, $p = 0.048$ and 0.004 respectively). However this difference was not observed at the other two cell concentrations and the third isolate of this strain, SC47, also shows no significant differences. To be sure that this difference was not significant, a total of eight biological replicates were carried out for these strains.

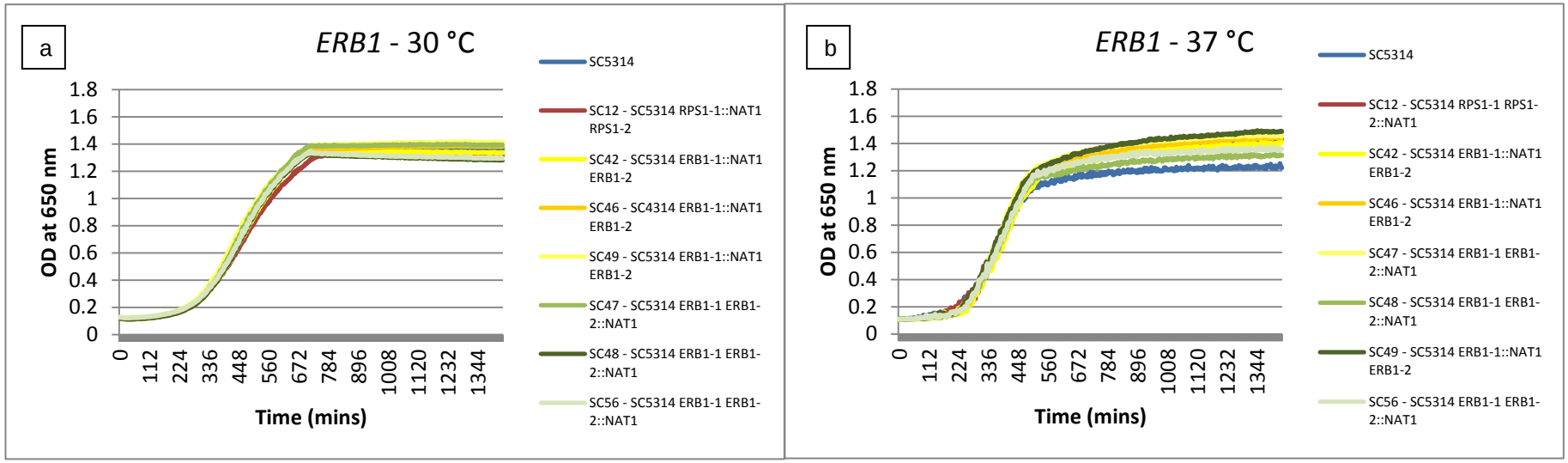
Other generic phenotypic screens also suggest that the alleles of *ERB1* do not differ in function. Growth at both 30 and 37 °C shows no significant differences in growth (ANOVA, $p > 0.05$; Figure 4.7a, Figure 4.7b, Table 4.4). And adhesion to buccal epithelial cells was comparable to the wild-type strain SC3514 for all three measures taken (Student's t test, d.f. = 2, $p > 0.05$; Figures 4.7c – e).

Table 4.4 Average generation times, times to maximum inflection and end-point optical densities of *ERB1* heterozygous knockout mutants at 30 °C and 37 °C (± one standard deviation)

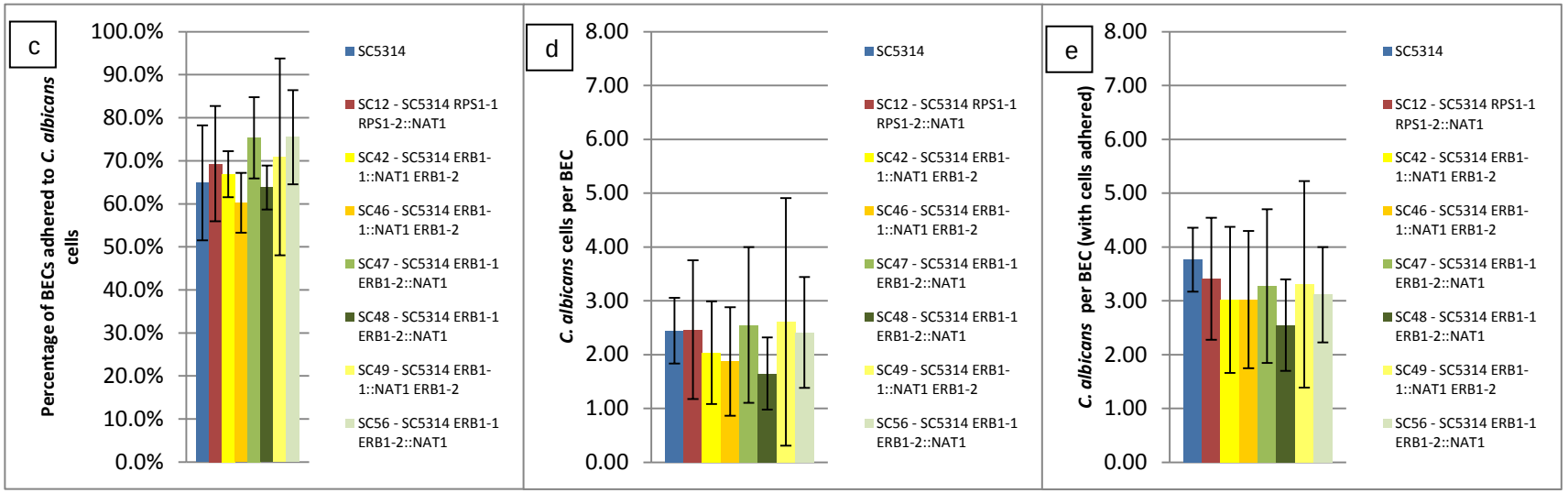
Growth Curve (from Figure 4.7)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	115.34 ± 9.88	325.06 ± 16.25	1.37 ± 0.15
	SC12	118.63 ± 8.37	328.13 ± 7.19	1.32 ± 0.09
Allele 1 knockout	SC42	117.94 ± 8.02	328.56 ± 11.45	1.35 ± 0.10
	SC46	111.24 ± 9.44	328.13 ± 10.54	1.38 ± 0.12
	SC49	111.64 ± 12.86	323.31 ± 11.67	1.39 ± 0.10
Allele 2 knockout	SC47	113.64 ± 9.02	328.13 ± 6.73	1.30 ± 0.11
	SC48	121.45 ± 6.00	312.81 ± 7.19	1.40 ± 0.03
	SC56	127.34 ± 13.85	321.50 ± 8.43	1.31 ± 0.02
b) 37 ° C	SC5314	54.86 ± 19.52	315.00 ± 33.44	1.20 ± 0.23
	SC12	60.35 ± 17.66	310.92 ± 30.40	1.38 ± 0.14
Allele 1 knockout	SC42	52.85 ± 21.37	314.96 ± 38.00	1.34 ± 0.16
	SC46	55.04 ± 15.89	302.17 ± 27.61	1.38 ± 0.16
	SC49	58.32 ± 13.02	294.54 ± 28.80	1.42* ± 0.11
Allele 2 knockout	SC47	65.07 ± 10.04	294.86 ± 29.96	1.42* ± 0.09
	SC48	61.66 ± 14.10	301.58 ± 25.63	1.27 ± 0.20
	SC56	59.57 ± 20.07	305.67 ± 26.88	1.32 ± 0.14

* Significantly different measurements from SC5314, identified by ANOVA followed by *post-hoc* analysis using a Dunnett's test, at $p < 0.05$, are annotated with an asterisk.

Figure 4.7 Phenotypic assays of *ERB1* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC42, SC46 and SC49 = knockout of “allele one” (yellow) and SC47, SC48 and SC56 = knockout of “allele two” (green).

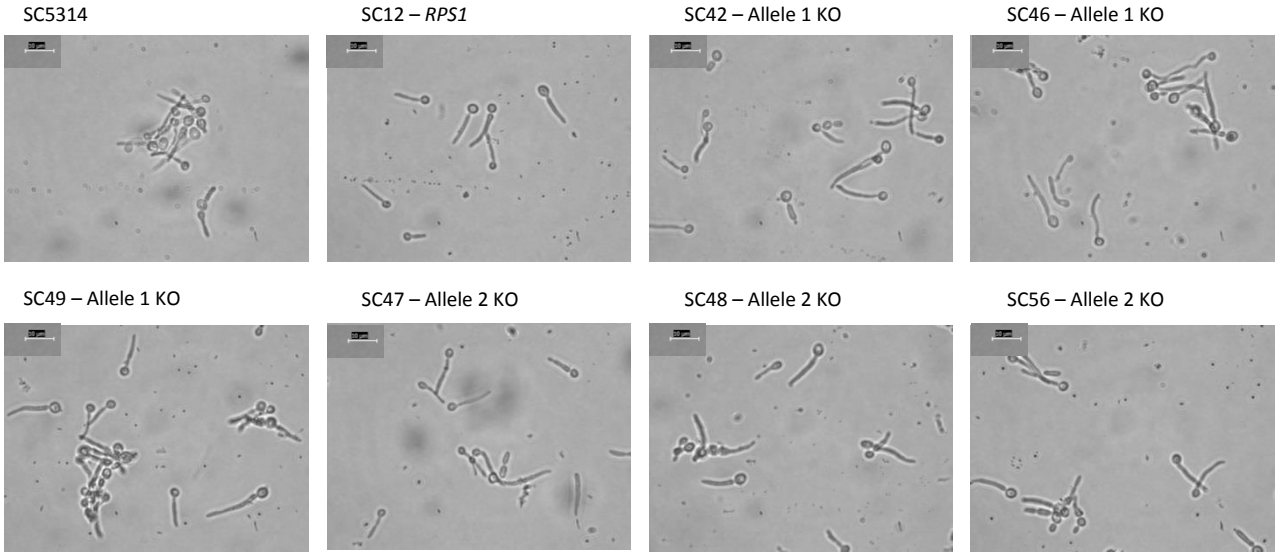


a) Growth rate at 30 °C and b) Growth rate at 37 °C



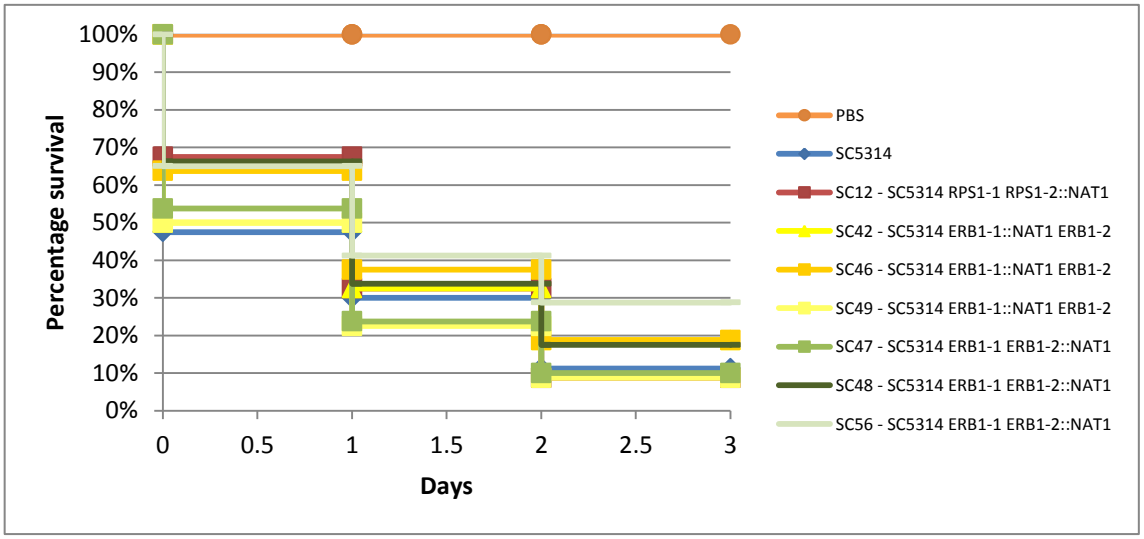
c) Percentage of BECs adhered to *C. albicans* cells. d) Number of *C. albicans* cells per BEC. e) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

f



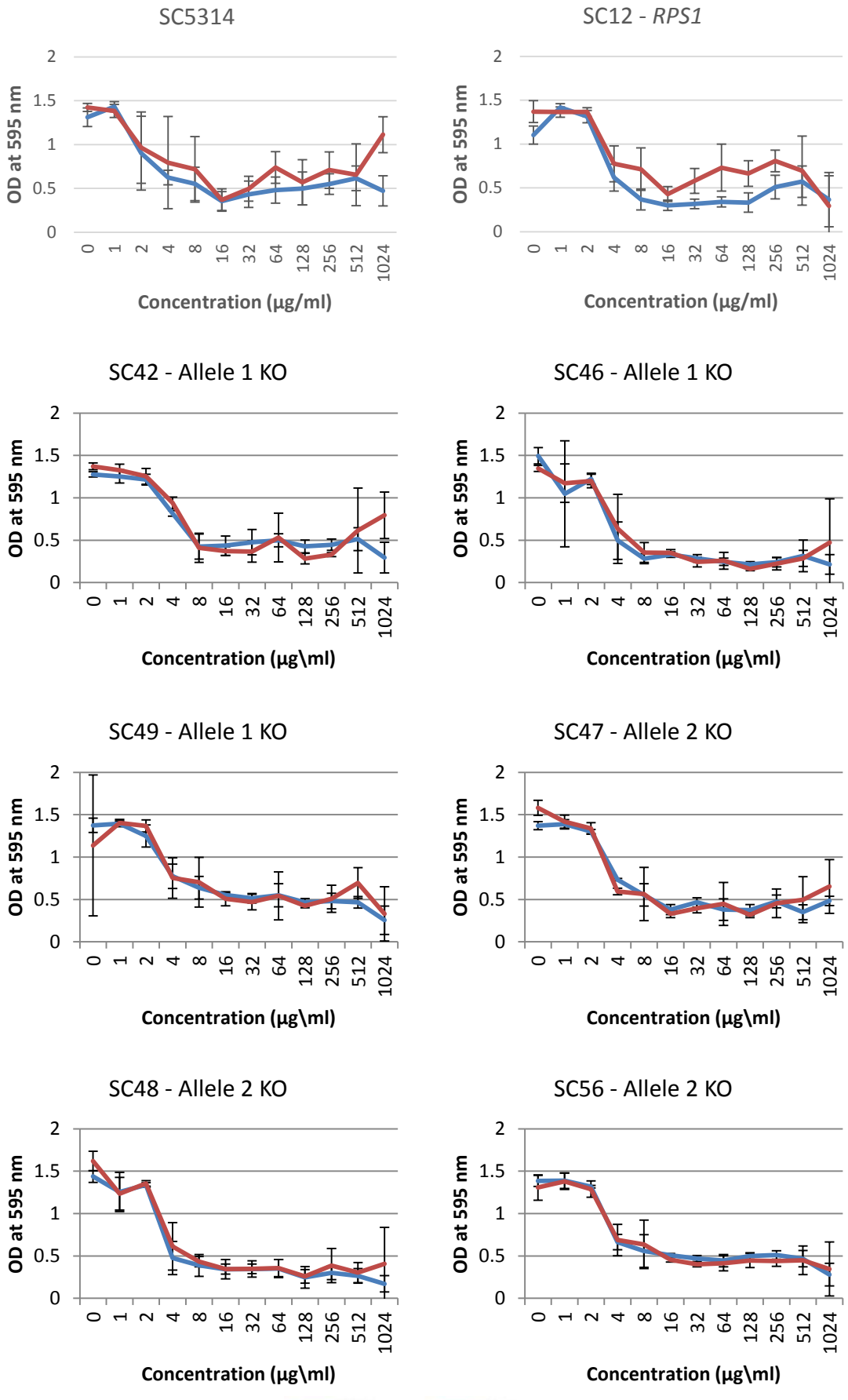
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

g

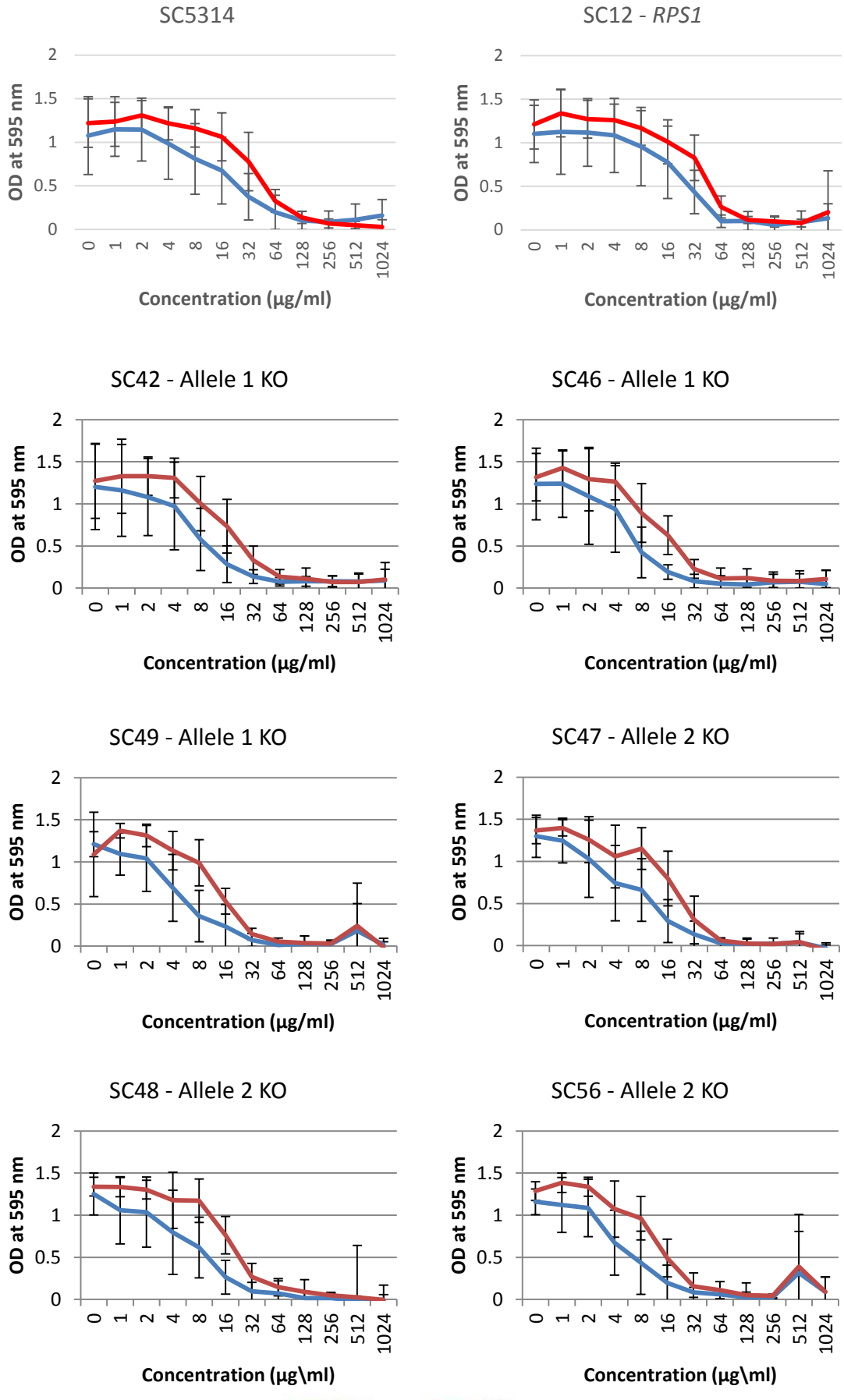


g) *Galleria mellonella* virulence assay at a cell concentration of 2×10^7 cells/ml.

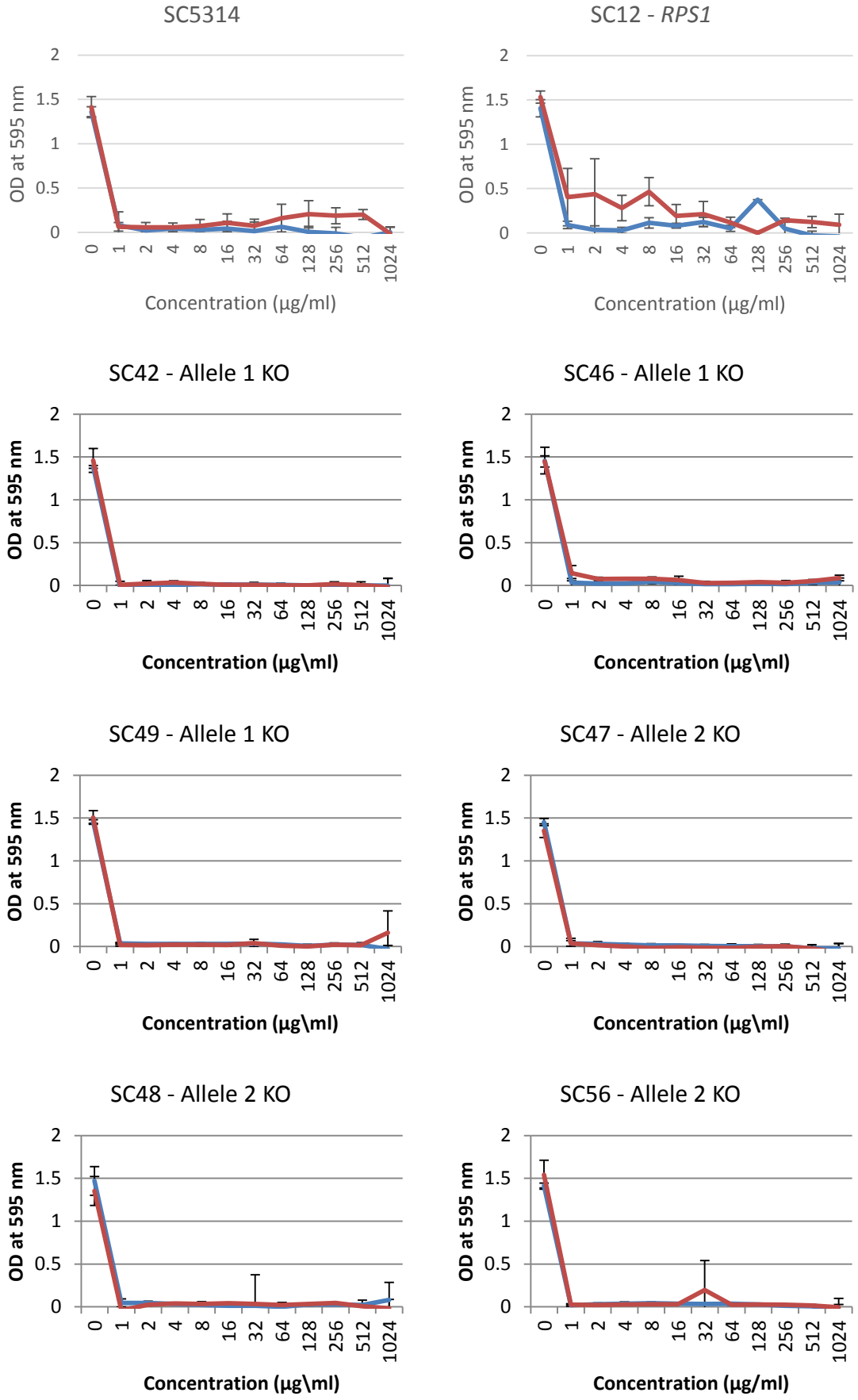
h) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



i) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



j) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



4.3.2.3 *RBT4* Phenotypic Screening

The gene *RBT4* has been associated with the transition from the yeast to hyphal form, and is named due to negative regulation by the gene *Tup1. Tup1* is a transcription factor which represses the switch of *C. albicans* cells from yeast to hyphal form. Deletion of *Tup1* induces the expression of *RBT4* and the switch to hyphal form (Braun *et al.*, 2000). Homozygous knockout strains of *RBT4* have been shown to have no differences in morphology or growth rate (Jackson *et al.*, 2007, Noble *et al.*, 2010), which is supported by our growth curve results as seen in Figures 4.8a and 4.8b and Table 4.5 (ANOVA, $p > 0.05$). However, knockout strains have been found to have reduced infectivity in BALB/c mice during both disseminated infection (Noble *et al.*, 2010) and corneal inflammation (Braun *et al.*, 2000, Jackson *et al.*, 2007). Conversely, here it is shown that heterozygous knockout mutants have no reduction in infectivity in a *Galleria mellonella* infection model (Kaplan-Meier test, d.f. = 4, $p > 0.05$; Figure 4.8g). During corneal infection with the *RBT4* deletion strain, it was also observed that only yeast cells were present, whereas many hyphal cells were seen during wild-type infection. The results here indicate that induction of hyphae is also unaffected in heterozygous knockout mutants (Figure 4.8f).

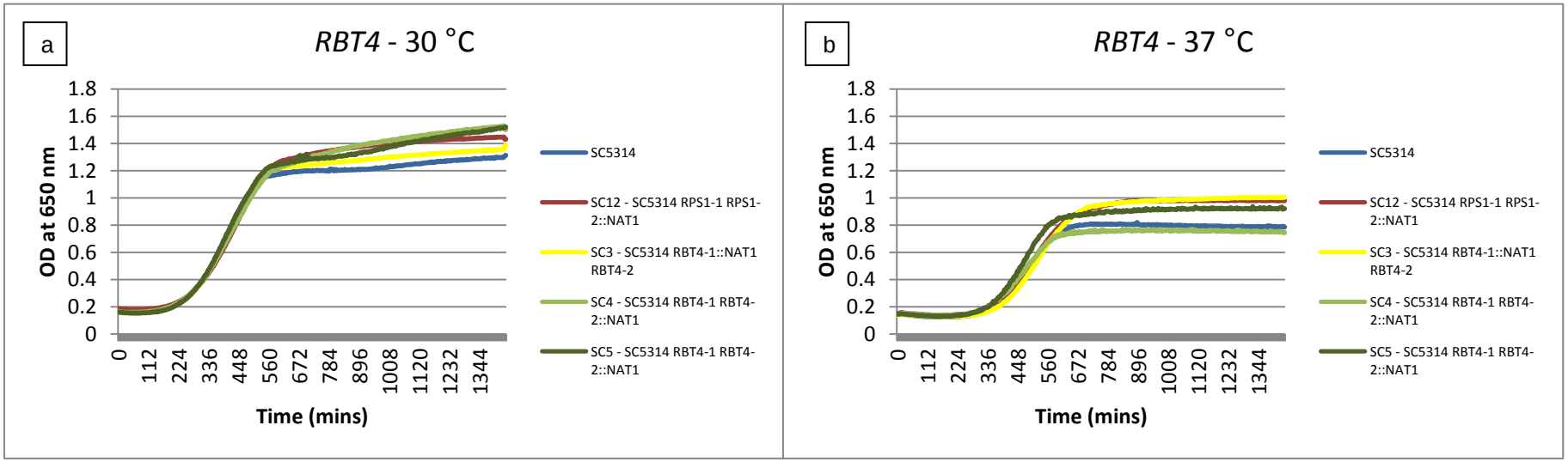
Screening in other phenotypic assays indicated no difference in function between the alleles of *RBT4.* All heterozygous knockout mutants showed similar capabilities as the wild-type strain to adhere to buccal epithelial cells for all three measures taken (Student t-test, d.f. = 2, $p > 0.05$; Figure 4.8c-e). All strains also showed no evidence of resistance or susceptibility to the antifungal drug treatments (Figure 4.8i-j). All three heterozygous knockout strains showed a marginal increase in growth in fluconazole at 48 hours when compared to the wild-type strain SC5314 and the control strain SC12 (Figure 4.8h), however this slight difference can most likely be accounted for by inter-plate variability.

Table 4.5 Average generation times, times to maximum inflection and end-point optical densities of *RBT4* heterozygous knockout mutants at 30 °C and 37 °C (± one standard deviation)

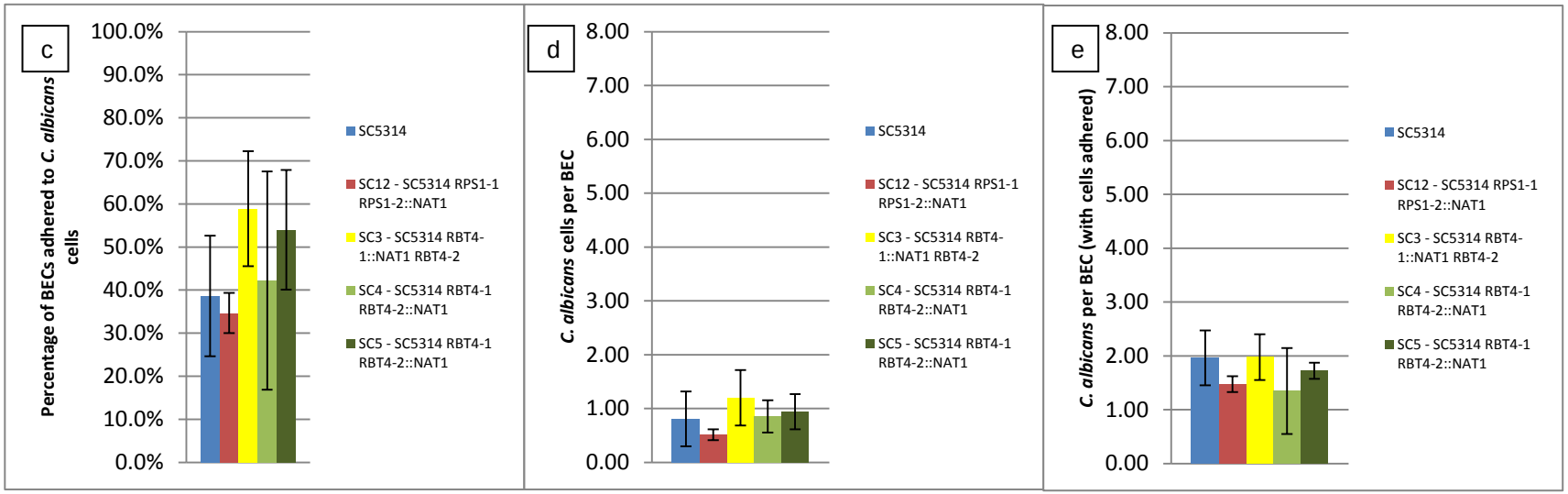
Growth Curve (from Figure 4.8)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	117.23 ± 36.54	306.69 ± 28.29	1.22 ± 0.17
	SC12	124.92 ± 59.84	326.20 ± 34.27	1.39* ± 0.20
Allele 1 KO	SC3	111.08 ± 17.51	311.50 ± 25.62	1.29 ± 0.15
Allele 2 KO	SC4	115.00 ± 19.22	322.00 ± 68.55	1.40* ± 0.14
	SC5	106.10 ± 18.54	310.41 ± 24.17	1.35 ± 0.16
b) 37 °C	SC5314	120.53 ± 29.04	448.58 ± 56.94	0.80 ± 0.23
	SC12	135.60 ± 28.20	449.75 ± 91.35	0.98 ± 0.19
Allele 1 KO	SC3	120.41 ± 18.63	461.71 ± 65.50	0.98 ± 0.34
Allele 2 KO	SC4	130.30 ± 25.66	415.33 ± 41.04	0.75 ± 0.28
	SC5	120.76 ± 17.52	404.83 ± 43.00	0.92 ± 0.26

* Significantly different measurements from SC5314, identified by ANOVA followed by post-hoc analysis using a Dunnett's test, at $p < 0.05$, are annotated with an asterisk.

Figure 4.8 Phenotypic assays of *RBT4* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC3 = knockout of “allele one” (yellow) and SC4 and SC5 = knockout of “allele two” (green).

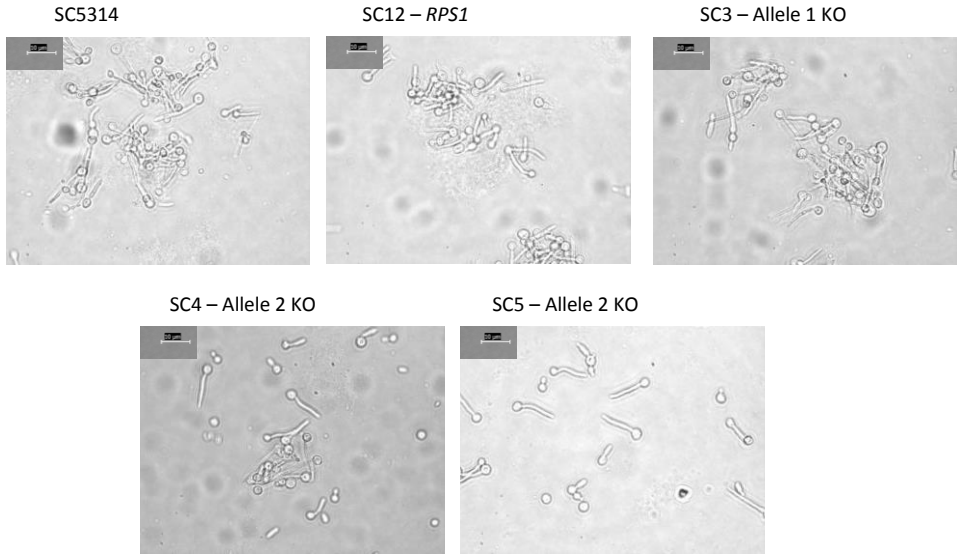


a) Growth rate at 30 °C. b) Growth rate at 37 °C.



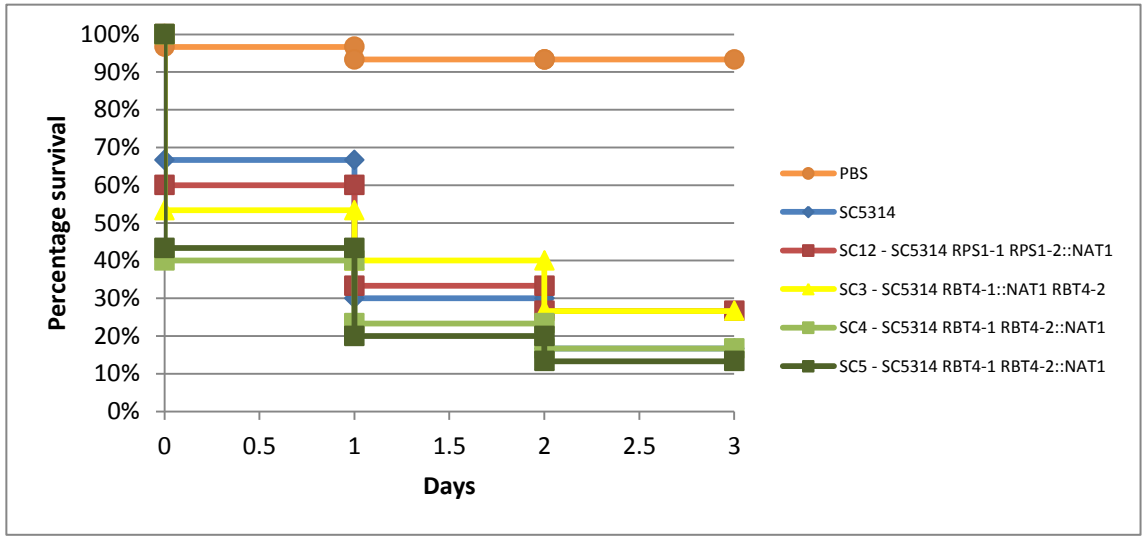
c) Percentage of BECs adhered to *C. albicans* cells. d) Number of *C. albicans* cells per BEC. e) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

f



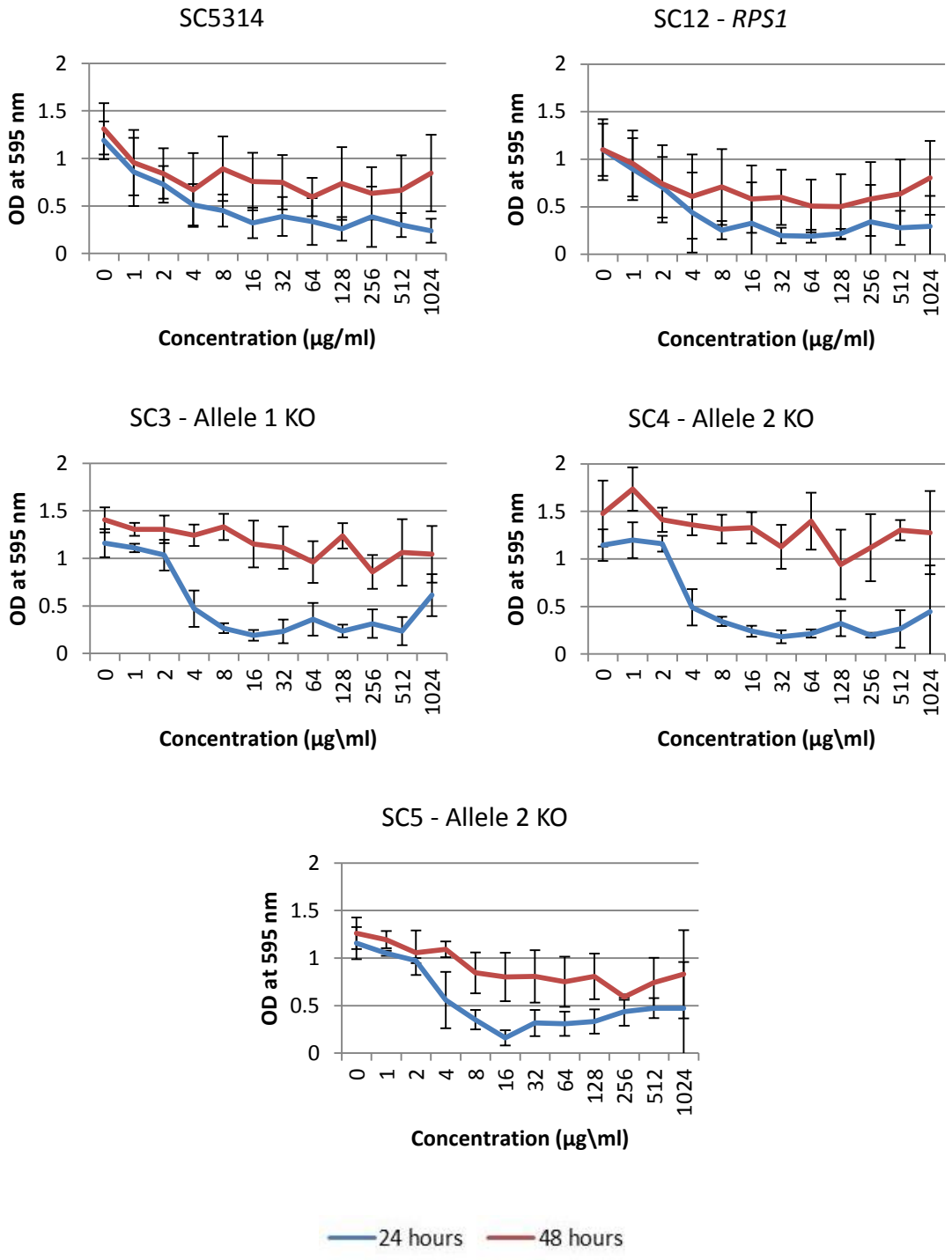
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

g

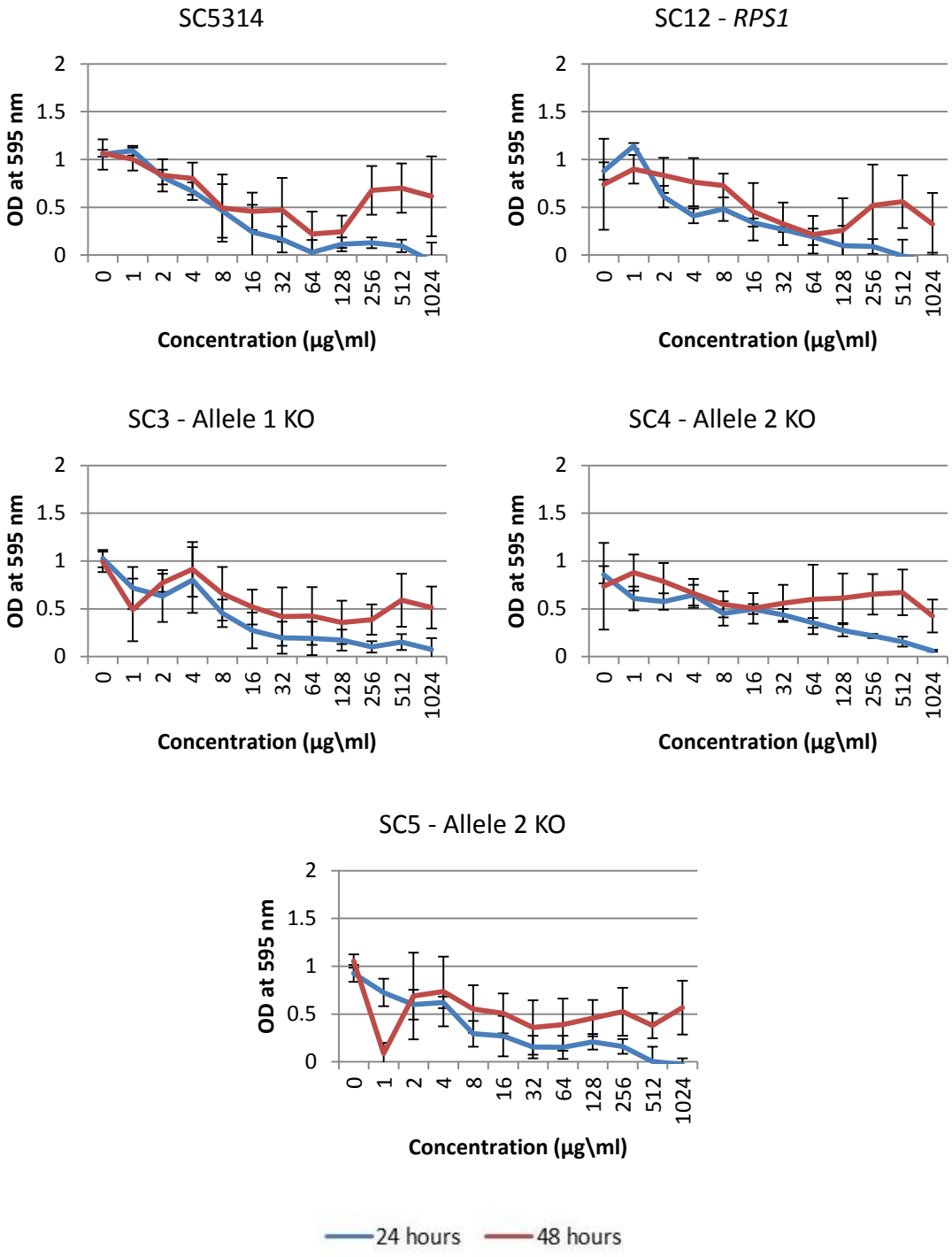


g) *Galleria mellonella* virulence assay at a cell concentration of 2×10^7 cells/ml.

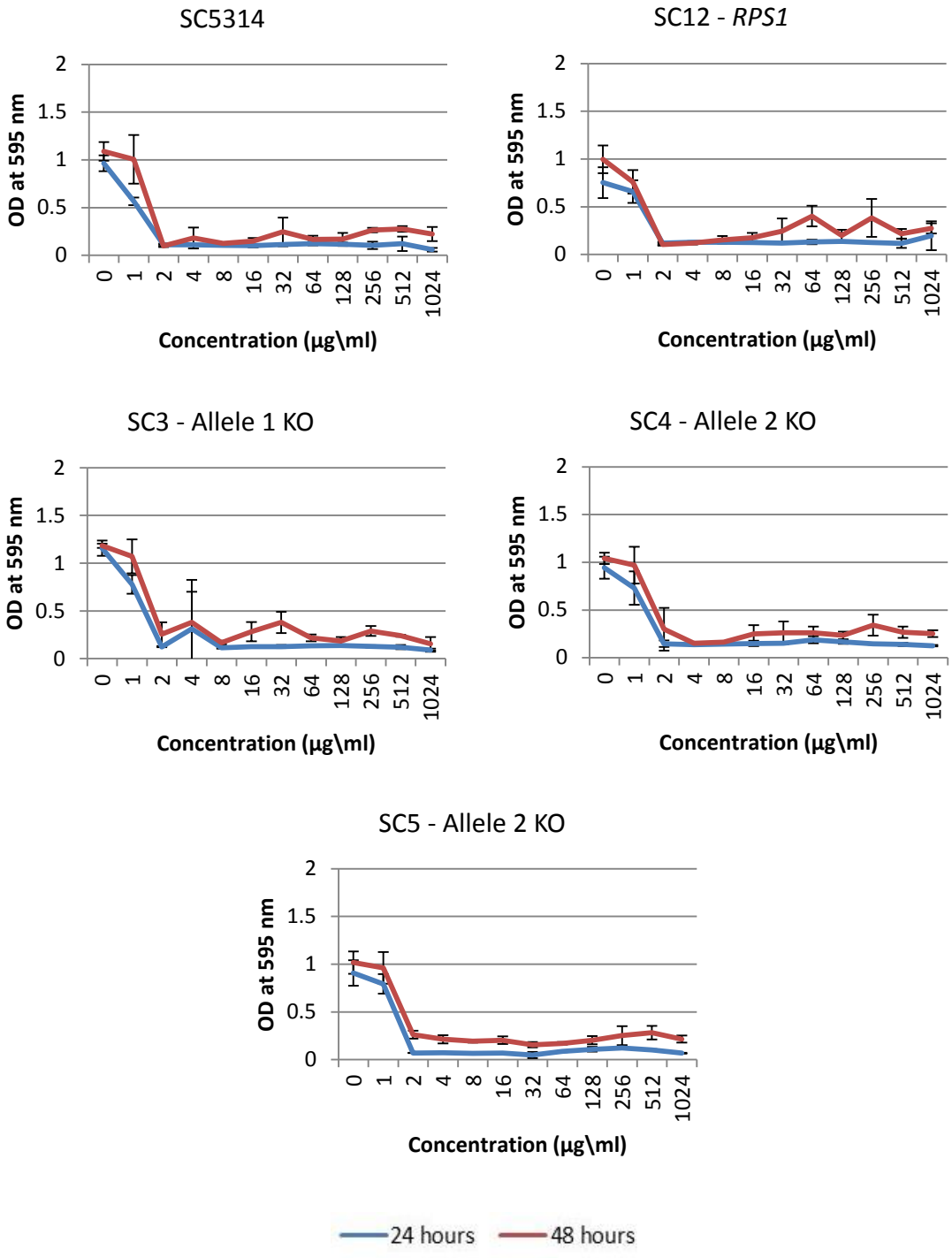
h) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



i) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



j) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



4.3.2.4 *SMI1* Phenotypic Screening

SMI1 is a gene involved in cell wall biosynthesis which appears to interact with the β-1,3-glucan synthase *FKS1* (Nett *et al.*, 2011). Homozygous deletion of the gene results in reduced biofilm antifungal resistance when exposed to fluconazole, amphotericin B and anidulafungin, as well as reduced β-1,3-glucan in the biofilm matrix, a thinner biofilm cell wall, and an increased susceptibility to calcofluor white stress. A heterozygous knockout strain also demonstrated the reduction in β-1,3-glucan but to a lesser extent. However growth rate, biofilm growth and virulence phenotypes were all unaffected by removal of *SMI1* (Nett *et al.*, 2011).

Phenotypic screening of heterozygous knockout mutants here also shows that growth rate (ANOVA, $p > 0.05$; Figures 4.9a and 4.9b; Table 4.5), virulence (Kaplan-Meier test, d.f. = 3; Figure 4.9g) and biofilm formation (ANOVA, $p > 0.05$; Figure 4.9k) are not altered significantly by removal of a copy of *SMI1.* Exposure of biofilms to the antifungal drugs fluconazole, 5-flucytosine and amphotericin B also showed no significant differences between heterozygous knockout mutants and the wild-type strain SC5314 (Figure 4.9l). However, both SC30 and SC33 appear to have a higher growth rate than SC5314 when exposed to all concentrations of amphotericin B (Figure 4.9l3). This results is surprising, and the opposite as to what is seen in the homozygous knockout strain (Nett *et al.*, 2011).

Investigations in the *S. cerevisiae* homolog of *SMI1* have also shown that gene knockouts have reduced β-glucan (Hong *et al.*, 1994, Dague *et al.*, 2010), reduced cell wall elasticity, increased mannan and chitin (Dague *et al.*, 2010), increased sensitivity to caspofungin (Markovich *et al.*, 2004), increased sensitivity to SDS and cercosporamide (Hong *et al.*, 1994), increased sensitivity to ethanol (Dudley *et al.*, 2005, van Voorst *et al.*, 2006, Yoshikawa *et al.*, 2009), and increased sensitivity to the antifungal drug hygromycin B and the stress condition calcium chloride (Dudley *et al.*, 2005).

Based upon these observations, growth of the heterozygous knockout mutants was monitored under a range of conditions, including SDS, ethanol, calcium chloride, hygromycin B and calcofluor white. Figure 4.9m shows that under all of

the conditions tested, there is no significant sensitivity of any of the mutants when compared to the wild-type strain SC5314, however SC33, the knockout of allele one shows a very minimal sensitivity to SDS and ethanol with reduced growth and colony size at the lowest cell concentration. It is difficult to infer from the results here whether there is a minor functional difference in the alleles or if this observation is due to uneven plating of the strains with the replicator.

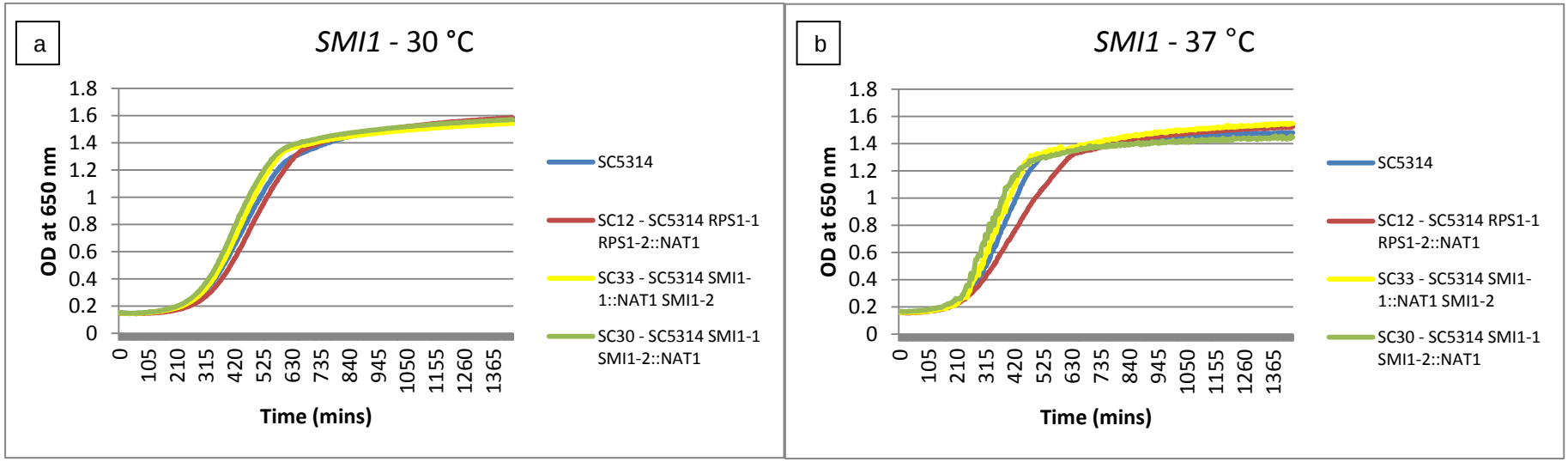
A possible assay that may have indicated phenotypic differences between the heterozygous knockout strains of each allele, but has not been included in this investigation, is the measurement of β-glucan. The methodology for this assay, as previously seen in (Hong *et al.*, 1994, Dague *et al.*, 2010), was not available for use here, but is something that could be investigated as part of further work.

Results of other phenotypic assays indicate no functional differences between the alleles of *SMI1.* All heterozygous knockout strains had comparable adhesion to buccal epithelial cells to the wild-type strain for all three measures taken (student's t-test, d.f. = 2, $p > 0.05$; Figures 4.9c-e). All strains were able to switch from the yeast to hyphal form (Figure 4.9f). Growth under exposure to the antifungal drug treatments fluconazole (Figure 4.9h), 5-flucytosine (Figure 4.9i) and amphotericin B (Figure 4.9j) is comparable for all strains at both 24 and 48 hours.

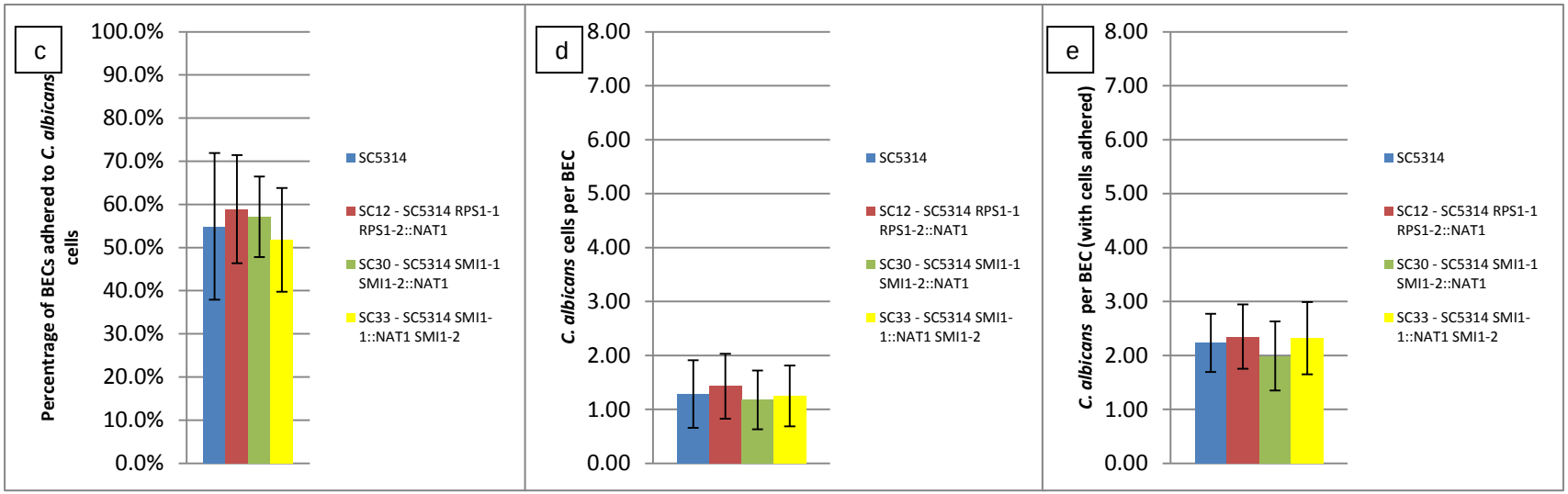
Table 4.6 Average generation times, times to maximum inflection and end-point optical densities of *SMI1* heterozygous knockout mutants at 30 °C and 37 °C (± one standard deviation)

Growth Curve (from Figure 4.9)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	127.49 ± 38.30	344.75 ± 80.57	1.48 ± 0.18
	SC12	123.68 ± 34.77	357.00 ± 43.54	1.49 ± 0.11
Allele 1 knockout	SC33	118.29 ± 31.14	306.54 ± 28.53	1.50 ± 0.12
Allele 2 knockout	SC30	121.59 ± 30.46	322.88 ± 21.01	1.47 ± 0.11
b) 37 ° C	SC5314	108.28 ± 36.15	258.13 ± 77.11	1.43 ± 0.21
	SC12	116.43 ± 31.22	304.06 ± 123.98	1.46 ± 0.16
Allele 1 knockout	SC33	87.36 ± 52.91	240.19 ± 38.97	1.48 ± 0.16
Allele 2 knockout	SC30	98.47 ± 48.64	228.44 ± 28.13	1.40 ± 0.12

Figure 4.9 Phenotypic assays of *SMI1* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC33 = knockout of “allele one” (yellow) and SC30 = knockout of “allele two” (green).

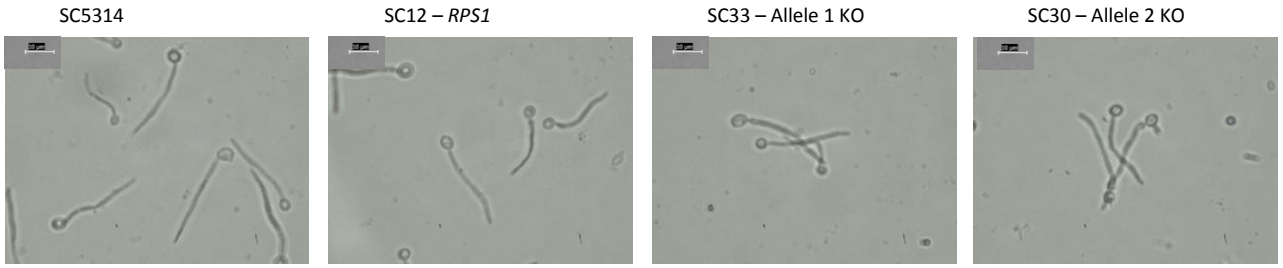


a) Growth rate at 30 °C. b) Growth rate at 37 °C.



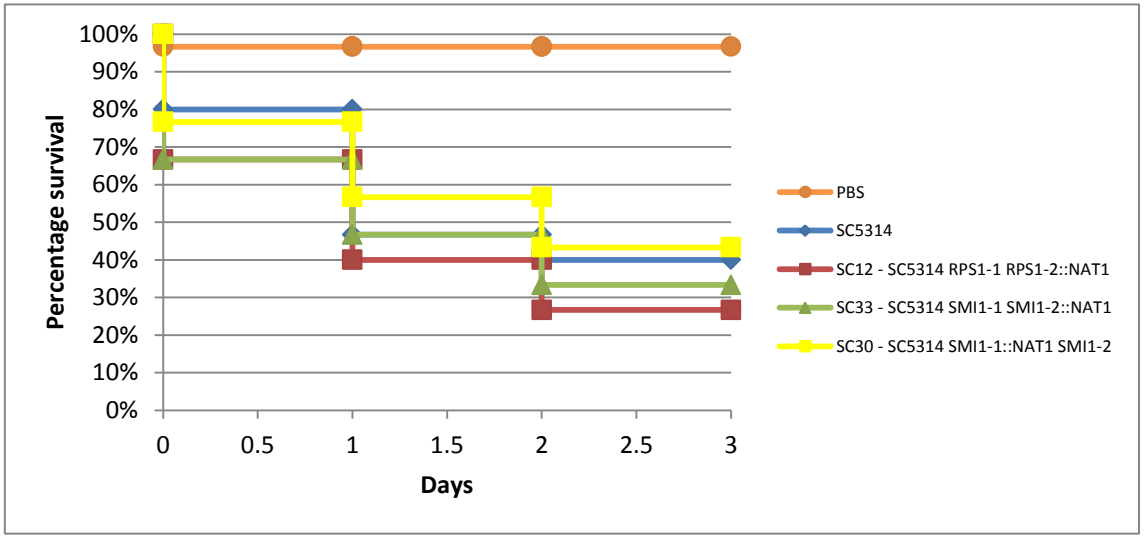
c) Percentage of BECs adhered to *C. albicans* cells. d) Number of *C. albicans* cells per BEC. e) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

f



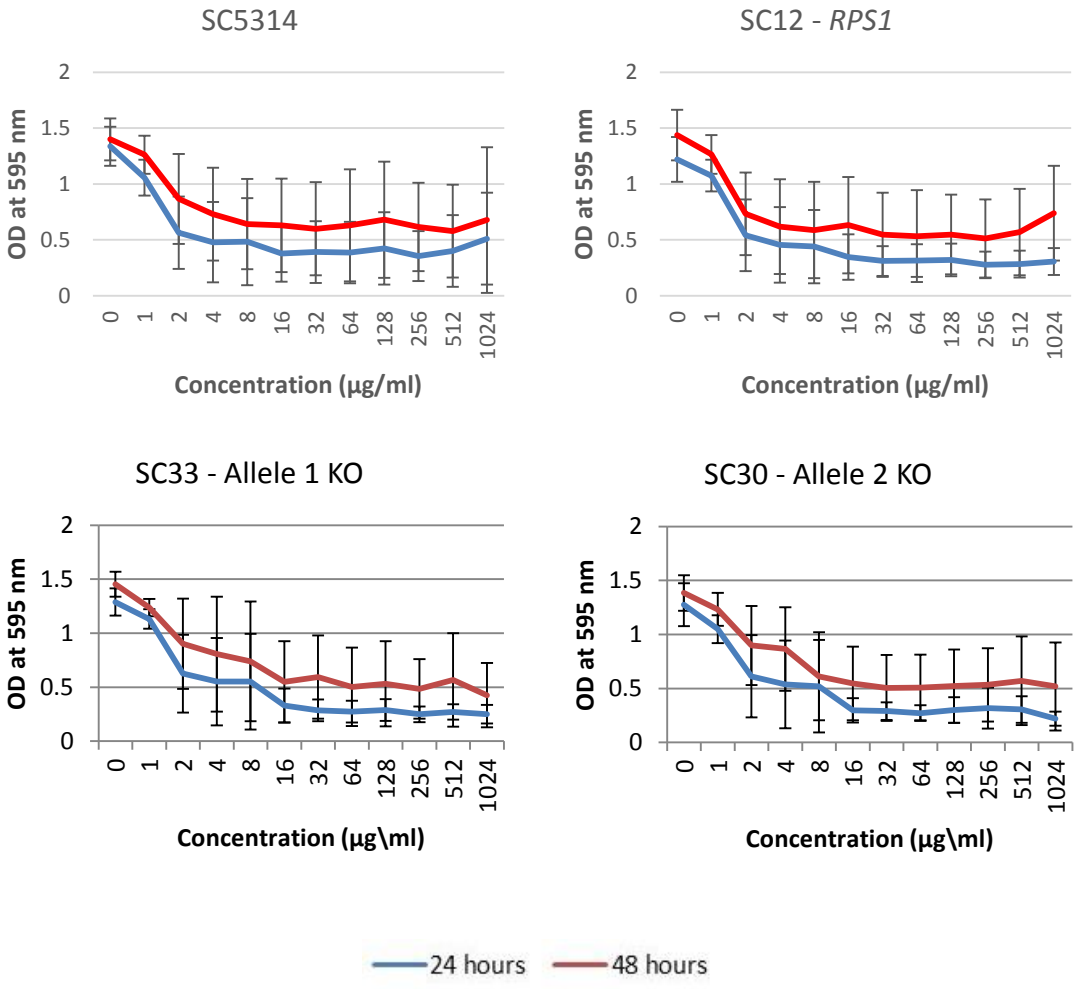
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

g

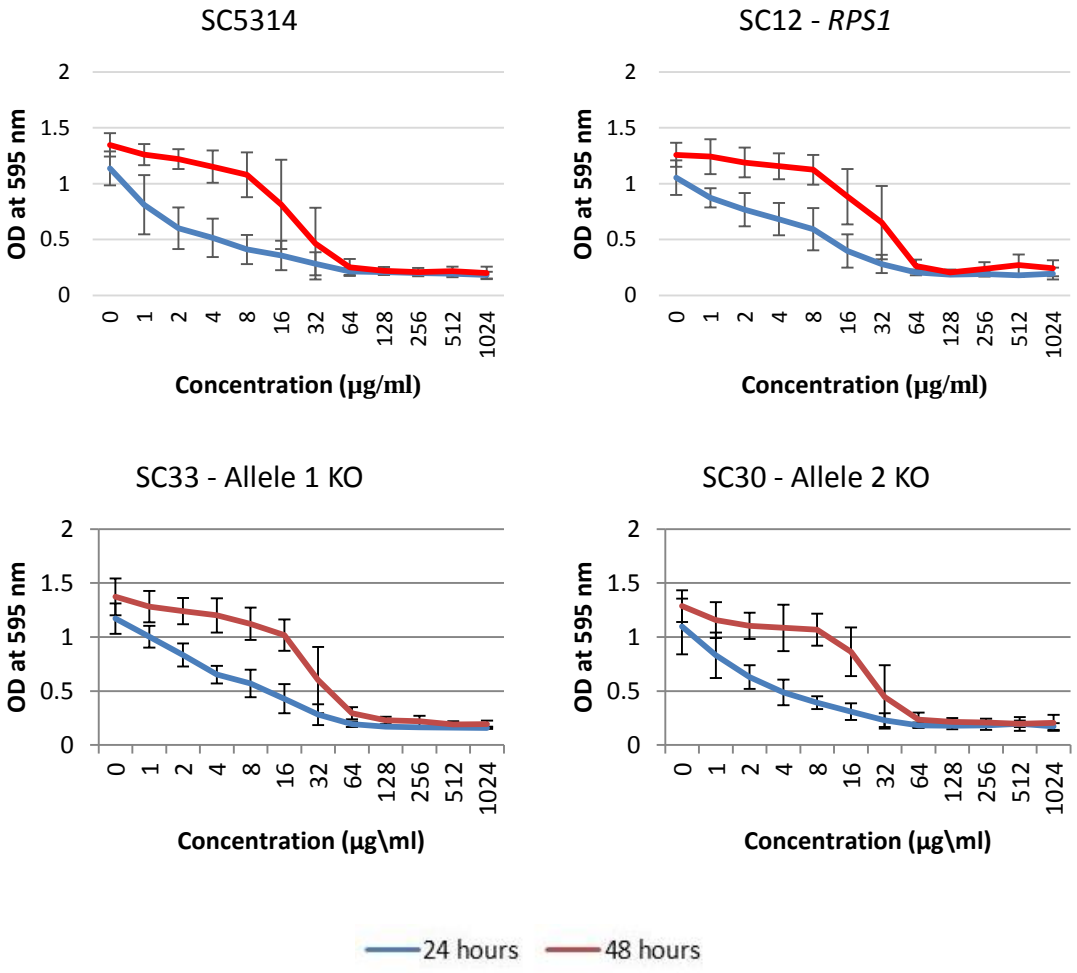


g) *Galleria mellonella* virulence assay at a cell concentration of 2×10^7 cells/ml.

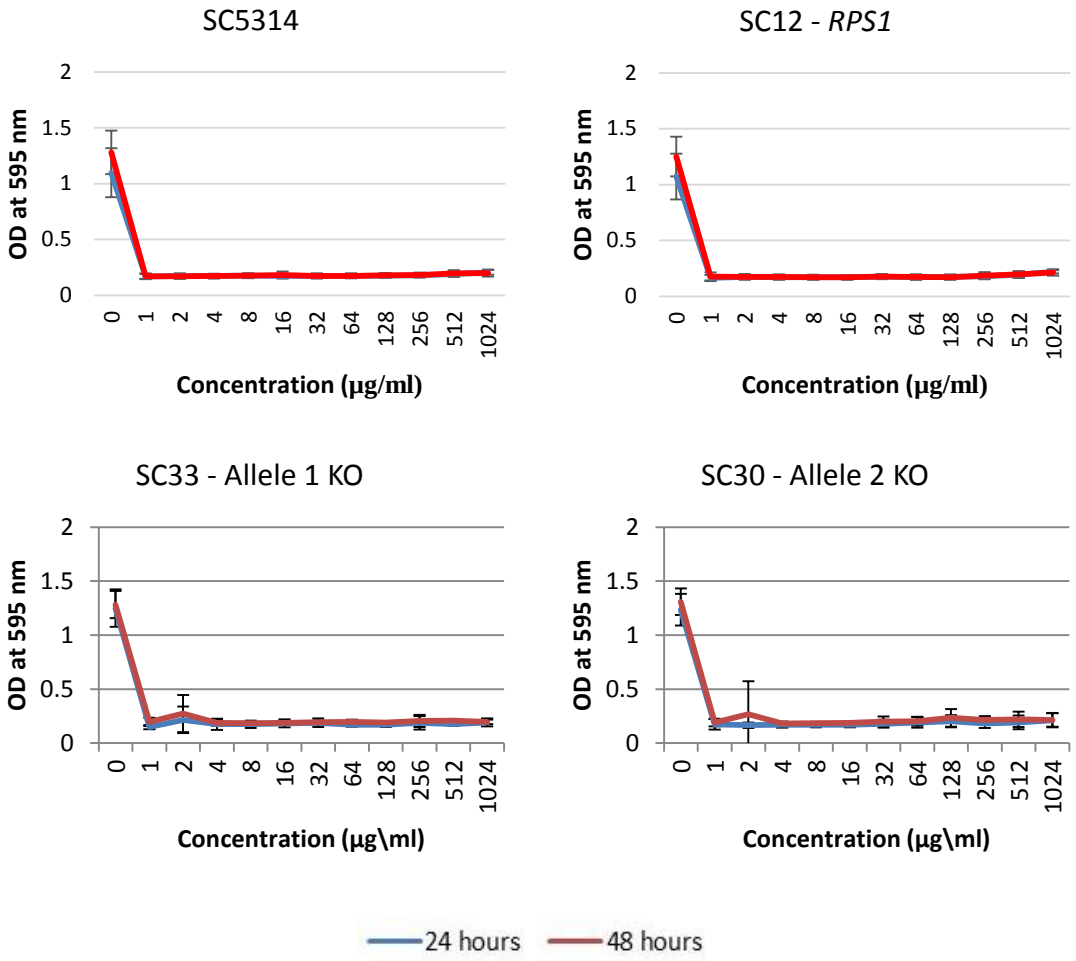
h) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



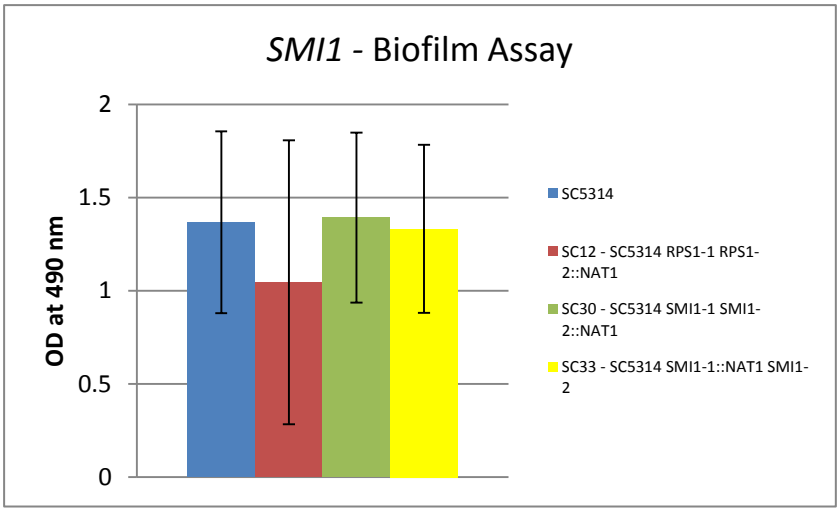
i) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



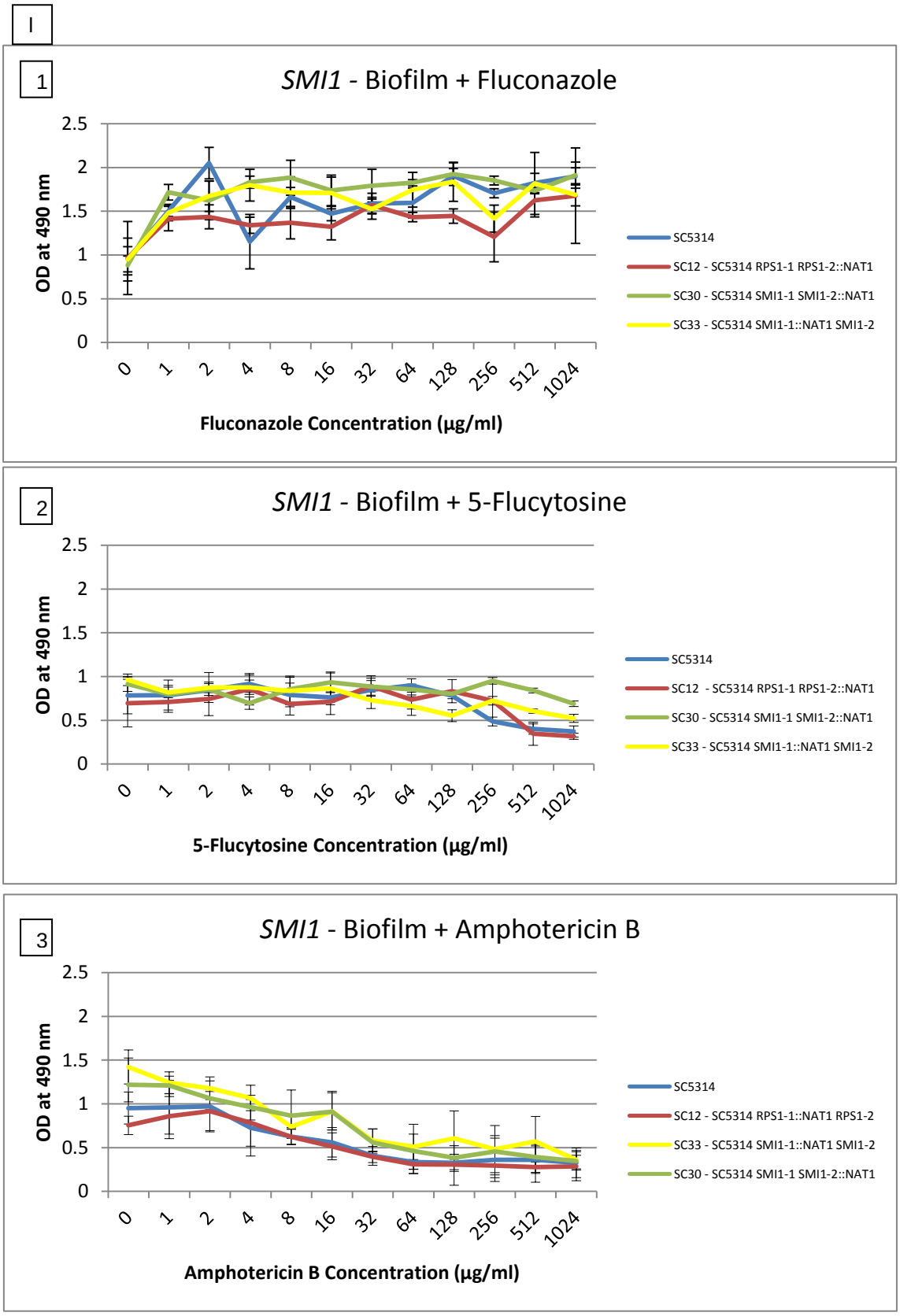
j) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



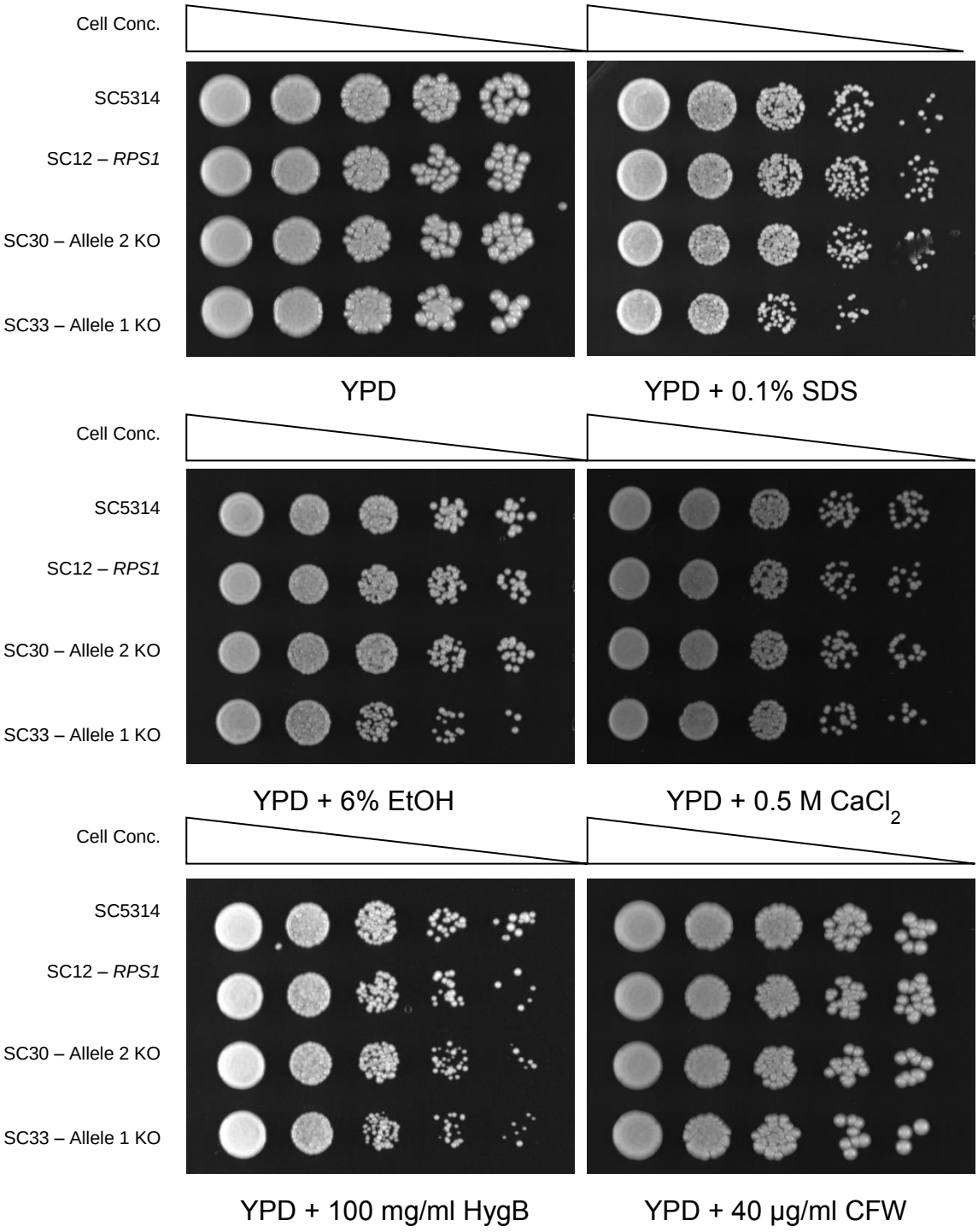
k



k) Biofilm production. Representative results for the 24 hour time point shown here. Error bars represent ± one standard deviation.



l) Biofilm production under exposure to antifungal drug treatments at a range of concentrations from 0 to 1024 µg/ml. 1) fluconazole, 2) 5-flucytosine and 3) amphotericin B. Error bars represent ± one standard deviation.



m) Growth under stress conditions. Cells are plated in tenfold dilution from 1 x 10^7 cells/ml to 1 x 10^3 cells/ml. Growth on YPD as a control compared to growth on YPD + 0.1% SDS, 6% ethanol, 0.5 M calcium chloride, 100 mg/ml hygromycin B and 40 µg/ml calcofluor white.

4.3.2.5 *VPS1* Phenotypic Screening

In *C. albicans,* the gene *VPS1* encodes a dynamin-like GTPase. This gene has been well characterised and has been found to functionally complement the *VPS1* gene in *S. cerevisiae* (Bernardo *et al.*, 2008). From this it has been inferred that *C. albicans VPS1* is involved in the sorting of vacuolar proteins. As an essential gene, the function has been tested using a conditional mutant under control of a tetracycline-regulatable promoter. This mutant was found to have normal growth rates, and a normal response to flucytosine and amphotericin B but was hyper-sensitive to the antifungal drug fluconazole, had abnormal vacuolar morphology, defects in filamentation and biofilm formation, and have a decrease in secretion of extracellular proteases such as aspartyl proteinases and lipase. (Bernardo *et al.*, 2008).

Interestingly, *VPS1* has not been found to be essential in *S. cerevisiae.* Removal of the gene has been associated with numerous phenotypic effects, some of which are similar to those seen in *C. albicans,* including mis-sorting of the vacuolar protein carboxypeptidase Y, abnormal fragmented vacuolar morphology (Peters *et al.*, 2004, Yu and Cai, 2004, Bernardo *et al.*, 2008) which is a classical phenotype of *VPS* genes, a decrease in number and increase in size of peroxisomes (Hoepfner *et al.*, 2001, Kuravi *et al.*, 2006, Vizeacoumar *et al.*, 2006), an increase in the time taken to internalise endocytic vesicles (Yu and Cai, 2004, Nannapaneni *et al.*, 2010, Rooij *et al.*, 2010) suggestive of a role in invagination, disruption of the actin cytoskeleton (Yu and Cai, 2004, Nannapaneni *et al.*, 2010) although normal actin has also been observed in similar mutants (Hoepfner *et al.*, 2001), a delay in cytokinesis (Hoepfner *et al.*, 2001), and a defect in Golgi protein retention (Wilsbach and Payne, 1993, Nothwehr *et al.*, 1996).

From these previous studies, the *VPS1* heterozygous knockout mutants constructed here were screened for a number of phenotypic defects. However, unlike the results found by Bernardo *et al.* (2008), the response of these strains to growth in fluconazole was less striking, with justSC7, a single isolate lacking allele two, which had increased sensitivity at 24 hours (Figure 4.10h). This is supported by growth on solid agar containing 64 µg/ml fluconazole (Figure 4.10k) The ability to form hyphae was also shown to be normal in these

heterozygous mutants (Figure 4.10f). Lipase secretion did not significantly differ between mutants (ANOVA, $p > 0.05$; Figure 4.10l), nor did the strains abilities to form biofilms (ANOVA, $p > 0.05$; Figure 4.10m). Vacuole morphology, as indicated by FUN-1 staining (detailed in section 4.2.3.4) shows that the knockout strains of allele two; SC7, SC18 and SC32; all have a larger vacuole than the wild-type strain. However this is also the case for strain SC17, a knockout of allele one (Figure 4.10n).

Despite some isolates showing significant differences in growth rate measurements at 30 °C, upon observation of the figure, the biological significance of these differences are not apparent (ANOVA followed by Dunnett's test, $p < 0.05$; Figure 4.10a and Table 4.7). Growth rates were found to not differ significantly at 37 °C (ANOVA, $p > 0.05$; Figure 4.10b and Table 4.7), with the exception of the strain SC7, the knockout of *VPS1* allele two, which has a significantly longer generation time and longer time to maximum inflection. However, the other two isolates of this allele knockout, SC18 and SC32, appear to have growth rates comparable to the wild-type strain.

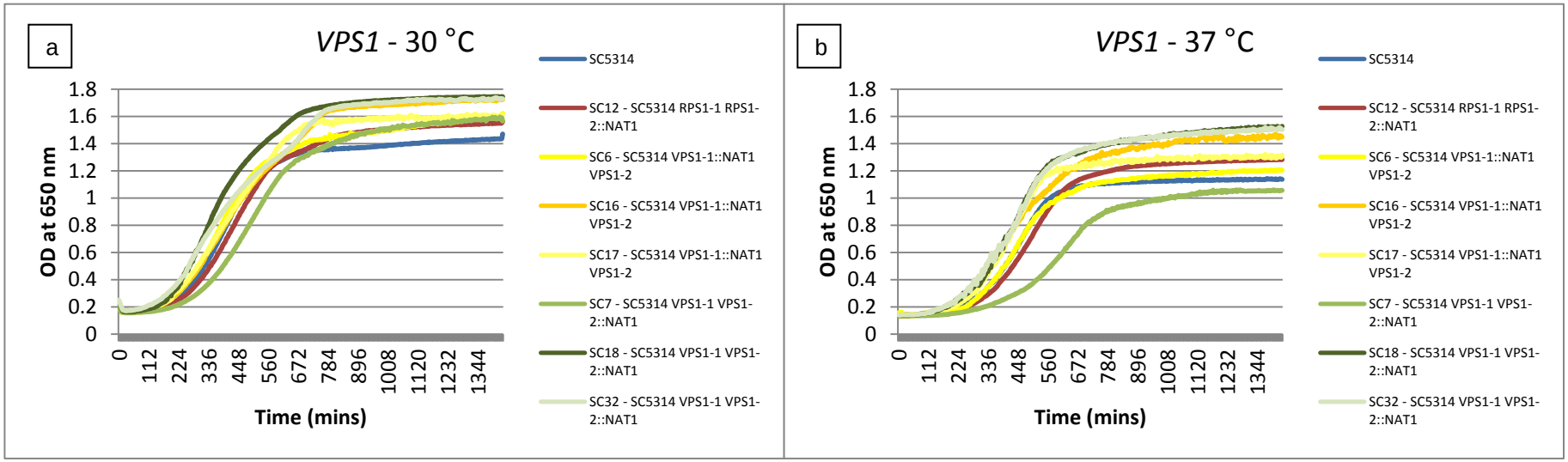
Results of other phenotypic assays indicate no functional differences between the alleles of *VPS1.* The ability to adhere to buccal epithelial cells was not significantly different for any of the heterozygous knockout mutants for any of the three measures taken (student t-test, $p > 0.05$; Figures 4.10c-e), suggesting that the alleles of *VPS1* do not have a functional role in adhesion. Virulence using the *Galleria mellonella* infection model showed that none of the heterozygous mutants were attenuated significantly for virulence (Kaplan-Meier test, d.f. = 7, $p > 0.05$; Figure 4.10g). Growth after exposure to the antifungal drug amphotericin B did not differ between all strains (Figure 4.10j) and the response to 5-flucytosine was generally comparable for all isolates lacking allele one. However, strains lacking allele two differed in their response to 5-flucytosine considerably with SC7 showing reduced sensitivity at 48 hours and SC18 and SC32 showing increased sensitivity at 48 (Figure 4.10i). However, as growth on solid agar containing 32 µg/ml 5-flucytosine shows that at 48 hours only SC7 has increased sensitivity (Figure 4.10k), these differences are likely to be accounted for by inter-plate variation in the assay used.

Table 4.7 Average Generation Times, Times to Maximum Inflection and End-Point Optical Densities of *VPS1* Heterozygous Knockout Mutants at 30 °C and 37 °C (± one standard deviation)

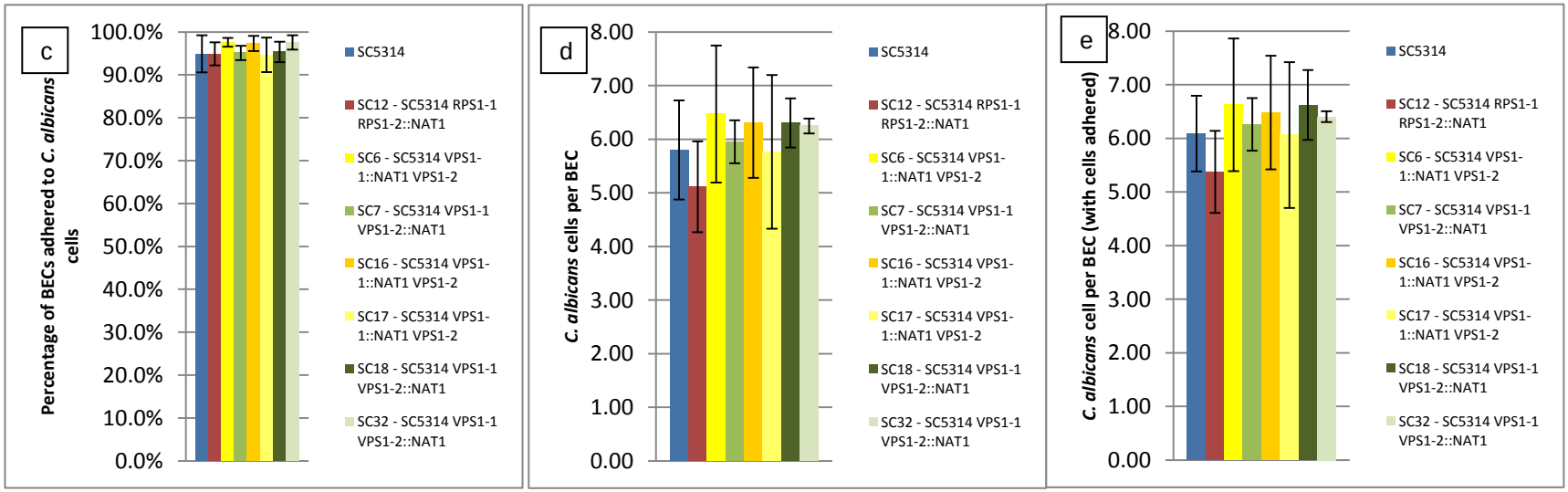
Growth Curve (from Figure 4.10)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	119.74 ± 30.63	280.88 ± 49.69	1.38 ± 0.27
	SC12	124.74 ± 50.20	304.96 ± 48.06	1.50 ± 0.24
Allele 1 KO	SC6	109.06 ± 16.80	263.96 ± 55.37	1.48 ± 0.20
	SC16	122.11 ± 15.54	247.63 ± 28.11	1.68* ± 0.07
	SC17	116.11 ± 18.26	245.44 ± 20.36	1.58 ± 0.19
Allele 2 KO	SC7	140.92 ± 27.49	336.44* ± 57.62	1.49 ± 0.16
	SC18	109.52 ± 25.32	232.75 ± 50.65	1.71* ± 0.07
	SC32	132.81 ± 9.19	215.69* ± 51.78	1.70* ± 0.09
b) 37 °C	SC5314	103.37 ± 28.92	367.35 ± 115.90	1.12 ± 0.41
	SC12	116.83 ± 30.55	388.35 ± 125.48	1.25 ± 0.33
Allele 1 KO	SC6	104.22 ± 44.59	347.93 ± 110.93	1.16 ± 0.41
	SC16	94.85 ± 27.38	308.29 ± 151.98	1.39 ± 0.13
	SC17	94.98 ± 30.46	290.21 ± 87.96	1.29 ± 0.21
Allele 2 KO	SC7	153.62* ± 37.86	488.48* ± 150.63	0.98 ± 0.36
	SC18	85.46 ± 27.03	286.13 ± 106.79	1.44* ± 0.15
	SC32	91.71 ± 32.16	282.63 ± 110.53	1.45* ± 0.16

* Significantly different measurements from SC5314, identified by ANOVA followed by *post-hoc* analysis using a Dunnett's test, at $p < 0.05$, are annotated with an asterisk.

Figure 4.10 Phenotypic assays of *VPS1* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC6, SC16 and SC17 = knockout of “allele one” (yellow) and SC7, SC18 and SC32 = knockout of “allele two” (green).

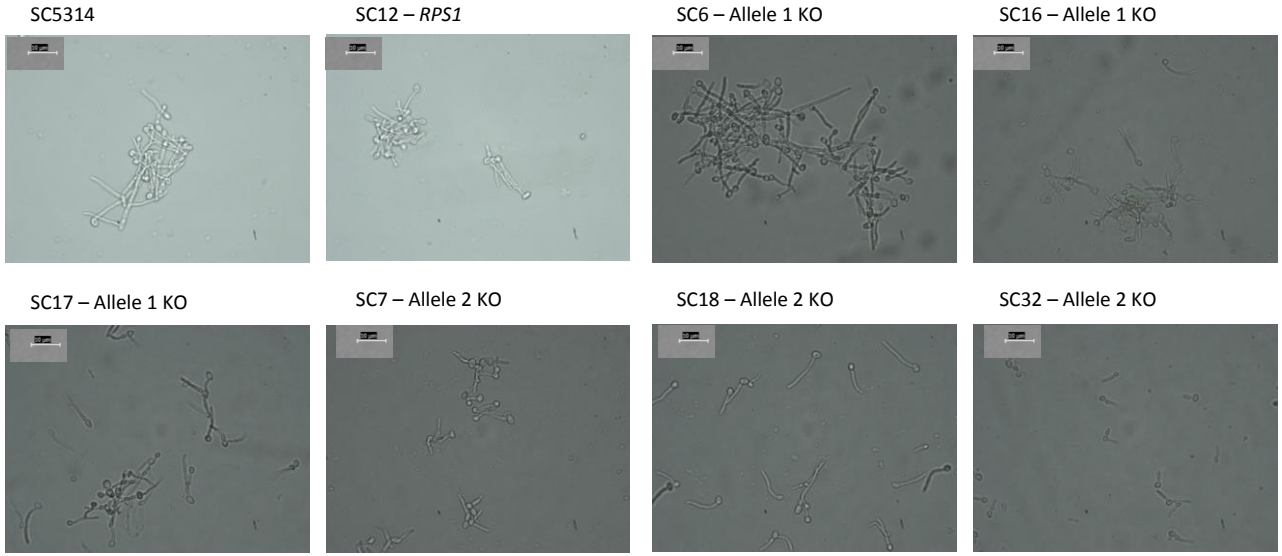


a) Growth rate at 30 °C. b) Growth rate at 37 °C.



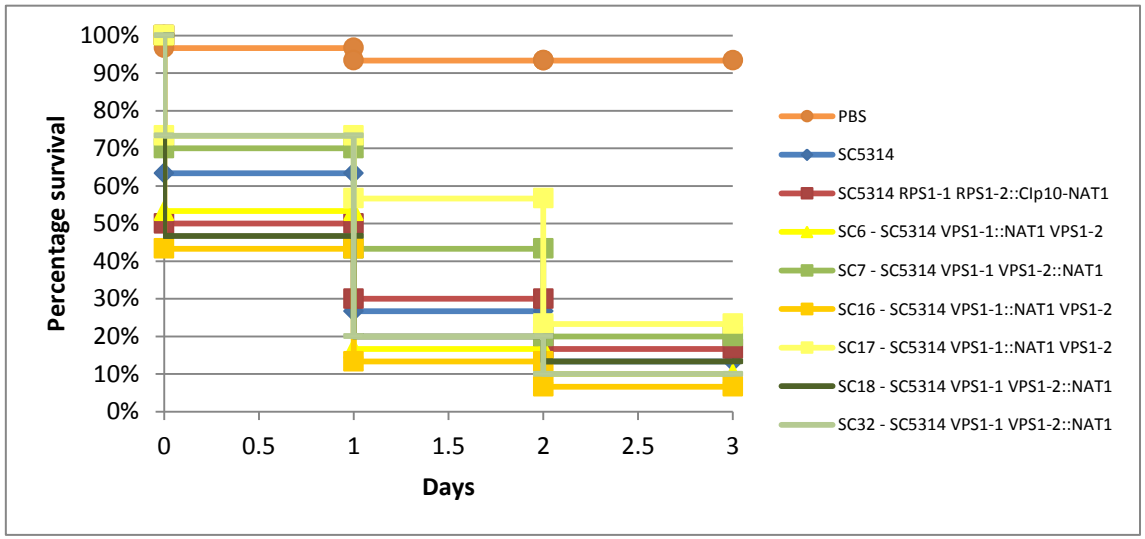
c) Percentage of BECs adhered to *C. albicans* cells. d) Number of *C. albicans* cells per BEC. e) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

f



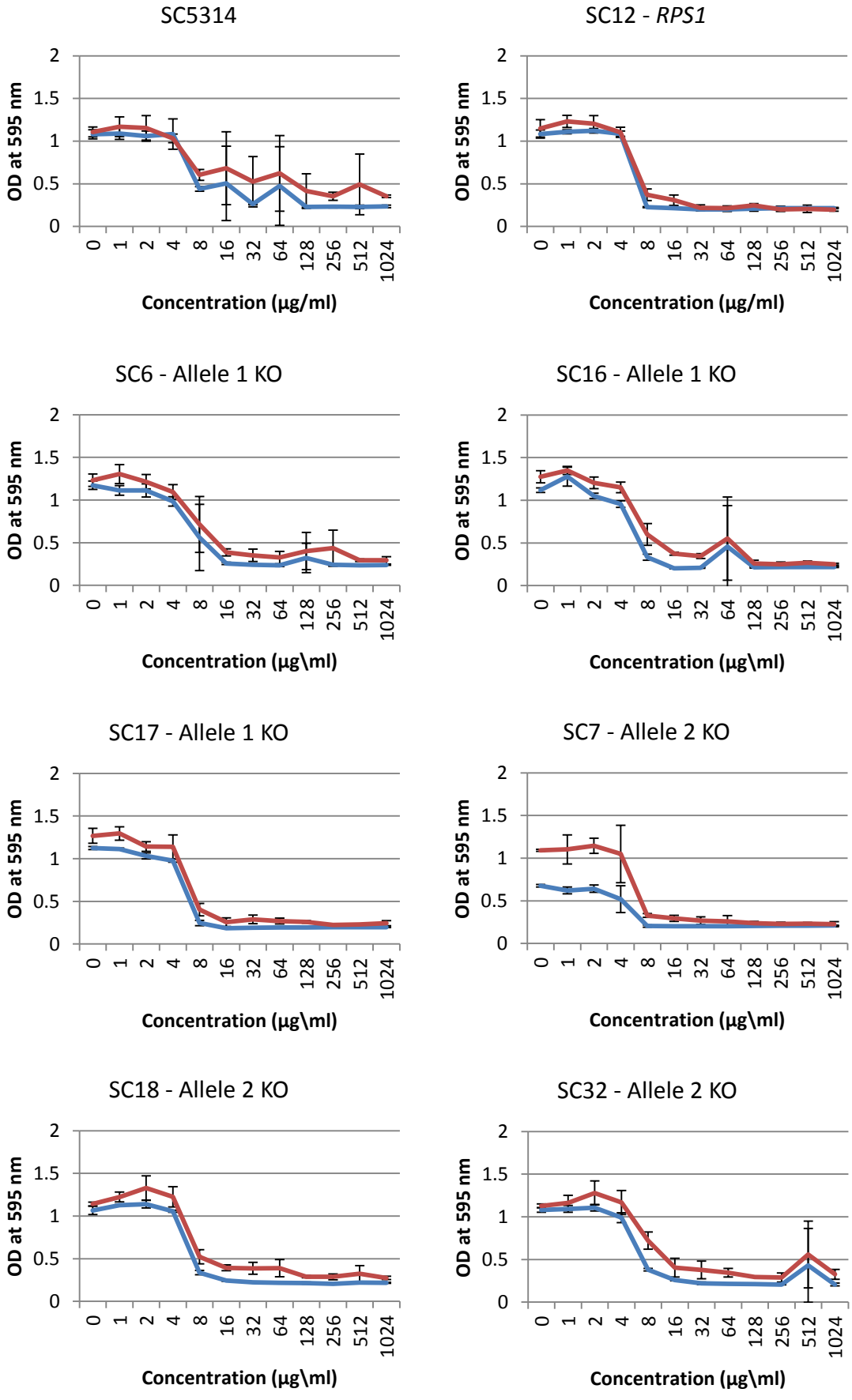
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

g

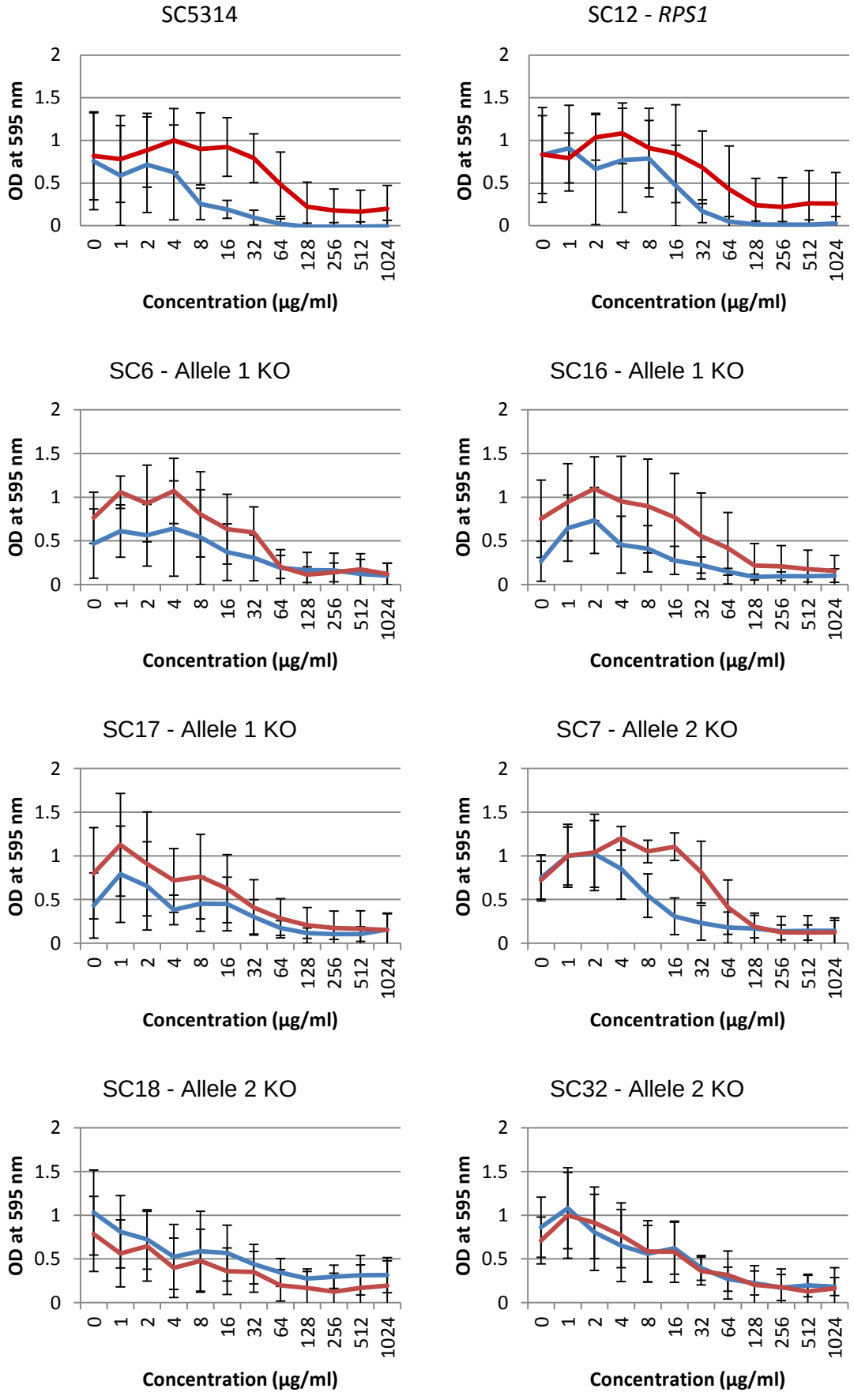


g) *Galleria mellonella* virulence assay at a cell concentration of 2×10^7 cells/ml.

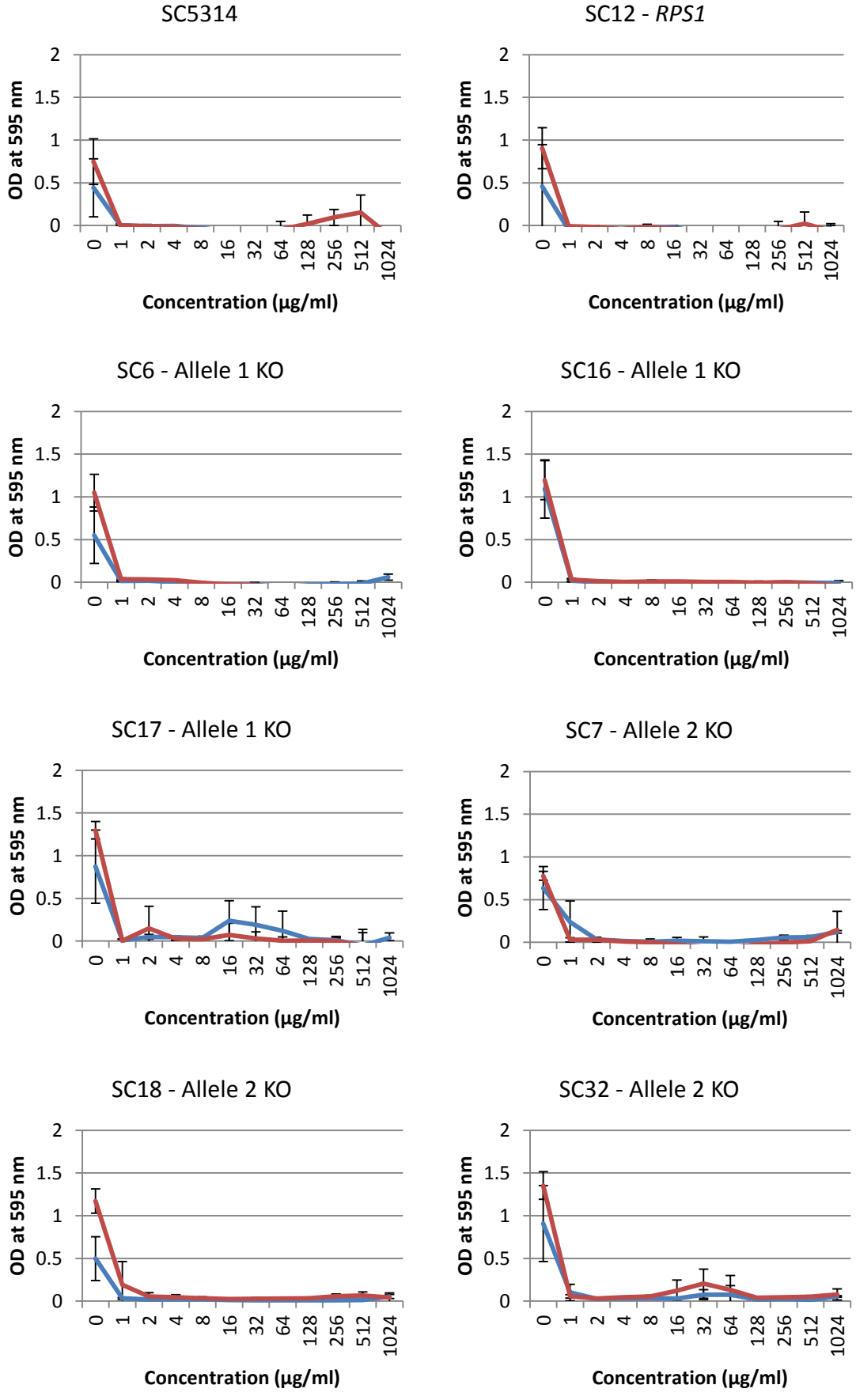
h) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.

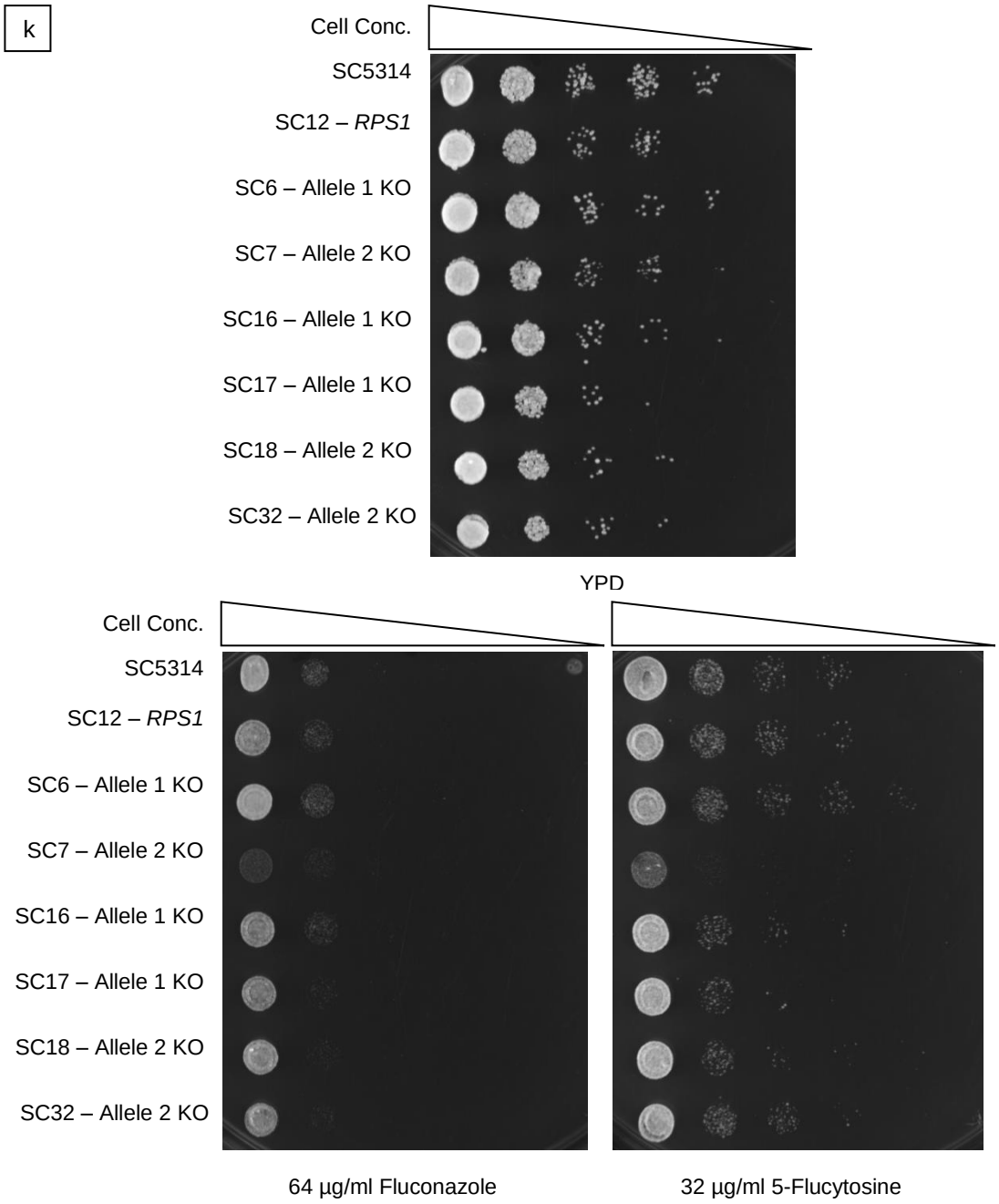


i) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



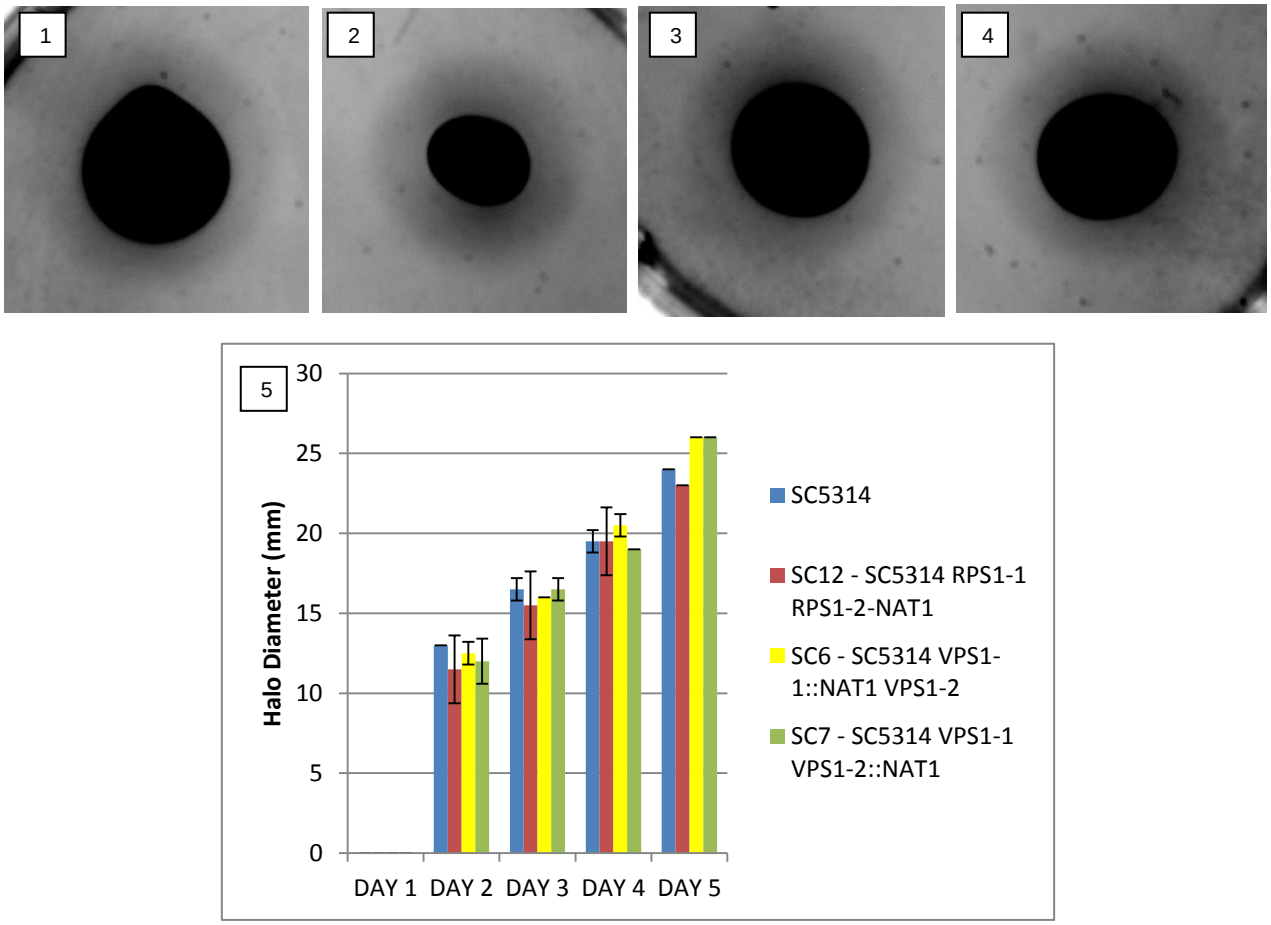
j) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.





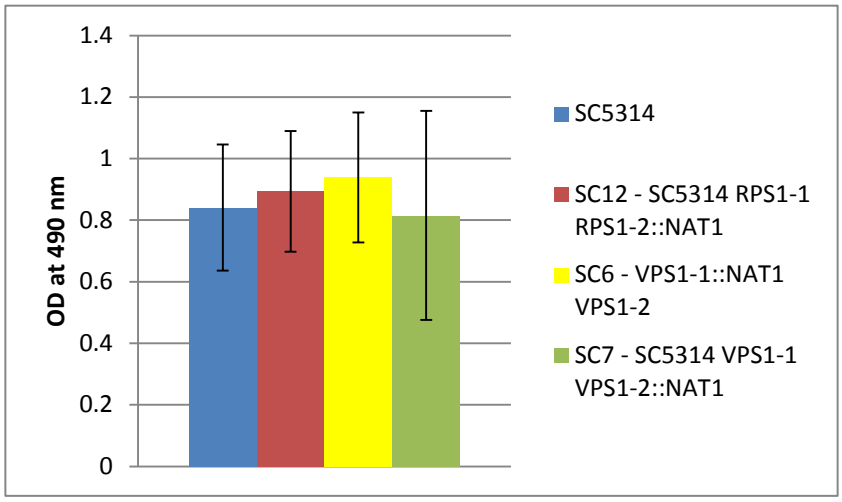
k) Growth on YPD + 64 µg/ml fluconazole agar plates at 24 hours and YPD + 32 µg/ml 5-flucytosine agar plates at 48 hours. YPD agar is present as a control. Cell concentrations range in tenfold dilutions from 1 x 10^7 to 1 x 10^3 cells/ml.

l

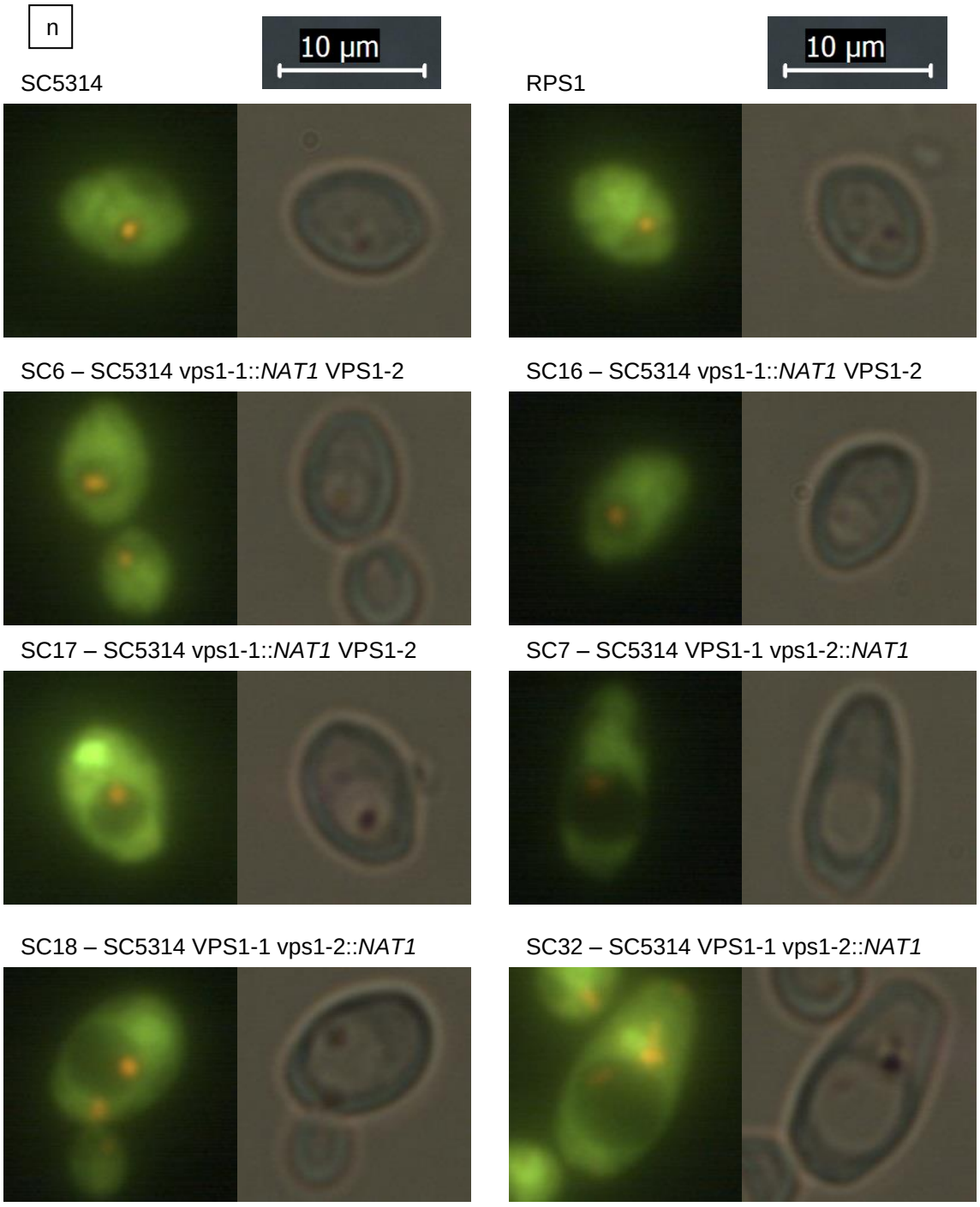


l) Lipase secretion analyses. Representative results for four strains shown here. Halos produced by degradation of egg yolk agar after 7 days for 1) SC5314, 2) SC12, 3) SC6 and 4) SC7. 5) Measurements of the halo diameters every 24 hours. Error bars represent ± one standard deviation.

m



m) Biofilm production. Representative results for four strains at $the 24 hour time point shown here. Error bars represent ± one standard deviation.



n) Vacuole morphology. Right hand side shows vacuole morphology using FUN-1 staining, visualised using fluorescence microscopy with excitation at 470 nm to 590 nm. Left hand side shows the corresponding bright field image. Scale bar = 10 µm.

4.3.2.6 Phenotypic Screening Summary

Table 4.8 summarises the results for all of the phenotypic screening carried out in this chapter.

Table 4.8 Summary of phenotypic screening. Strains with significant results are highlighted with arrows indicating the direction of the results.

GENE	*CDC6*		*ERB1*		*RBT4*		*SMI1*		*VPS1*	
ALLELE KO	1	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
Growth at 30 °C Generation Time										
Endpoint Density	↑SC9 ↑SC12					↑SC4 ↑SC12			↑SC16	↑SC18 ↑SC32
Time to Maximum Inflection										↑SC7 ↓SC32
Growth at 37 °C Generation Time										↑SC7
Endpoint Density			↑SC49	↑SC47						↑SC18 ↑SC32
Time to Maximum Inflection										↑SC7
Adhesion										
Hyphal Induction										
Virulence (at 2 x 10^7 cells/ml)				↓SC48 ↓SC56						
Sensitivity to Fluconazole			↑All at 48 hours		↓All 48 hours					↑SC7 at 24 hours
Sensitivity to 5-Flucytosine (liquid)										↓SC7 ↑SC18 ↑SC32 at 48 hours

Sensitivity to 5-flucytosine (agar)										↑SC7 at 48 hours
Sensitivity to Amphotericin B										
Cell Cycle (DNA content)	↑All									
Biofilm										
Biofilm + Fluconazole										
Biofilm + 5-Flucytosine										
Biofilm + Amphotericin B							↑All growth			
Sensitivity to EtOH and SDS							↑SC33			
Lipase Secretion										
Vacuole Size									↑SC17	↑SC7 ↑SC18 ↑SC32

4.3.3 Errors in the Reference Genome

From the construction of heterozygous knockout mutants, two unusual observations were made. Upon sequence confirmation of the *VPS1* heterozygous knockout mutants, it was noted that five of the SNPs in the reference genome do not exist, and after construction of nine strains, only allele one knockouts were produced for the gene *RCK2.*

These observations were followed up using sequencing methods to compare the gene sequences of *VPS1* and *RCK2* to the reference genome. As heterozygous knockout mutants of both alleles were available for *VPS1,* sequencing of the gene was carried out by colony PCR amplification of the entire length of the gene in six sections from strains SC6 and SC7 (see section 2.11, and for oligonucleotides used see Table 2.7). Samples were sent for Sanger sequencing (see section 2.12) and results were aligned against the reference genome, downloaded from the *Candida* genome database

(http://www.candidagenome.org/) (Inglis *et al.*, 2012). Results confirmed that of the 20 SNPs in the reference genome, five do not exist (Appendix I Figure III). The remaining 15 SNPs only lead to synonymous amino acid changes. Therefore the protein sequences of the two alleles do not differ. Consequently, the unusual phenotypic results seen when screening the knockout strains of allele two, such as the larger vacuole size, are likely to be due to secondary mutations elsewhere in the strains and not due to a difference in function of the two alleles. However as the chitin synthesis gene *CHS7* has been shown to have identical allele sequences with functional differences due to disparity in expression level (Sanz *et al.* 2007), functional differences between the alleles of *VPS1* cannot be ruled out.

A different method was adopted for *RCK2* as only knockouts of allele one were obtained. PCR amplification and sequencing of the wild-type strain SC5314, similar to sequencing *VPS1* in SC6 and SC7, would produce unclear sequencing results with two peaks at SNP locations, and sequencing of the allele one knockout strains would not confirm the sequence of the missing allele. Therefore a conventional cloning technique was used, as described in section 2.8. To summarise, the gene sequence for *RCK2* was amplified using colony PCR from the wild-type strain SC5314, with the assumption that both alleles would amplify with equal efficiency. These PCR fragments were ligated into a pGEM-T easy vector and transformed into *E. coli* cells. Only the sequence from one allele is ligated and transformed into a single *E. coli* cell. Positive transformants were selected based on disruption of the *LacZ* operon and antibiotic resistance. Allele sequences were then amplified via PCR and sent for conventional Sanger sequencing, with a probability of 0.5 that the sequence is allele one and 0.5 that the sequence is allele two. All seven positive transformations came back matching the sequence for allele two. Based on the probability of 0.5 that the sequence is allele two, the probability that all seven strains contained allele two equates to:

$$0.5^7 = 7.8125 \times 10^{-3}$$

With such a small probability of this occurring by chance, and all of the heterozygous knockout strains matching the sequence for allele two

(suggesting that allele one had been removed), it is sensible to assume that *RCK2* is in fact a homozygous gene with an identical sequence for both alleles, matching that of allele two (Appendix I Figure IV).

4.4 Discussion

4.4.1 Is AEI Linked to Functional Differences of Alleles?

The functional consequences of AEI have been investigated here through use of heterozygous knockout mutant strains. From the list of genes with AEI identified in chapter three, potential targets for knockout construction were initially selected based upon a 2x fold difference in allele expression, a minimum allele count of 20 and a pathogenesis-related function. Genes matching 2 out of 3 of these criteria were selected. Heterozygous knockouts of both alleles were produced for five genes. Phenotypic screening under a wide range of general and gene-specific conditions indicated that in most cases the alleles are not functionally distinct, with no sets of heterozygous knockout mutants showing segregation of phenotypic differences from each other or from the wild-type strain SC5314. This is with the exception of *VPS1* which showed some phenotypic differences of some strains, such as a decreased growth rate of SC7 at 37 °C (Figure 4.10b) and unusual vacuolar morphology of a number of strains (Figure 4.10n). However, due to the sequencing results showing that *VPS1* alleles in fact translate to two identical proteins, and with the inconsistencies in the phenotypes of identical isolates, it is unlikely that these phenotypes are due to functional differences in the alleles and are in fact a consequence of secondary mutations elsewhere in the genome. However functional differences cannot be firmly ruled out, as differences in function due solely to expression levels of identical sequences have been observed in the past, as is the case for the previously mentioned gene *CHS7* (Sanz *et al.*, 2007).

Initially this could suggest that AEI is not linked to function, and is present for another purpose which is currently unclear. However, this cannot be claimed with 100% confidence. Of the heterozygous knockout mutants constructed, phenotypic tests were selected based upon pathogenesis related functions or phenotypes identified in previous studies. There is a chance that the alleles do differ in function, but the right phenotypic tests to demonstrate this were not

used. For example, it was not practical to test if the β-glucan amounts in the cell wall of the *SMI1* mutants differed even though it had been shown to be reduced in mutants previously (Hong *et al.*, 1994, Dague *et al.*, 2010). To increase the chances of identifying the most suitable conditions to show differences in allele function, strains could be tested using a high-throughput screen testing an extensive number of conditions such as the screen used by Homann *et al.* (2009) testing a transcription factor knockout library on 55 conditions (Homann *et al.*, 2009). However, this type of investigation was also not feasible during this project but is an area that could be followed in the future. The next chapter will discuss the use of RNA sequencing data collected from *C. albicans* under different conditions to identify if AEI is condition specific, and therefore identify possible conditions to assay to demonstrate differences in allele function. Of the 233 genes identified with AEI in chapter three, functional contributions of the alleles were only investigated for five genes. Attempts were made to knockout two other genes, *RCK2* and orf19.2051, however *RCK2* was later identified as homozygous and no successful strains were ever obtained for orf19.2051 despite numerous attempts at transformation. It could also be the case that if further heterozygous knockout mutants were made, functional differences in the alleles will be identified. Again, this is an area for further investigation.

Variability in the phenotypic assays used is unlikely to be masking phenotypic differences in the heterozygous knockout mutants, but is something should be highlighted from this investigation. Although biological, and often technical, replicates were used for all assays, high levels of variance are still observed. This is particularly clear in measures of growth in response to the antifungal compounds fluconazole and 5-flucytosine, where large error bars are present. Strains were assayed in 96 well plates with two strains per plate. Clear inter plate variation is also observed suggesting false positive interpretations of functional differences between strains, as can be seen for the knockout strains of the gene *VPS1* when grown in 5-flucytosine (Figure 4.10i). The adhesion assay using buccal epithelial cells shows considerable variability between the "adhesiveness" of BECs on samples taken on different days. For example, the average percentage of BECs adhered to wild-type *C. albicans* cells is 44% in the assay using *CDC6* knockouts and 94.9% in the assay using *VPS1* knockouts. This should be considered when comparing results from across

studies. The survival rates of the wax moth larvae used for the virulence assay often varied between batches. Although using three separate cell concentrations aimed to account for this, some *Galleria* did die in the PBS control suggesting that death by natural causes may have some minor impacts upon results.

Another possible explanation as to why no phenotypic differences were seen in the heterozygous knockout mutants is that when one allele is removed, the other allele increases expression levels to compensate for the loss. Functional redundancy is observed within *C. albicans,* especially with regards to gene families, where removal of one gene is compensated for by other genes. This has been reported for the phosphatase gene *PTC6* (Yu *et al.*, 2010), the mannosyltransferase *MNN1* gene family (Bates *et al.*, 2013), and *ALS* genes *ALS2* and *ALS4* (Zhao *et al.*, 2005). Due to a lack of a method which verifies allele-specific expression, as discussed in chapter three, it was not possible to assess if allele expression levels were altered in knockouts and if a mechanism of functional redundancy occurs.

As demonstrated by sequencing of *VPS1* and *RCK2,* errors are present in the diploid reference genome suggesting that some genes are incorrectly annotated as heterozygous. This directly impacts upon the initial identification of genes with AEI, as measurements of allele expression are based upon reads which align to SNP positions. If these SNPs are errors, reads will be aligning to the incorrect position in the genome, producing incorrect allele counts. In the future, it would be beneficial to verify the heterozygosity of all of the genes identified as having significant levels of AEI to show that the AEI is not a false positive due to incorrect alignment of reads at sequencing errors. Alternatively, re-alignment of reads to the recently published phased diploid reference genome may also improve estimations of AEI (Muzzey *et al.*, 2013). Within this paper, evidence was shown that the reliability of AEI measurements were significantly improved using the new reference genome.

4.4.2 Conclusion

To conclude, the functional consequences of AEI are still unclear after phenotypic screening of heterozygous knockout mutants. Although no functional

differences in alleles were identified here, the correct conditions or gene to show this may have been overlooked. Conversely, AEI may not be linked to function and could in fact be due to errors in the initial identification of genes with AEI due to errors in the reference genome. However, sequencing errors were not present during verification of all heterozygous knockout strains and therefore this cannot explain the lack of functional differences in all cases.

Chapter 5: AEI in Different Growth Conditions

5.1 Introduction

The identification of allelic expression imbalance in chapter three and the subsequent investigations of the functional consequences associated with this phenomenon in chapter four have so far been based upon a single RNA sequencing data-set. However, changes in gene expression due to a shift in growth conditions occurs readily in *C. albicans* (Enjalbert *et al.*, 2003, Bensen *et al.*, 2004, Fradin *et al.*, 2005, Biswas *et al.*, 2007, Nobile *et al.*, 2012, Tierney *et al.*, 2012). Therefore it would be sensible to infer that levels of AEI may also be responsive to the growth environment. Identification of changes in AEI responses may therefore shed some light upon the functional purpose of AEI itself. To investigate this hypothesis, RNA sequencing data-sets obtained from *C. albicans* grown under different conditions were examined from both in house experiments and public databases. A computational pipeline was developed to identify levels of AEI from this data, and the condition-specific response of AEI was investigated. Heterozygous knockout mutants of genes indicating condition specific allelic expression responses were then phenotypically screened to further investigate the functional impact of allelic expression imbalance.

5.1.1 Condition-Specific Gene Expression in *Candida albicans*

Extensive evidence is present to demonstrate that *Candida albicans* gene expression levels alter in response to environmental cues. As discussed in section 1.1.1, a complex network of signalling pathways centred on the MAPK pathway and the cAMP pathway are responsible for the up-regulation of hypha specific genes leading to the morphological switch from a yeast to a hyphal form following exposure to a number of different growth conditions (Biswas *et al.*, 2007). The genome-wide transcriptional response of this morphological transition has been elucidated through use of microarrays, which identified key genes involved in this response (Nantel *et al.*, 2002). Additionally, six key regulator genes have been identified in the control of biofilm formation, which

consequently alter the expression level of approximately 1000 genes (Nobile *et al.*, 2012), as discussed in section 1.1.4.

As well as morphological transitions, gene expression levels in *C. albicans* have shown to be altered under growth in differing environments, especially those inducing a stress response. A shift in temperature from 23 °C to 37 °C induces the expression of heat shock proteins *HSP12, HSP70, HSP78* and *HSP104* (Enjalbert *et al.*, 2003). Hyperosmotic stress causes an increase in the expression of *ENA1, GPP1* and *GPD2* to protect the cell from increases in ionic strength (Enjalbert *et al.*, 2003). Both acidic and alkaline pH induce changes in gene expression, with 514 genes having been identified as pH responsive (Bensen *et al.*, 2004). Oxidative stress leads to an increase in expression of the glutathione reductase *TTR1* and the thioredoxin gene *TRX1* (Enjalbert *et al.*, 2003). A core set of stress response genes, which alter their transcriptional profile under osmotic, heavy metal and oxidative stress, have also been identified (Enjalbert *et al.*, 2006). Growth of the closely related species *C. parapsilosis* under hypoxic conditions has been associated with an increase in expression of genes involved with ergosterol biosynthesis and carbohydrate metabolism, and a decrease in expression of genes involved in cellular respiration and the tricarboxylic acid cycle (Guida *et al.*, 2011). However, the most comprehensive investigation of the condition-specific transcriptional response in *C. albicans* was published by Bruno *et al.* in 2010. Here RNA sequencing was used to identify the gene expression patterns of the wild-type strain SC5314 when grown under nine different conditions, including hypha specific conditions, oxidative stress, nitrosative stress and cell wall damaging conditions. The results were taken forward and used to identify novel transcripts which are regulated in a condition specific manner, and used to uncover novel functions and pathways of previously annotated genes (Bruno *et al.*, 2010).

The transcriptional response under conditions closely resembling *in vivo* infections have also demonstrated that gene expression levels respond to environmental cues that would occur naturally. For example, 545 genes have been found to alter expression levels when cells are co-cultured with mouse dendritic cells (Tierney *et al.*, 2012), upon phagocytosis by mammalian dendritic cells, gene expression shifts to increase gluconeogenic growth and fatty acid

degradation (Lorenz *et al.*, 2004, Fernández-Arenas *et al.*, 2007), and resistance to fungal clearance by neutrophils requires gene expression profiles that resemble growth in carbohydrate starvation, nitrosative and oxidative stress conditions (Fradin *et al.*, 2005, Miramón *et al.*, 2012).

Although condition-specific shifts in AEI are yet to be reported, it is sensible to infer from the above examples that this could occur during growth both under *in vitro* conditions and during the infection process.

5.1.2 Computational Tools Available for Identifying AEI

The identification of AEI from RNA sequencing data in chapter three used the software package CLCBio software (www.clcibio.com). However this is a closed source software package, which is no longer available for use in this project. Therefore a new computational tool to identify AEI needed to be devised. Various different software packages have been previously developed to identify levels of AEI using RNA sequencing data. However, most of these programmes are designed for use with human sequencing data, a haploid reference genome, or full genomes for both parental strains in the case of diploid hybrid offspring, most of which are unavailable for *Candida albicans* (Rozowsky *et al.* 2011; Turro *et al.* 2011; Krueger, 2012; Pandey *et al.* 2013). Although a haploid reference genome is available for *C. albicans*, significant biases have been observed when identifying AEI using this methodology (Degner *et al.*, 2009, Stevenson *et al.*, 2013) as discussed in section 3.1.

As mentioned above, it has been reported that bias can occur when mapping reads against a haploid reference genome with an unequal success rate of read mapping producing a skewed number of reads towards the reference (Stevenson *et al.*, 2013). The software package Allim (Allelic imbalance metre) (Pandey *et al.*, 2013) has been developed to overcome this mapping bias by initially constructing a polymorphism-aware reference genome, then mapping bias is estimated using a sequence-specific simulation tool, followed by a G-test which corrects the alignment for this bias (Pandey *et al.*, 2013). Unfortunately, this software was not suitable for use here as it assumes the diploid strain is a hybrid of two parental strains and therefore requires full genome or transcriptome data (with full phasing) of both parents.

AlleleSeq has been developed for use with the human reference genome and also aims to overcome mapping bias seen when using a haploid reference genome (Rozowsky *et al.*, 2011). In a similar way to Allim, Alleleseq constructs a “personal” diploid reference genome, with maternal and paternal haplotypes. This is achieved using fully phased variant information obtained from the 1000 genomes consortium alongside “equivalence maps” listing base-pair locations for both parental genomes. Reads are then aligned against the maternal and paternal haplotypes and quantified at SNP locations to give allele-specific counts (Rozowsky *et al.*, 2011). Again, this level of phased variant information is unavailable for *C. albicans,* and equivalence maps are not available as there are not two separate parental genomes, making Alleleseq unsuitable for use here.

MMSEQ aims to identify haplotype specific isoforms and takes a similar approach to Allim and Alleleseq, but includes an initial alignment of RNA sequencing reads against a haploid reference genome using TopHat (Turro *et al.*, 2011). Variants are then called from this alignment and phased using population level genotype data before editing of the reference genome into a haplotype specific reference and realignment of the reads to this reference using Bowtie (Turro *et al.*, 2011). As this software uses a haploid reference with variant calling, it is subjected to the mapping biases mentioned above. Additionally, population level genotype data is unavailable for *Candida albicans* and therefore, for these reasons, MMSEQ was not suitable for use in this investigation.

The final piece of computational software available is ASAP (Allele-specific alignment pipeline). ASAP has been developed for identification of allele-specific expression in heterozygous individuals where reads are aligned to two separate haploid reference genomes (Krueger, 2012). As only one reference genome is available for *C. albicans,* this software is also unsuitable for use.

Due to these reasons, development of a new computational pipeline is detailed in this chapter to identify AEI in *Candida albicans* using alignment against the diploid reference genome and avoiding the use of variant calling. This pipeline

can then be applied to a number of RNA sequencing data-sets to evaluate if AEI changes in response to environmental cues.

5.1.3 Aims of this Chapter

This chapter aims to achieve three research objectives:

1. Development of a computational pipeline to identify AEI from RNA sequencing data.
2. Use of RNA sequencing data to identify how AEI changes in a condition-specific manner.
3. Construction and phenotypic screening of heterozygous knockout mutants to further investigate the functional consequence of AEI.

5.2 Materials and Methods

5.2.1 Computational Pipeline

The computational pipeline developed here works around a modified version of standard RNA sequencing analysis. Figure 5.1 details the pipeline in a step-by-step manner. Initially, the analysis was repeated on the data-set collected in chapter three (see section 3.2.1 for methodology) to compare computational analyses.

Raw sequencing reads were filtered to improve quality, removing sequencing adaptors and bases with a phred value of less than 20 using the software Fastq-mcf (Aronesty, 2011). Filtered reads were then aligned against the diploid reference genome (Jones *et al.*, 2004) (indexed using BWA index (Li and Durbin, 2009)) using Bowtie (Langmead *et al.*, 2009). To replicate the original CLCBio analysis as closely as possible, parameters were set to allow up to two mismatches (see section 3.2.1.4).

Bowtie was selected as the alignment software for a number of reasons. Due to an algorithm that combines the Burrows-Wheeler transformation and a suffix array, Bowtie converts the reference genome to an FM-index. This transforms large reference genomes to a manageable size, increasing the efficiency of short-read mapping (Berger *et al.*, 2013). Bowtie is also adopted by the allele-specific software MMSEQ (Turro *et al.*, 2011) and has been used in previous studies identifying AEI (Li *et al.*, 2012), including the investigations of AEI in

Candida albicans (as discussed in section 1.8) (Muzzey *et al.*, 2013, Muzzey *et al.*, 2014)

Alignments were then converted from SAM to BAM file formats and sorted using the SAMTools package (Li *et al.*, 2009). The numbers of reads aligned at every base-pair location throughout the genome were then identified using SAMTools mpileup (Li *et al.*, 2009).

To quantify allele-specific expression levels, reads aligning at SNP locations were quantified. To achieve this, all of the SNP locations were first identified using a custom Perl script (Appendix II.II), which takes the list of allele pairs and the diploid reference genome as the input. Alignments of each allele pair for a gene were carried out using MUSCLE (Edgar, 2004) and SNP locations were recorded. Only SNP locations were used and INDEL locations were ignored to prevent a bias in genes identified with AEI that differ in length. This is discussed in more detail in section 5.3.1. The mpileup file listing all of the reads across the genome was then filtered to list just the reads at SNP locations using a custom Perl script written with the help of Paul O'Neill (Appendix II.III). Again, a custom Perl script (Appendix II.IV) was used to total the number of reads aligned to each allele.

From here, three statistical measures were implemented to identify genes with significant levels of AEI. Three tests were chosen to ensure that the pipeline was stringent and the number of false positive results are minimal. The choice to use multiple statistical tests also reflects the lack of consistency in methods which identify AEI as discussed in section 1.5.2. Raw allele counts of each gene were paired using a custom Perl script (Appendix II.V) and then statistically analysed using DESeq (Anders and Huber, 2010) implemented in R (The R Foundation for Statistical Computing, 2010). DESeq works on the assumption of an underlying negative binomial distribution and has been shown to be a powerful tool for analysing differential expression (Anders and Huber, 2010). Alleles of a single gene with a p value of less than 3.67107×10^{-5} (set using Bonferroni correction) were identified as having statistically significant allelic expression imbalance.

To reproduce the original analysis, raw counts for each allele were also normalised to RPKM values (reads per kilobase per million mapped reads) (Mortazavi *et al.*, 2008) using a custom Perl script (Appendix II.VI). RPKM values for each gene were paired using a custom Perl script (Appendix II.V) and statistically compared in two ways: to replicate the original analysis, a Fisher's Exact test was used, and to replicate previous investigations of AEI a chi-square test was used (Gregg *et al.*, 2010, Tuch *et al.*, 2010a, Zhang *et al.*, 2011). In both cases, alleles of a single gene with a p value of less than 3.67107×10^{-5} (set using Bonferroni correction) were identified as having statistically significant allelic expression imbalance.

Only genes identified as having significant levels of AEI by all three statistical tests and within all replicates of the experiment were taken forward as genes with AEI.

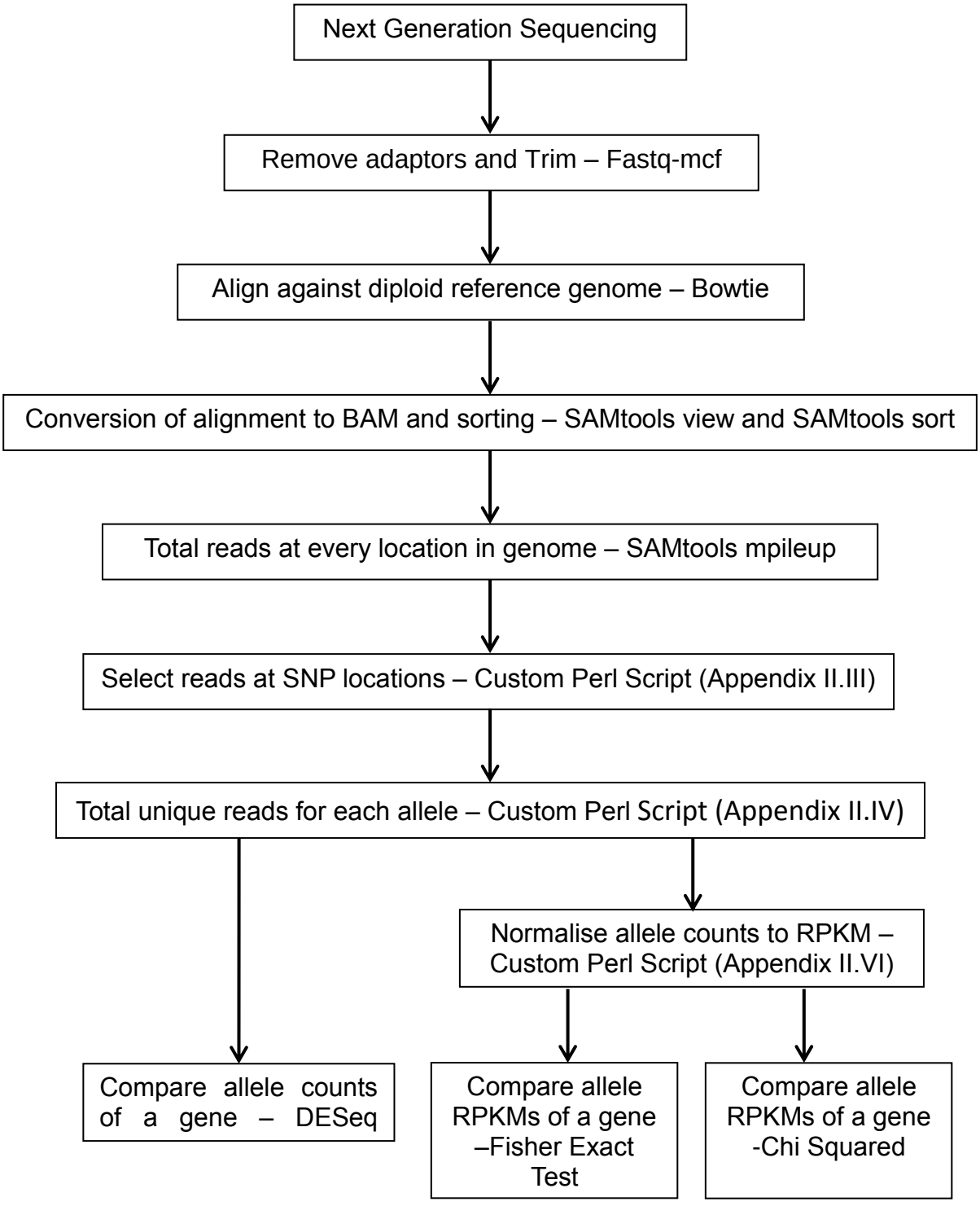


Figure 5.1 Computational pipeline devised to identify allelic expression imbalance from RNA sequencing data.

5.2.2 Acquisition of RNA Sequencing Data from Cells Grown Under Different Conditions

To assess if condition specific patterns of allelic expression imbalance occur, RNA sequencing data-sets from *C. albicans* grown under different conditions was obtained and analysed with the pipeline described above.

5.2.2.1 Data from Bruno *et al.* (2010)

As described in section 5.1.1, Bruno *et al.* (2010) carried out RNA sequencing on *C. albicans* grown under different conditions. Methodologies describing sample preparation and sequencing can be found in Bruno *et al.* (2010). It is important to note, that in the Bruno *et al.* study, short read, 30 bp single end sequencing was carried out using an Illumina GAII sequencer. The raw sequencing reads for this paper were downloaded from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under the accession number SRA020929. Table 5.1 details the runs downloaded and analysed for each condition. If more than one run was available for a replicate, these were concatenated into a single file before processing via the computational pipeline.

5.2.2.2 Data from Co-Culture of *C. albicans* with *Streptococcus gordonii*

In collaboration with Professor Howard Jenkinson and Dr. Lindsay Dutton from the University of Bristol, RNA samples were collected from *C. albicans* hyphae co-cultured with the oral bacterium *Streptococcus gordonii.* Interactions between these species are often observed *in vivo* during oral candidiasis infections (Wright *et al.*, 2013). RNA samples were taken in triplicate from *C. albicans* cell harvests grown under three conditions: alone in hypha-inducing conditions (37 °C in glucose) for two hours, in hypha-inducing conditions (37 °C with glucose) for two hours before co-culture with *S. gordonii* for one hour in a 2:1 ratio, and in hypha-inducing conditions (37 °C with glucose) for two hours before culture alone in *S. gordonii* media for one hour.

Total RNA was extracted using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions for yeast, using mechanical disruption and including on-column DNase digestion using RNase-free DNase (Qiagen). RNA concentration and quality was quantified using a NanoDrop Spectrophotometer

and formaldehyde agarose gel electrophoresis (see section 3.2.6.1.3). Ribosomal RNA was depleted using a RiboZero Magnetic Gold Kit (Epicentre) and Illumina sequencing libraries were prepared using ScriptSeq v2 (Epicentre). The quality and quantity of each library was analysed using an Agilent 2100 Bioanalyser (samples prepared according to the manufacturer's instructions). 100 bp paired-end sequencing was then performed using the Illumina HiSeq2500 platform in high output mode using Truseq v3 reagents (Illumina).

Table 5.1 Sequencing Runs Downloaded and Analysed from the Bruno *et al.* (2010) paper

Condition	Replicate	Run Number
YPD	Replicate 1	SRR060099
		SRR060100
		SRR060101
	Replicate 2	SRR060102
		SRR060124
		SRR060125
		SRR060126
Hyphae Inducing (YPD + 10% foetal calf serum)	Replicate 1	SRR060087
		SRR060088
	Replicate 2	SRR060089
		SRR060090
		SRR060091
		SRR060092
Tissue Culture Media pH 4 (M199 pH 4)	Replicate 1	SRR060127
	Replicate 2	SRR060128
Tissue Culture Media pH 8 (M199 pH 8)	Replicate 1	SRR060129
	Replicate 2	SRR060130
High Oxidative Stress (YPD + 5 mM Hydrogen Peroxide)	Replicate 1	SRR060131
	Replicate 2	SRR060134
Low Oxidative Stress (YPD + 0.5 mM Hydrogen Peroxide)	Replicate 1	SRR060135
	Replicate 2	SRR060136
No Oxidative Stress Control (YPD)	Replicate 1	SRR060143
	Replicate 2	SRR060144
Cell Wall Damaging (YPD + 100 µg/ml Congo Red)	Replicate 1	SRR060146
	Replicate 2	SRR060145
No Cell Wall Damage Control (YPD)	Replicate 1	SRR060148
	Replicate 2	SRR060147
Nitrosative Stress (YPD + 1 mM Dipropylenetriamine Nonoate dissolved in 10 mM NaOH)	Replicate 1	SRR063952
	Replicate 2	SRR063953
No Nitrosative Stress Control (YPD + 10 mM NaOH)	Replicate 1	SRR063986
	Replicate 2	SRR063987

5.2.3 Calculation of Allele Lengths

Allele lengths were obtained from the 'Chromosomal Features File' available from the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012) for "allele one" and calculated manually for "allele two" based on chromosomal coordinates.

5.2.4 Gene Ontology (GO) Analysis

Analysis for over representation of Gene Ontology (GO) terms within genes identified to have AEI was carried out using "CGD GO Term Finder" at the *Candida* genome database (www.candidagenome.org) (Inglis *et al.*, 2012). Lists of GO terms were created using "CGD Gene Ontology Slim Mapper" also available from the *Candida* genome database.

5.2.5 Calculating Differences in Promoter Sequences

As in chapter three, the promoter regions of differentially expressed alleles were compared to assess if differences in this region were linked to differences in expression. As promoter sequences in *C. albicans* are currently undefined, the 1000 bp of DNA sequence upstream of each allele were downloaded from the *Candida* Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012). If the neighbouring open reading frames were within the 1000 bp upstream, the sequence up to the neighbouring ORF was taken. Promoter sequences of each pair of alleles of a gene were aligned used ClustalW (http://www.genome.jp/tools/clustalw/) (Kyoto University Bioinformatics Center, 2010) and sequences were recorded as different if one of more SNPs or INDELs were observed. The probability of observing SNPs across a region of 1000 bp was calculated based upon the observed average level of heterozygosity across the genome of one SNP in every 237 bp (Jones *et al.*, 2004). The observed and expected number of promoter sequences with SNPs was compared statistically using a chi-square test.

5.2.6 Identification of Genes with Uneven Changes in Allele Expression Levels between Growth Conditions

Despite a lack of significant AEI, as determined by the statistical tests described in section 5.2.1, for some genes the ratio of allele one expression to allele two expression will differ significantly when growth conditions change. To identify these genes where the growth condition causes a change in expression of one allele greater than the change in the other allele, the fold difference in RPKM values (see section 5.2.1) of each allele between conditions was calculated for every gene, for 16 different condition comparisons taken from the analysis of the Bruno *et al.* (2010) data (Table 5.2). The Log_{10} fold change in allele one expression level was then plotted against the Log_{10} fold change in allele two expression level for each comparison and linear trend lines were fitted. The equation of the linear trend line was used to calculate the predicted y values. Any genes where the observed y value was greater than two standard deviations away from the predicted y value were identified as having a significantly larger change in one allele expression than the other allele between those growth conditions. To identify genes which frequently have a disparity in the fold difference of alleles between conditions, the frequency of genes identified as significant for each condition comparison was calculated and multiplied for each gene to give a significance measure. All genes with $p < 1.95 \times 10^{-4}$, as determined by Bonferroni correction for multiple comparisons, were identified as having alleles whose changes in expression levels differ significantly in response to a change in growth condition.

Table 5.2 Condition Comparisons Used to Identify Genes with Uneven Differences in Allele Expression Levels between Conditions

Condition One	Condition Two
YPD	Serum
YPD	Congo Red
YPD	No Congo Red (YPD)
Congo Red	No Congo Red (YPD)
YPD	High Oxidative Stress
YPD	Low Oxidative Stress
High Oxidative Stress	Low Oxidative Stress
YPD	No Oxidative Stress (YPD)
Low Oxidative Stress	No Oxidative Stress (YPD)
High Oxidative Stress	No Oxidative Stress (YPD)
YPD	M199 pH 4
YPD	M199 pH 8
M199 pH 4	M199 pH 8
YPD	Nitrosative Stress
YPD	No Nitrosative Stress (YPD + NaOH)
Nitrosative Stress	No Nitrosative Stress (YPD + NaOH)

5.2.7 Heterozygous Knockout Mutant Construction

Heterozygous knockout mutants of *ADH2, GPX1, RPS7A* and orf19.5648 were constructed as described in sections 2.8 to 2.11. For a full list of strains refer to Table 2.1.

5.2.8 Phenotypic Screening

For the methods describing general phenotypic assays used here and in chapter four, see section 2.14. Listed below are the phenotypic screens used in just this chapter.

5.2.8.1 Growth with Ethanol as the Sole Carbon Source

Conditions for the growth of *ADH2* heterozygous knockout strains on solid media containing ethanol as the sole carbon source is detailed in section 2.14.4.

For growth in liquid media, the growth rate of *ADH2* heterozygous knockout strains in ethanol as the sole carbon source was measured using a liquid assay in a 96-well plate format. A single colony of the appropriate strain was grown overnight in either 10 ml of YPD or 10 ml of YPE (2% (w/v) Bacto-peptone, 2% (v/v) ethanol, 1% (w/v) yeast extract) at 30 °C, 180 rpm. 10 µl of this culture was taken and diluted into 1 ml of fresh YPE. 100 µl of the cell suspension was then transferred to a single well of a 96 well plate. Each strain was plated in technical quadruplicate and sterile YPE was used as a control. The experiment was then carried out in biological duplicate.

Optical density at 650 nm was measured every 3½ minutes using a kinetic read microplate spectrophotometer held at 30 °C for a total of 48 hours (Molecular Devices VersaMax Microplate Reader). Plates were shaken for three minutes in between reads.

5.2.9 Cloning of Genes

Cloning of *RPS7A* and orf19.5648 to confirm allele sequence polymorphisms was carried out as described in section 2.8.

5.3 Results

5.3.1 Comparison of Computational Pipelines

To assess the validity of the computational pipeline designed to identify AEI, the RNA sequencing data collected in chapter three was reanalysed with the new pipeline (see section 5.2.1) and the results were compared. Initially, the computational pipeline totalled reads for each allele at both SNP and INDEL locations. In total, 175 genes were identified as having significant AEI from all three statistical tests in all three replicates. However 168 of these genes showed a difference in the length of the alleles, with 159 of these genes showing higher expression levels of the longer allele. Although gene length has been shown to impact upon expression levels (as discussed in chapter three), the relationship is well established as being the opposite of what is shown here, with the shorter genes showing increased transcriptional efficiency and higher expression levels (Coghlan and Wolfe, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003). Therefore, it is proposed that although allele counts are

normalised for length using RPKM, the pipeline devised has biases surrounding alleles which differ in length, producing a high number of false positive results. Biases for length during the sequencing process which are not sufficiently corrected for using RPKM have been previously reported (Bullard *et al.* 2010). This may be due to the bias seen in cDNA library preparation which leads to a higher coverage of reads at the 3' end of a transcript (as discussed in section 3.4.3.4) (Wilhelm *et al.*, 2008), with alleles that differ in length at the 3' end being the most susceptible to this bias.

To correct for this, the pipeline was revised and reads aligning at INDEL locations were removed from the allele specific counts. This may lead to some false negative results, missing genes with AEI which only differ in length. However, alleles which just differ in length are more likely to be a consequence of annotation errors during the construction of the reference genome, making screening these genes out of the identification for AEI advantageous.

Repeating the analysis of the RNA sequencing data collected in chapter three, using the SNP locations alone, identified 24 genes with AEI from replicate one, 29 genes with AEI from replicate two and 26 genes with AEI from replicate three. Of these genes, a total of 21 were identified in all three replicates (Figure 5.2 and Table 5.3). Two thirds (14) of these genes were also identified in the analysis carried out in chapter three, including *RCK2* and *RBT4* which were taken forward for heterozygous knockout construction in chapter 4. Length differences between the alleles were also closer to what was expected with 9 of the 21 genes differing in length, of which only 4 showed higher expression of the longer allele, against the trend that has been previously observed in yeast.

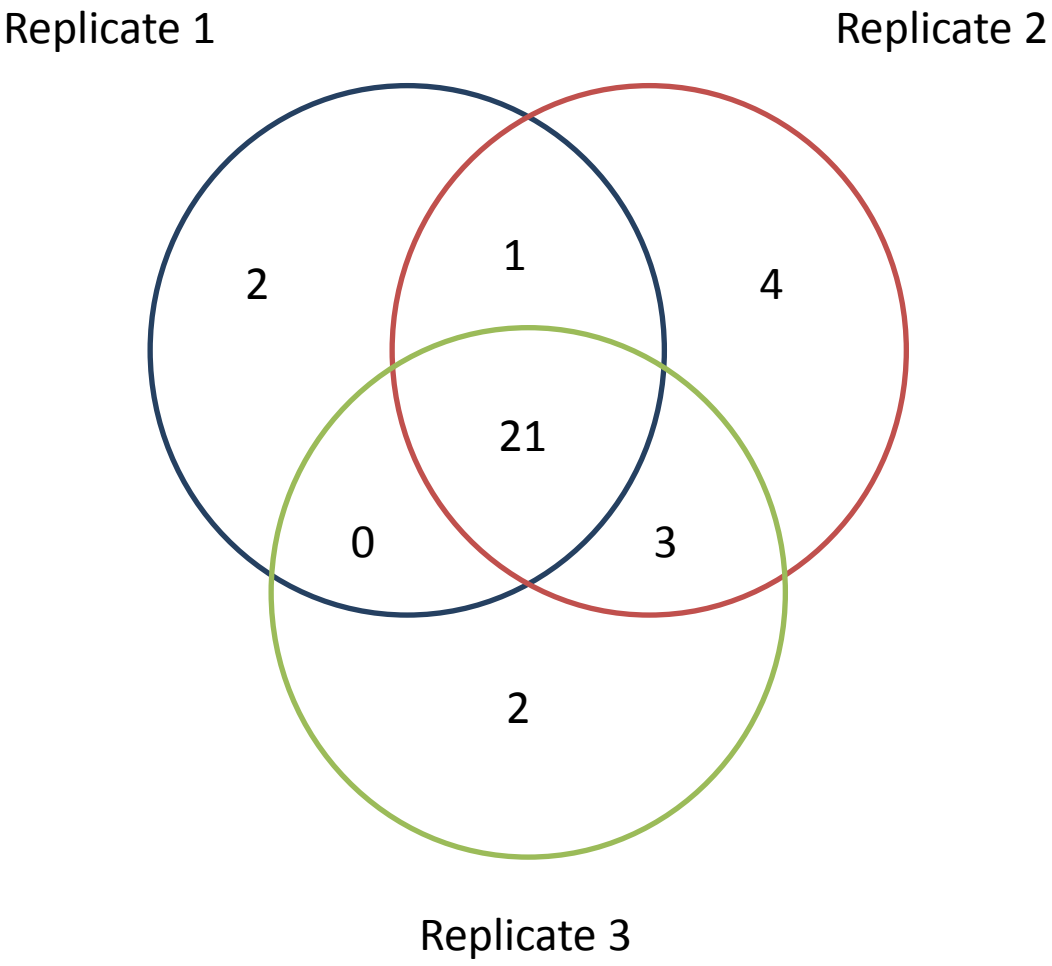


Figure 5.2 Number of genes identified with AEI using the new computational pipeline. A total of 21 genes with AEI were identified in *C. albicans* cells grown in YPD at 30 °C from the three RNA sequencing replicates isolated in chapter three.

Table 5.3 Genes identified with AEI in *C. albicans* cells grown in YPD at 30 °C using the new computational pipeline. Counts and RPKM values for each replicate are detailed.

Allele 1	Count Replicate 1	Count Replicate 2	Count Replicate 3	RPKM Replicate 1	RPKM Replicate 2	RPKM Replicate 3	Allele 2	Count Replicate 1	Count Replicate 2	Count Replicate 3	RPKM Replicate 1	RPKM Replicate 2	RPKM Replicate 3	Average Fold Difference
orf19.10681	66	134	123	4.58	7.53	7.51	orf19.3171	556	655	666	38.54	36.82	40.66	6.24
orf19.11687	26887	45984	46492	1578.35	2188.57	2403.68	orf19.4212	124	352	294	7.20	16.57	15.03	170.44
orf19.11957	1480	1296	928	159.37	113.14	88.01	orf19.4476	97	137	149	10.23	11.72	13.84	10.53
orf19.12579	525	356	233	54.75	30.10	21.40	orf19.5113	1	1	0	0.10	0.08	0.00	-
orf19.13093	41	70	38	7.50	10.38	6.12	orf19.5648	2789	3256	2761	500.04	473.30	435.97	61.17
orf19.13163	255	380	299	7.36	8.89	7.60	orf19.5741	2782	4509	4031	80.30	105.51	102.47	12.09
orf19.13213	2526	2750	1882	248.48	219.32	163.05	orf19.5791	13	5	1	1.28	0.40	0.09	875.44
orf19.1357	19	50	34	1.34	2.86	2.12	orf19.8937	279	472	348	26.65	36.56	29.28	15.48
orf19.13891	2	1	2	0.44	0.18	0.39	orf19.6538	284	329	282	63.03	59.20	55.12	204.00
orf19.14144	1	3	5	0.10	0.24	0.43	orf19.6854	3327	4152	3697	169.51	171.52	165.90	938.50
orf19.1915	0	4	1	0.00	0.21	0.06	orf19.9471	1890	2875	2806	125.53	154.81	164.13	-
orf19.2268	47	90	68	2.90	4.50	3.69	orf19.9808	748	1287	1191	46.14	64.37	64.71	15.91
orf19.4212	124	352	294	7.20	16.57	15.03	orf19.11689	1144	1117	1326	66.62	52.74	68.01	5.65
orf19.4213	12	13	32	0.70	0.61	1.64	orf19.11689	1144	1117	1326	66.62	52.74	68.01	74.23
orf19.465	919	1782	1592	35.51	55.82	54.17	orf19.8096	13	66	61	0.50	2.067	2.08	41.26
orf19.4959	388	732	664	24.60	37.63	37.08	orf19.12424	3	8	7	0.19	0.41	0.39	105.23
orf19.5602	624	575	475	35.10	26.22	23.53	orf19.13045	16	4	2	1.16	0.24	0.13	108.47
orf19.6202	35	118	98	3.55	9.70	8.75	orf19.13583	301	806	710	30.51	66.25	63.39	7.56
orf19.8644	214	305	307	27.52	31.80	34.77	orf19.1042	951	764	770	122.31	79.66	87.22	3.15
orf19.9267	1955	261	289	380.50	41.19	49.53	orf19.1700	10	2	1	1.95	0.32	0.17	205.00
orf19.9571	14641	14629	12283	984.98	797.93	727.78	orf19.2022	1142	1192	1148	103.91	87.94	92.00	8.82

- Indicates that expression levels for one allele were below the detectable limit and therefore the fold difference could not be calculated.

Key		Allele 1 Higher		Allele 1 Monoallelic		Allele 2 Higher		Allele 2 Monoallelic

Despite the analysis in chapter three identifying 233 genes with AEI, and this pipeline identifying only 21 genes, the high stringency of statistical tests used could explain this difference. Genes need to be identified as significant in all three statistical tests and in all three replicates in this pipeline to be considered as having a significant level of AEI. In the analysis in chapter three, only one statistical test, the Fisher Exact test, was applied to the data. When looking at the genes identified just by Fisher Exact test in the new pipeline, a much higher number is identified (69 in replicate one, 55 in replicate two and 47 in replicate three), accounting for some of the differences between the two analyses. The analysis here discounted any genes with alleles that only differed by INDELs. If this criteria was applied to the data from chapter three, the total number of significant genes is reduced to 193. Although this doesn't solely account for the differences in significant genes between the two pipelines, 99 genes identified in chapter three contain both SNPs and INDELs, and therefore measurements of allelic expression may be significantly altered using the new methodology. Although the high stringency of statistical tests here may be losing some genes as false negatives, it can increase the certainty of the genes that have been identified as having real disparities in allele expression levels.

5.3.2 Identification of Condition Specific AEI

Condition specific allelic expression imbalance was investigated by applying the new computational pipeline to RNA sequencing data from *C. albicans* grown under different conditions including 11 conditions from Bruno *et al.* (2010) (see section 5.2.2.1) and from co-culture with the bacterial species *Streptococcus gordonii* (Jenkinson, personal communication, see section 5.2.2.2). Genes identified as having significant levels of AEI from analysis of both data-sets are listed in Table 5.4 and Table 5.5 respectively. In each case, a comparison to the analysis of the wild-type strain SC5314, as detailed in section 5.3.1, is made. Genes are defined as monoallelic if the read count for all three replicates of one allele is less than five, and the other allele is more than five. In the re-analysis of the data from chapter three, four genes were defined as monoallelic, and a further two genes had very low expression (<10 counts) of one allele.

Table 5.4 Genes identified with AEI from RNA sequencing data obtained from Bruno *et al.* (2010). Numbers indicate fold difference in RPKM values (to 2 d.p.), with dashes present where one replicate has an RPKM value equal to zero.

Allele 1	Allele 2	Genes from 5.3.1	YPD	Serum	Congo Red	No Congo Red	High Oxidative	Low Oxidative	No Oxidative	M199 pH 8	M199 pH 4	Nitrosative	No Nitrosative
orf19.1048	orf19.8650						20.83			6.77			
orf19.10681	orf19.3171	6.24											
orf19.10952	orf19.3448			-			-	295.00	-	-	-		-
orf19.11687	orf19.4212	170.44								91.18			64.23
orf19.11957	orf19.4476	10.53					23.80						11.82
orf19.11980	orf19.4504									37.44	40.24		
orf19.12237	orf19.4773											51.09	
orf19.12579	orf19.5113	-	-		45.56		220.75		59.25	6.10		87.58	48.83
orf19.12758	orf19.5299												3.23
orf19.13093	orf19.5648	61.17		30.89	21.13		-	-	268.24	59.86	29.82	34.35	52.91
orf19.13163	orf19.5741	12.09											
orf19.13213	orf19.5791	875.44											
orf19.1357	orf19.8937	15.48											
orf19.13840	orf19.6486						14.29						
orf19.13891	orf19.6538	204.00								-			
orf19.14144	orf19.6854	938.50					-	-	-	-			-
orf19.1763	orf19.9332						44.97						
orf19.1915	orf19.9471	-							-				-
orf19.2023	orf19.9571												6.38
orf19.2023	orf19.9572												6.27
orf19.2268	orf19.9808	15.91					11.18			39.25			

orf19.2787	orf19.10303						187.91						
orf19.3365	orf19.10873						12.22						
orf19.4212	orf19.11689	5.65								77.65			7.67
orf19.4213	orf19.11689	74.23		46.60	36.30		35.32		33.45	42.04	25.36	144.10	47.02
orf19.4505	orf19.11981									49.18			
orf19.465	orf19.8096	41.26											
orf19.4959	orf19.12424	105.23											
orf19.5145	orf19.12611						17.55						
orf19.5602	orf19.13045	108.47											
orf19.6202	orf19.13583	7.56											
orf19.8644	orf19.1042	3.15											
orf19.8714	orf19.1117										47.58		
orf19.917	orf19.8532												3.23
orf19.9267	orf19.1700	205.00	503.50	523.58	320.50	183.50	-	-	-	-		-	-
orf19.9571	orf19.2022	8.82											4.08

- Indicates that expression levels for one allele were below the detectable limit and therefore the fold difference could not be calculated.

Key		Allele 1 Higher		Allele 1 Monoallelic		Allele 2 Higher		Allele 2 Monoallelic		No AEI

Table 5.5 Genes identified with AEI from RNA sequencing data obtained from co-culture with *S. gordonii.* Numbers indicate fold difference in RPKM values (to 2 d.p.), with dashes present where one replicate has an RPKM value equal to zero.

Allele 1	Allele 2	Genes from 5.3.1	*C. albicans* alone	Co-Culture	*C. albicans* in *S. gordonii* media
orf19.11075	orf19.3593		27.19	19.41	25.48
orf19.11087	orf19.3604		-	-	-
orf19.11553	orf19.4072		175.63	-	194.72
orf19.11972	orf19.4496		72.03	-	29.85
orf19.11990	orf19.4515		-	-	-
orf19.1219	orf19.8806		57.35	-	54.86
orf19.13069	orf19.5624		-	-	-
orf19.13093	orf19.5648	61.17	-	-	1330.02
orf19.13211	orf19.5789		13.36	12.26	11.66
orf19.13213	orf19.5791	875.44	-	-	435.75
orf19.13289	orf19.5867		95.98	144.88	93.80
orf19.13290	orf19.5869		24.44	17.50	14.98
orf19.1351	orf19.8931		-	-	-
orf19.1357	orf19.8937	15.48	-	-	-
orf19.14178	orf19.6889		144.00	-	-
orf19.1434	orf19.9008		60.84	128.46	75.46
orf19.1516	orf19.9091		24.55	20.94	13.38
orf19.1556	orf19.9129		100.48	96.78	92.47
orf19.1744	orf19.9311		20.86	38.25	19.45
orf19.1864	orf19.9420		105.52	236.75	65.19
orf19.2268	orf19.9808	15.91	-	796.50	851.83
orf19.2841	orf19.10359		-	-	-
orf19.3561	orf19.11045		279.30	166.00	315.31
orf19.3776	orf19.11257		-	-	-
orf19.4068	orf19.11551		40.79	45.35	56.49
orf19.4118	orf19.11600		26.22	26.06	21.02
orf19.4213	orf19.11689	74.23	38.54	138.29	29.12
orf19.4488	orf19.11964		20.41	28.56	18.28
orf19.5095	orf19.12561		-	-	257.87
orf19.6080	orf19.13499		-	-	-
orf19.6346	orf19.13702		-	-	-
orf19.797	orf19.8417		-	-	-
orf19.807	orf19.8426		40.45	37.98	48.43
orf19.8214	orf19.581		-	-	-
orf19.841	orf19.8461		105.22	237.19	82.59
orf19.8420	orf19.801		40.97	29.08	37.68
orf19.8421	orf19.802		46.78	53.89	27.97
orf19.8421	orf19.803		-	71.95	-
orf19.8428	orf19.808		46.07	37.89	30.62
orf19.8875	orf19.1295		36.09	22.31	38.06
orf19.8930	orf19.1350		268.92	283.00	160.61
orf19.8963	orf19.1383		55.84	6.32	17.95
orf19.8963	orf19.1384		99.15	-	-
orf19.8971	orf19.1393		16.82	22.40	-
orf19.9071	orf19.1494		-	-	54.13
orf19.9267	orf19.1700	205.00	41.23	52.75	37.93
orf19.9315	orf19.1747		39.87	50.33	53.74
orf19.9345	orf19.1779		-	-	-
orf19.9428	orf19.1872		12.31	16.83	10.75
orf19.9469	orf19.1913		31.13	-	38.81
orf19.9470	orf19.1914		62.83	-	39.51
orf19.9565	orf19.2015		24.71	36.61	26.05
orf19.9569	orf19.2019		42.94	66.00	202.17

orf19.9783	orf19.2242		-	-	418.00
orf19.9806	orf19.2266		207.70	39.30	58.67
orf19.11979	orf19.4503		13.72	13.44	
orf19.11980	orf19.4504		56.51	61.41	
orf19.11980	orf19.4505		-	-	
orf19.13747	orf19.6389		-	-	
orf19.7788	orf19.148		13.73	-	
orf19.10860	orf19.3352		10.02		12.82
orf19.11037	orf19.3554		15.25		10.35
orf19.11560	orf19.4079		62.86		101.58
orf19.1356	orf19.8936		28.91		31.07
orf19.1372	orf19.8952		38.81		22.41
orf19.14144	orf19.6854	938.50	-		-
orf19.1759	orf19.9328		19.89		16.38
orf19.182	orf19.7812		9.60		14.93
orf19.185	orf19.7816		44.50		39.72
orf19.1911	orf19.9467		34.24		18.81
orf19.1995	orf19.9547		61.06		73.21
orf19.3189	orf19.10699		-		-
orf19.3605	orf19.11088		-		57.50
orf19.3733	orf19.11218		6.03		4.88
orf19.4689	orf19.12158		-		-
orf19.4901	orf19.12367		-		-
orf19.5863	orf19.13285		-		-
orf19.7836	orf19.206		21.60		18.51
orf19.8753	orf19.1161		-		-
orf19.11687	orf19.4212	170.44	43.18		
orf19.13909	orf19.6556		44.21		
orf19.1495	orf19.9072		-		
orf19.1782	orf19.9348		-		
orf19.2018	orf19.9568		-		
orf19.3550	orf19.11034		-		
orf19.3934	orf19.11416		9.94		
orf19.6202	orf19.13583	7.56	13.05		
orf19.9578	orf19.2030		13.61		
orf19.11233	orf19.3746			-	75.20
orf19.11988	orf19.4513			11.78	15.82
orf19.1915	orf19.9471	-		44.51	31.47
orf19.2005	orf19.9556			34.90	36.17
orf19.216	orf19.7848			40.39	33.69
orf19.7828	orf19.198			15.80	19.56
orf19.8414	orf19.795			50.63	30.79
orf19.9563	orf19.2012			203.50	83.75
orf19.11110	orf19.3627			-	
orf19.1386	orf19.8964			-	
orf19.1479	orf19.9054			56.55	
orf19.1765	orf19.9334			20.12	
orf19.220	orf19.7851			20.44	
orf19.465	orf19.8096	41.26		-	
orf19.5225	orf19.12690			11.59	
orf19.8485	orf19.866			14.10	
orf19.9825	orf19.2285			-	
orf19.1048	orf19.8650				6.67
orf19.11254	orf19.3770				10.17
orf19.1390	orf19.8968				8.90
orf19.1531	orf19.9106				17.34
orf19.1873	orf19.9429				72.50
orf19.2116	orf19.9664				8.21
orf19.3555	orf19.11038				6.39
orf19.3590	orf19.11072				5.41

orf19.3771	orf19.11254				26.37
orf19.4280	orf19.11756				18.28
orf19.4506	orf19.11982				31.98
orf19.7708	orf19.35				7.56
orf19.8776	orf19.1185				20.22
orf19.8951	orf19.1371				8.86
orf19.9090	orf19.1515				20.16
orf19.9323	orf19.1754				44.75
orf19.9663	orf19.2115				7.03
orf19.10681	orf19.3171	6.24			
orf19.11957	orf19.4476	10.53			
orf19.12579	orf19.5113	-			
orf19.13163	orf19.5741	12.09			
orf19.13891	orf19.6538	204.00			
orf19.4212	orf19.11689	5.65			
orf19.4959	orf19.12424	105.23			
orf19.5602	orf19.13045	108.47			
orf19.8644	orf19.1042	3.15			
orf19.9571	orf19.2022	8.82			

- Indicates that expression levels for one allele were below the detectable limit and therefore the fold difference could not be calculated.

Key		Allele 1 Higher		Allele 1 Monoallelic		Allele 2 Higher		Allele 2 Monoallelic		No AEI

Analysis of the data from Bruno *et al.* (2010) identified a small number of genes with AEI (27 across all 11 conditions), similar to the analysis carried out in section 5.3.1. Of these genes, five were monoallelic in all conditions identified and a further seven were identified as monoallelic in some condition but not all conditions. Low expression (<10 counts) of one allele was observed in three genes out of the seven with inconsistent monoallelism and in one gene with AEI. Again, this smaller than expected number may be due to the high stringency of statistical tests used as discussed above. Additionally, each condition had only two replicates, reducing the amount of data available, possibly impacting upon the amount of genes with AEI identified.

On the other hand, a larger number of genes (123) were identified from the three separate conditions of the co-culture with *S. gordonii* experiment. 31 of these genes were monoallelic in all conditions and a further 12 genes were monoallelic in at least one but not all conditions. Of these 12, ten genes had low expression (<10 counts) of one allele in the other conditions and a further 20 genes had low expression levels of one allele. The larger number of genes identified could be due to several reasons surrounding the sequencing process itself. As opposed to the analysis carried out in section 5.3.1 and carried out by

Bruno *et al.* (2010), where an Illumina GAII sequencer was used, the RNA sequencing for the co-culture experiment used an Illumina HiSeq2500 platform. This platform has been shown to produce a much higher number of reads than the Genome Analyser (Minoche *et al.*, 2011). Additionally, longer 100 bp paired end reads were obtained for the co-culture experiment, compared to shorter 76 bp and 30 bp reads for the in-house and Bruno *et al.* (2010) analyses. In all, this increase in the volume and length of reads gives a larger amount of data in which AEI may be detected, possibly explaining the larger number of genes identified. This has been reported to be the case for general studies investigating differential gene expression, where an increase in replicate number and sequencing depth increases statistical power (Zhang *et al.*, 2014).

Where genes have been identified in more than one condition, the allele with the higher level of expression remains the same and does not differ from condition to condition. However for some genes, such as orf19.5113, expression from the allele with lower levels is detected in some conditions and the gene is identified as monoallelic in other conditions. In other cases, the fold change in RPKM values is also variable between conditions, as an example orf19.5648 in the Bruno *et al.* (2010) data shows a much higher fold change under growth in no oxidative stress. These results suggest that in cases of genes with AEI, one allele has a greater functional contribution than the other, but this remains constant across all conditions, as opposed to the alleles having distinct condition-specific functions.

Interestingly, few similarities in the genes identified with AEI were seen across similar growth conditions, such as low and high oxidative stress, or the four control conditions, YPD, no Congo red, no oxidative stress and no nitrosative stress. It is difficult to determine why this is the case, but it is possible to infer that although the genes with AEI do differ between growth conditions, the conditions themselves are unlikely to trigger this difference.

5.3.3 Gene Ontology (GO) Analysis

Despite the lack of evidence for a condition specific response in AEI, the functions of genes with AEI were determined using Gene Ontology analysis on the separate gene lists, using the GO Term Finder tool available from the

Candida Genome Database (www.candidagenome.org) (Inglis *et al.*, 2012). This would indicate whether all genes with AEI have similar functions, or functions related to the growth condition.

An over representation of genes involved in oxidation-reduction processes and activities, iron transport, and locations to the cell surface and plasma membrane was observed for the analysis of the genes identified in section 5.3.1 where the wild-type strain was grown in YPD (Table 5.6). Although, in the analysis of the Bruno *et al.* (2010) data-set, just two genes (orf19.1700 and orf19.5113) were identified with AEI from growth in YPD, when combining the results from all of the control conditions used by Bruno *et al.* (2010) (YPD, no Congo red, no oxidative stress and no nitrosative stress) similar Gene Ontology terms were identified as over represented (Table 5.7). Additionally, an over representation of genes involved in carbohydrate transport was observed.

GO analysis of the genes identified in other conditions from the Bruno *et al.* (2010) experiment showed that there appears to be no condition specific roles of AEI. Similar processes and functions to those found in the control conditions above were found; metal ion binding and transport processes were identified in growth in serum and Congo red, oxidation and reduction processes were identified in growth in Congo red, high oxidative stress, nitrosative stress and tissue culture media (M199) at both pH 4 and pH 8, and plasma membrane localisation was identified in growth in nitrosative stress (Table 5.8).

Conversely, GO analysis of the 123 genes identified with AEI from the three conditions of the co-culture experiment show over representation of different processes to the genes identified from Bruno *et al.* (2010), with functions involved in protein kinase and transferase activity being over represented (Table 5.9). Again, condition specific analysis showed no over representation of functions when *C. albicans* is cultured with *S. gordonii* nor when it is grown in *S. gordonii* media alone. However growth in hypha inducing conditions *per se* did identify similar functions to those identified from the Bruno *et al.* (2010) data-set with enrichment of metal ion binding (Table 5.10).

Taken together these results suggest that AEI does not differ in a condition specific manner. However, contrary to the findings of chapter three, it does suggest that genes with AEI may be enriched for functional roles in metal binding and transport, and oxidation and reduction processes.

Table 5.6 GO Terms enriched for genes exhibiting AEI when SC5314 is grown in YPD. GO terms identified using "CGD Gene Ontology Term Mapper" (www.candidagenome.org).

Ontology	GOID	GO term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Gene(s) annotated to the term
Process	70627	"ferrous iron import"	2 out of 22 genes, 9.1%	2 out of 6517 background genes, 0.0%	0.0011	0.00%	*FET3 FET31*
	97286	"iron ion import"	2 out of 22 genes, 9.1%	2 out of 6517 background genes, 0.0%	0.0011	0.00%	*FET3 FET31*
	15684	"ferrous iron transport"	2 out of 22 genes, 9.1%	3 out of 6517 background genes, 0.0%	0.00331	0.00%	*FET3 FET31*
	55085	"transmembrane transport"	7 out of 22 genes, 31.8%	363 out of 6517 background genes, 5.6%	0.01035	1.00%	*POR1 FCY21* orf19.2022 *FET3 FET31 VMA11 ATP1*
	34755	"iron ion transmembrane transport"	2 out of 22 genes, 9.1%	10 out of 6517 background genes, 0.2%	0.04907	8.80%	*FET3 FET31*
Function	5507	"copper ion binding"	3 out of 22 genes, 13.6%	22 out of 6517 background genes, 0.3%	0.00175	4.00%	*FET3 FET99 FET31*
	16724	"oxidoreductase activity, oxidizing metal ions, oxygen as acceptor"	2 out of 22 genes, 9.1%	4 out of 6517 background genes, 0.1%	0.00242	4.00%	*FET3 FET31*
	4322	"ferroxidase activity"	2 out of 22 genes, 9.1%	4 out of 6517 background genes, 0.1%	0.00242	2.67%	*FET3 FET31*

	16722	“oxidoreductase activity, oxidizing metal ions”	2 out of 22 genes, 9.1%	12 out of 6517 background genes, 0.2%	0.02624	7.50%	*FET3 FET31*
Component	71944	“cell periphery”	10 out of 22 genes, 45.5%	685 out of 6517 background genes, 10.5%	0.0011	0.00%	*POR1 FCY21* orf19.2022 *FET3 FET99 FET31 IFF9 ADH2 ALS1 ATP1*
	9986	“cell surface”	6 out of 22 genes, 27.3%	204 out of 6517 background genes, 3.1%	0.00192	0.00%	*RPS7A FET99 IFF9 ALS1 RBT4 ATP1*
	5886	“plasma membrane”	8 out of 22 genes, 36.4%	475 out of 6517 background genes, 7.3%	0.00392	0.67%	*POR1 FCY21* orf19.2022 *FET3 FET99 FET31 ADH2 ATP1*

Table 5.7 GO terms enriched for genes exhibiting AEI when SC5314 is grown in control conditions (YPD, No Congo Red, No Oxidative Stress and No Nitrosative Stress) from Bruno *et al* (2010). GO terms identified using "CGD Gene Ontology Term Mapper" (www.candidagenome.org).

Ontology	GOID	GO term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Gene(s) annotated to the term
Process	70627	"ferrous iron import"	2 out of 14 genes, 14.3%	2 out of 6517 background genes, 0.0%	0.00024	0.00%	*FET3 FET31*
	97286	"iron ion import"	2 out of 14 genes, 14.3%	2 out of 6517 background genes, 0.0%	0.00024	0.00%	*FET3 FET31*
	15684	"ferrous iron transport"	2 out of 14 genes, 14.3%	3 out of 6517 background genes, 0.0%	0.00072	0.00%	*FET3 FET31*
	34755	"iron ion transmembrane transport"	2 out of 14 genes, 14.3%	10 out of 6517 background genes, 0.2%	0.01081	7.00%	*FET3 FET31*
	55085	"transmembrane transport"	5 out of 14 genes, 35.7%	363 out of 6517 background genes, 5.6%	0.03054	11.20%	orf19.2022 *HGT7 FET3 FET31 ATP1*
	6826	"iron ion transport"	2 out of 14 genes, 14.3%	19 out of 6517 background genes, 0.3%	0.04067	11.33%	*FET3 FET31*
Function	5507	"copper ion binding"	3 out of 14 genes, 21.4%	22 out of 6517 background genes, 0.3%	0.00018	0.00%	*FET3 FET99 FET31*

	16724	"oxidoreductase activity, oxidizing metal ions, oxygen as acceptor"	2 out of 14 genes, 14.3%	4 out of 6517 background genes, 0.1%	0.00043	1.00%	*FET3 FET31*
	4322	"ferroxidase activity"	2 out of 14 genes, 14.3%	4 out of 6517 background genes, 0.1%	0.00043	0.67%	*FET3 FET31*
	16722	"oxidoreductase activity, oxidizing metal ions"	2 out of 14 genes, 14.3%	12 out of 6517 background genes, 0.2%	0.00479	2.00%	*FET3 FET31*
	5355	"glucose transmembrane transporter activity"	2 out of 14 genes, 14.3%	19 out of 6517 background genes, 0.3%	0.01232	1.60%	orf19.2022 *HGT7*
	15149	"hexose transmembrane transporter activity"	2 out of 14 genes, 14.3%	20 out of 6517 background genes, 0.3%	0.01367	1.33%	orf19.2022 *HGT7*
	15145	"monosaccharide transmembrane transporter activity"	2 out of 14 genes, 14.3%	22 out of 6517 background genes, 0.3%	0.01659	1.14%	orf19.2022 *HGT7*
	15144	"carbohydrate transmembrane transporter activity"	2 out of 14 genes, 14.3%	28 out of 6517 background genes, 0.4%	0.02697	1.00%	orf19.2022 *HGT7*
	1901476	"carbohydrate transporter activity"	2 out of 14 genes, 14.3%	28 out of 6517 background genes, 0.4%	0.02697	0.89%	orf19.2022 *HGT7*
	51119	"sugar transmembrane transporter activity"	2 out of 14 genes, 14.3%	28 out of 6517 background genes, 0.4%	0.02697	0.80%	orf19.2022 *HGT7*
Component	5886	"plasma membrane"	7 out of 14 genes, 50.0%	475 out of 6517 background genes, 7.3%	0.00051	0.00%	orf19.2022 *HGT7 FET3 FET99 FET31 ADH2 ATP1*

	71944	“cell periphery”	7 out of 14 genes, 50.0%	685 out of 6517 background genes, 10.5%	0.00561	1.00%	orf19.2022 *HGT7 FET3 FET99 FET31 ADH2 ATP1*

Table 5.8 GO terms enriched for genes exhibiting AEI when SC5314 is grown in different conditions from Bruno *et al* (2010). GO terms identified using "CGD Gene Ontology Term Mapper" (www.candidagenome.org).

Condition/ Ontology	GOID	GO term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Gene(s) annotated to the term
Serum							
Function	5507	"copper ion binding"	2 out of 5 genes, 40.0%	22 out of 6517 background genes, 0.3%	0.00075	2.00%	*FET99 FET31*
Congo Red							
Process	55114	"oxidation-reduction process"	3 out of 5 genes, 60.0%	418 out of 6517 background genes, 6.4%	0.02614	88.00%	*FET99 FET31 ADH2*
Function	5507	"copper ion binding"	2 out of 5 genes, 40.0%	22 out of 6517 background genes, 0.3%	0.00075	0.00%	*FET99 FET31*
	46914	"transition metal ion binding"	3 out of 5 genes, 60.0%	401 out of 6517 background genes, 6.2%	0.01474	8.00%	*FET99 FET31 ADH2*
	16491	"oxidoreductase activity"	3 out of 5 genes, 60.0%	421 out of 6517 background genes, 6.5%	0.01698	5.33%	*FET99 FET31 ADH2*
	46872	"metal ion binding"	3 out of 5 genes, 60.0%	502 out of 6517 background genes, 7.7%	0.02827	6.00%	*FET99 FET31 ADH2*
	43169	"cation binding"	3 out of 5 genes, 60.0%	510 out of 6517 background genes, 7.8%	0.02959	5.20%	*FET99 FET31 ADH2*

M199 pH 4							
Process	55114	"oxidation-reduction process"	4 out of 6 genes, 66.7%	418 out of 6517 background genes, 6.4%	0.0009	14.00%	orf19.1117 *FET99 FET31* orf19.4504
Function	5507	"copper ion binding"	2 out of 6 genes, 33.3%	22 out of 6517 background genes, 0.3%	0.00113	0.00%	*FET99 FET31*
	16491	"oxidoreductase activity"	4 out of 6 genes, 66.7%	421 out of 6517 background genes, 6.5%	0.00162	0.00%	orf19.1117 *FET99 FET31* orf19.4504
	46914	"transition metal ion binding"	3 out of 6 genes, 50.0%	401 out of 6517 background genes, 6.2%	0.02814	4.67%	*FET99 FET31* orf19.4504
M199 pH 8							
Process	70627	"ferrous iron import"	2 out of 13 genes, 15.4%	2 out of 6517 background genes, 0.0%	0.00023	4.00%	*FET3 FET31*
	97286	"iron ion import"	2 out of 13 genes, 15.4%	2 out of 6517 background genes, 0.0%	0.00023	2.00%	*FET3 FET31*
	15684	"ferrous iron transport"	2 out of 13 genes, 15.4%	3 out of 6517 background genes, 0.0%	0.0007	1.33%	*FET3 FET31*
	34220	"ion transmembrane transport"	4 out of 13 genes, 30.8%	116 out of 6517 background genes, 1.8%	0.00385	2.00%	*FET3 FET31 VMA11 ATP1*
	55114	"oxidation-reduction process"	6 out of 13 genes, 46.2%	418 out of 6517 background genes, 6.4%	0.00501	1.60%	*FET3 FET99 FET31* orf19.4504 *ADH3 ADH2*

	6812	"cation transport"	4 out of 13 genes, 30.8%	145 out of 6517 background genes, 2.2%	0.0092	2.33%	*FET3 FET31 VMA11 ATP1*
	34755	"iron ion transmembrane transport"	2 out of 13 genes, 15.4%	10 out of 6517 background genes, 0.2%	0.01048	2.00%	*FET3 FET31*
	15988	"energy coupled proton transmembrane transport, against electrochemical gradient"	2 out of 13 genes, 15.4%	16 out of 6517 background genes, 0.2%	0.02777	5.25%	*VMA11 ATP1*
	15991	"ATP hydrolysis coupled proton transport"	2 out of 13 genes, 15.4%	16 out of 6517 background genes, 0.2%	0.02777	4.67%	*VMA11 ATP1*
	6826	"iron ion transport"	2 out of 13 genes, 15.4%	19 out of 6517 background genes, 0.3%	0.03944	6.20%	*FET3 FET31*
Function	16491	"oxidoreductase activity"	7 out of 13 genes, 53.8%	421 out of 6517 background genes, 6.5%	0.00023	0.00%	*IFD6 FET3 FET99 FET31* orf19.4504 *ADH3 ADH2*
	5507	"copper ion binding"	3 out of 13 genes, 23.1%	22 out of 6517 background genes, 0.3%	0.0004	0.00%	*FET3 FET99 FET31*
	16724	"oxidoreductase activity, oxidizing metal ions, oxygen as acceptor"	2 out of 13 genes, 15.4%	4 out of 6517 background genes, 0.1%	0.00094	0.00%	*FET3 FET31*
	4322	"ferroxidase activity"	2 out of 13 genes, 15.4%	4 out of 6517 background genes, 0.1%	0.00094	0.00%	*FET3 FET31*
	46914	"transition metal ion binding"	6 out of 13 genes, 46.2%	401 out of 6517 background genes, 6.2%	0.00266	0.80%	*FET3 FET99 FET31* orf19.4504 *ADH3 ADH2*

	46872	"metal ion binding"	6 out of 13 genes, 46.2%	502 out of 6517 background genes, 7.7%	0.00936	0.67%	*FET3 FET99 FET31* orf19.4504 *ADH3 ADH2*
	43169	"cation binding"	6 out of 13 genes, 46.2%	510 out of 6517 background genes, 7.8%	0.01022	0.57%	*FET3 FET99 FET31* orf19.4504 *ADH3 ADH2*
	16722	"oxidoreductase activity, oxidizing metal ions"	2 out of 13 genes, 15.4%	12 out of 6517 background genes, 0.2%	0.0103	0.50%	*FET3 FET31*
	43167	"ion binding"	8 out of 13 genes, 61.5%	1140 out of 6517 background genes, 17.5%	0.02061	0.44%	*RCK2 FET3 FET99 FET31* orf19.4504 *ADH3 ADH2 ATP1*
High Oxidative							
Process	55114	"oxidation-reduction process"	5 out of 15 genes, 33.3%	418 out of 6517 background genes, 6.4%	0.07621	100.00%	*IFR1 DAO2 FET99 FET31 ADH2*
Function	16491	"oxidoreductase activity"	6 out of 15 genes, 40.0%	421 out of 6517 background genes, 6.5%	0.00659	6.00%	*IFD6 IFR1 DAO2 FET99 FET31 ADH2*
	5507	"copper ion binding"	2 out of 15 genes, 13.3%	22 out of 6517 background genes, 0.3%	0.03448	8.00%	*FET99 FET31*
Nitrosative							
Process	55114	"oxidation-reduction process"	4 out of 6 genes, 66.7%	418 out of 6517 background genes, 6.4%	0.00293	4.00%	*FET99 FET31 AOX2 ADH2*
Function	5507	"copper ion binding"	2 out of 6 genes, 33.3%	22 out of 6517 background genes, 0.3%	0.00113	4.00%	*FET99 FET31*

	16491	“oxidoreductase activity”	4 out of 6 genes, 66.7%	421 out of 6517 background genes, 6.5%	0.00162	3.00%	*FET99 FET31 AOX2 ADH2*
	46914	“transition metal ion binding”	3 out of 6 genes, 50.0%	401 out of 6517 background genes, 6.2%	0.02814	9.33%	*FET99 FET31 ADH2*
Component	5886	“plasma membrane”	4 out of 6 genes, 66.7%	475 out of 6517 background genes, 7.3%	0.00668	0.00%	*FET99 FET31 AOX2 ADH2*
	71944	“cell periphery”	4 out of 6 genes, 66.7%	685 out of 6517 background genes, 10.5%	0.02746	1.00%	*FET99 FET31 AOX2 ADH2*

Table 5.9 GO terms enriched for genes exhibiting AEI when SC5314 is grown in hypha inducing conditions, co-cultured with *S. gordonii* and grown in *S. gordonii* media. GO terms identified using "CGD Gene Ontology Term Mapper" (www.candidagenome.org).

Ontology	GOID	GO term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Gene(s) annotated to the term
Function	4672	"protein kinase activity"	9 out of 130 genes, 6.9%	115 out of 6525 background genes, 1.8%	0.03399	6.00%	orf19.148/orf19.7788 orf19.1754 orf19.2015 *RCK2* orf19.35 *CDC7* orf19.3776 *PKH2 MKK2*
	16769	"transferase activity, transferring nitrogenous groups"	4 out of 130 genes, 3.1%	21 out of 6525 background genes, 0.3%	0.06694	22.00%	*AAT1* orf19.3771 *BAT21* orf19.803

Table 5.10 GO terms enriched for genes exhibiting AEI when SC5314 is grown in hypha inducing conditions. GO terms identified using "CGD Gene Ontology Term Mapper" (www.candidagenome.org).

Ontology	GOID	GO term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Gene(s) annotated to the term
Function	5507	"copper ion binding"	2 out of 10 genes, 20.0%	22 out of 6525 background genes, 0.3%	0.00384	6.00%	*FET3 FET99*
	46872	"metal ion binding"	5 out of 10 genes, 50.0%	502 out of 6525 background genes, 7.7%	0.00384	3.00%	orf19.1495 orf19.2018 *CAR1 FET3 FET99*
	43169	"cation binding"	5 out of 10 genes, 50.0%	510 out of 6525 background genes, 7.8%	0.00414	2.00%	orf19.1495 orf19.2018 *CAR1 FET3 FET99*
	46914	"transition metal ion binding"	4 out of 10 genes, 40.0%	401 out of 6525 background genes, 6.1%	0.01753	8.00%	orf19.2018 *CAR1 FET3 FET99*

5.3.4 Differences in Promoter Sequences

As in chapter three, the promoter regions of the genes identified with AEI from the three data-sets were analysed for sequence differences to assess if polymorphisms in these regions are linked to differential allele expression. As promoter regions are currently undefined in *C. albicans,* the 1000 bp upstream of each allele of a gene (or the distance to the adjacent ORF) were compared. Of the 21 genes identified with AEI from repeating the analysis of chapter three, 14 genes (66 %) differed in their upstream region by at least a single SNP. 23 of the 27 (85 %) genes identified from the Bruno *et al.* (2010) study had at least one polymorphism in the upstream region of the alleles. Polymorphisms were found in the upstream region of 100 of the 123 (82 %) genes identified with AEI from the co-culture experiment. Although differences in these regions were identified in a high number of genes, approximately a quarter of promoter regions were homozygous. Based upon the observed overall level of heterozygosity - one SNP in every 237 bp (Jones *et al.*, 2004), the probability that a 1000 bp region is homozygous can be calculated as follows:

$$\frac{236}{237}^{1000} = 0.015$$

Therefore, based on this probability, it would be expected that less than 1 promoter region would be homozygous out of the 21 and 27 genes identified from the re-analysis of the data from chapter three and from the data from Bruno *et al.* (2010) respectively. Just 2 promoter regions would be expected to be homozygous from the 123 genes identified from the co-culture experiment. However, a significantly larger proportion of promoter sequences were found to be homozygous in all three data-sets (chi-square test, d.f. = 1, $p = 3.61 \times 10^{-34}$, $p = 7 \times 10^{-9}$, and $p = 2.64 \times 10^{-57}$ respectively). Therefore, as opposed to differences in promoter regions driving differences in allele expression levels, it can be concluded that polymorphisms in the promoter regions of these genes are actually selected against. Based upon the observation in chapter three that genes with equal allele expression also have a similar rate of polymorphisms in promoter regions at 78% (section 3.3.1.3) it could be suggested that polymorphisms in upstream regions are selected against in all genes. The reasoning behind this is unclear, but may be due to the pressure to maintain gene expression levels. It should be noted that this probability has been

calculated for the simplest model where 1000 bp upstream were compared for each in gene. In reality, upstream regions varied in length dependent upon the distance to the adjacent open reading frame which may impact upon this probability estimate.

5.3.5 Identification of Genes with Differing Allele Expression Levels between Growth Conditions

Some genes which have not been identified as having significant levels of allelic expression imbalance by the computational pipeline will still have uneven changes in levels of allele expression when growth conditions change. To identify genes where the growth condition causes a change in expression of one allele greater than the change in the other allele, the fold difference in RPKM values of each allele between conditions was calculated for every gene for 16 different condition comparisons taken from the analysis of the Bruno *et al.* (2010) data (as described in section 5.2.6). The Log_{10} fold change in allele one expression level was then plotted against the Log_{10} fold change in the allele two expression level for each condition comparison and linear trend lines were fitted. Figure 5.3 demonstrates an example of the relationship observed for the change in expression of alleles between growth in YPD and growth in the cell wall damaging agent Congo Red.

The equation of the linear trend line was used to calculate the predicted y values. Any genes where the observed y value was greater than two standard deviations away from the predicted y value were identified as having a significantly larger change in one allele expression than the other allele between those growth conditions. Appendix I Table V lists all 256 genes identified as significant in comparisons of two conditions or more.

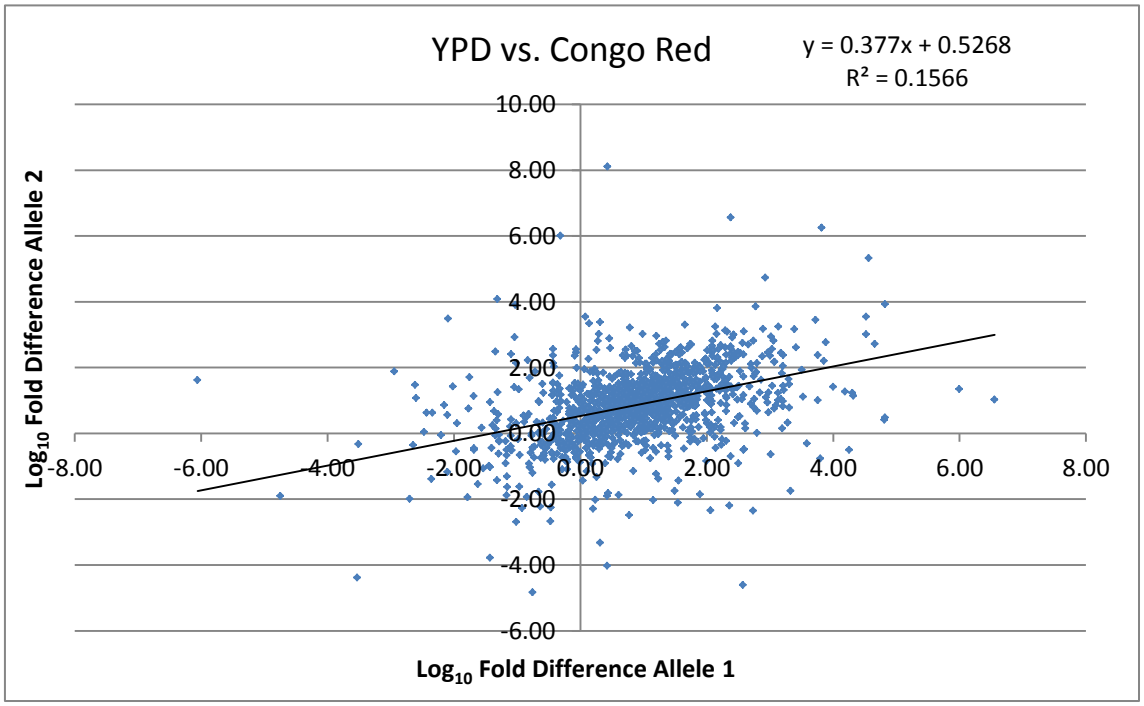


Figure 5.3 Log_{10} fold differences in allele expression levels between growth in YPD and growth in Congo Red. A linear trend line has been fitted.

To identify genes which frequently have a disparity in the fold difference of alleles between conditions, the frequency of genes identified as significant for each condition comparison was calculated and multiplied by each other for each gene to give a significance measure. To give an example, Table 5.11 gives an explanation of this calculation, looking at the gene *GPX1* that was identified as having significant disparity in the change in allele expression levels between 12 different conditions. The probability of this occurring by chance is calculated at $p = 3.758 \times 10^{-17}$ and is therefore significantly unlikely, suggesting that the two alleles of this gene have different responses to the change in growth conditions. All genes with $p < 1.95 \times 10^{-4}$, as determined by Bonferroni correction for multiple comparisons, were identified as significant by this analysis (Appendix I Table VI). This list includes the gene *ERB1* which was also identified as having significant levels of AEI in chapter three.

Table 5.11 Example of calculating the probability that a gene has significant differences in change of allele expression across multiple condition comparisons. The example here shows the results for the gene *GPX1.*

	YPD vs. Congo Red	YPD vs. No Congo Red	YPD vs. Low Oxidative	YPD vs. High Oxidative	YPD vs. No Oxidative	YPD vs Serum	YPD vs. No Nitrosative	YPD vs. M199 pH 8	YPD vs. M199 pH 4	High Oxidative vs. Low Oxidative	YPD vs. Nitrosative	Congo Red vs. No Congo Red	Nitrosative vs. No Nitrosative	Low Oxidative vs. No Oxidative	High Oxidative vs. No Oxidative	M199 pH 4 vs. M199 pH 8
No. of Genes Identified as Significant	56	55	55	67	60	47	49	51	53	46	55	43	47	62	57	43
Total Number of Genes Analysed	1265	1263	1171	1156	1269	1304	1149	1121	1276	1130	1233	1239	1112	1152	1137	1113
Frequency of Significant Genes	0.044	0.044	0.047	0.058	0.047	0.036	0.043	0.045	0.042	0.041	0.045	0.035	0.042	0.054	0.050	0.039
Is *GPX1* significant in this comparison?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N
Probability = 3.758×10^{-17}	0.044 x	0.044 x	0.047 x	0.058 x	0.047 x	0.036 x	0.043 x	0.045 x	0.042 x	0.041 x	0.045 x	0.035 x	0.958 x	0.946 x	0.950 x	0.961 x

5.3.6 Target Genes for Heterozygous Knockout Construction

From the condition-specific data compiled above, four genes were selected for heterozygous knockout construction with a hypothesis that these alleles are likely to differ in function. *ADH2* was identified as having significantly higher expression of allele one or monoallelic expression of allele one from the re-analysis of the data from chapter three and from seven conditions of the Bruno *et al.* (2010) study (section 5.3.2). *ADH2* was also identified as having significant differences in the change of allele expression across nine different growth condition comparisons (section 5.3.5). *GPX1* was not identified as having significant AEI but was found to have significant differences in changes in allele expression across 12 different condition comparisons, making it the most significant gene in the analysis in section 5.3.5. *RPS7A* was identified as having significantly higher expression levels of allele one or monoallelic allele one expression in all data-sets analysed in section 5.3.2 except from growth in M199 pH 4. Finally, orf19.5648 was found to have higher expression levels of allele two or monoallelic allele two expression in all data-sets analysed in section 5.3.2 except for growth in YPD and the no Congo red control from Bruno *et al.* (2010).

Heterozygous knockout mutants were successfully constructed for just two of these four genes; *ADH2* and *GPX1.* Only constructs lacking *RPS7A* allele one and constructs lacking orf19.5648 allele two were obtained after numerous transformation attempts. Due to the inability to produce knockouts of both alleles, there are no phenotypic screens for these genes. Reasons explaining this are further discussed in section 5.3.9. Validation of insertion of knockout cassettes at the correct genomic location was carried out using colony PCR as described in sections 2.9 and 2.10 (Figure 5.4). Southern blotting was used to further validate the *ADH2* and *GPX1* mutants, showing insertion in the correct position and that only one copy of the nourseothricin cassette had been inserted, as described in section 2.15 (Figure 5.5).

To show that the nourseothricin cassette was not causing phenotypic differences, the control strain used in chapter three containing nourseothricin at the RPS1 locus (SC12) was used as described in section 2.14.1.

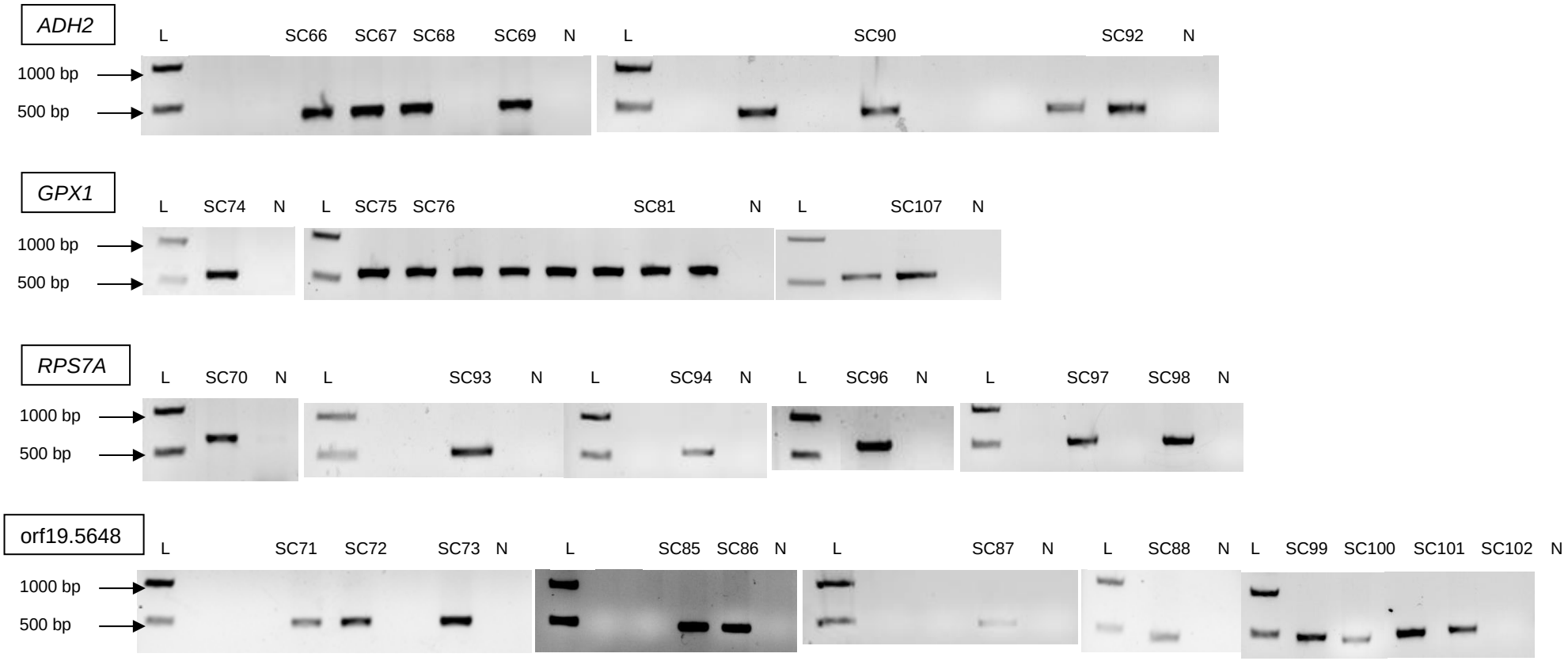


Figure 5.4 PCR validations of heterozygous knockout mutants. L = 1 kb DNA ladder (Fermentas, UK). N = negative control containing no DNA. SC66, SC67, SC68, SC69, SC90 and SC92 = *ADH2* knockouts with a band expected at 434 bp. SC74, SC75, SC76, SC81 and SC107 = *GPX1* knockouts with a band expected at 563 bp. SC70, SC93, SC94, SC96, SC97 and SC98 = *RPS7A* knockouts with a band expected at 550 bp. SC71, SC72, SC73, SC85, SC86, SC87, SC88, SC99, SC100, SC101 and SC102 = orf19.5648 knockouts with a band expected at 487 bp.

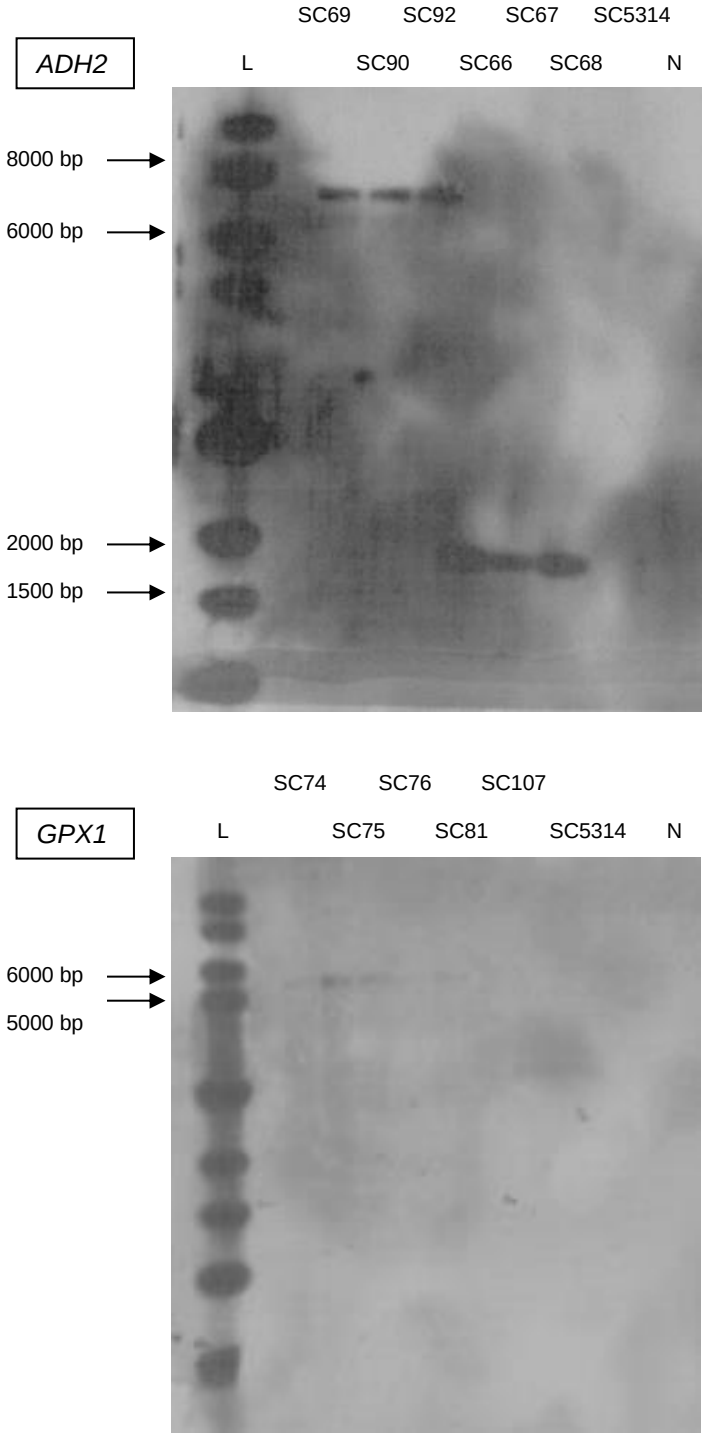


Figure 5.5 Southern blotting validations of heterozygous knockout mutants. L = 1 kb DNA ladder (Fermentas, UK). SC5314 = negative control containing untransformed wild-type DNA. N = negative control containing all reaction components except DNA. SC69, SC90 and SC92 = *ADH2* allele one heterozygous knockouts showing a single band at 7145 bp. SC66, SC67 and SC68 = *ADH2* allele two heterozygous knockouts showing a single band at 1910 bp. SC74, SC75, SC76, SC81 and SC107 = *GPX1* heterozygous knockouts showing a single very faint band at 5342 bp.

5.3.7 Phenotypic Screening

5.3.7.1 *ADH2* Phenotypic Screening

The *ADH2* gene encodes an alcohol dehydrogenase which works closely with the related gene product of *ADH1* to convert ethanol to acetaldehyde, allowing *Candida albicans* to utilise ethanol as a sole carbon source (Bertram *et al.*, 1996, Marttila *et al.*, 2013). Screening of heterozygous knockout mutants using generic assays showed little differences between the knockout strains of either allele and the wild-type strain SC5314. Growth at both 30 °C and 37 °C did not differ significantly from the wild-type strain SC5314 for any of the measures taken (ANOVA, $p > 0.05$; Figure 5.6a, Figure 5.6b, Table 5.12). Adhesion to buccal epithelial cells also did not differ significantly from the wild-type strain SC5314 for any of the three measures taken (two sample t-test, $p > 0.05$, Figures 5.6c-e). All strains were able to switch to the hyphal form at the same rate as the wild-type strain SC5314 (Figure 5.6f), showing that although over expression of *ADH2* protein has been reported in hyphae (Hernández *et al.*, 2004) the individual alleles are not functionally important for this morphological switch. Virulence of the heterozygous knockout mutants using a *Galleria mellonella* infection model was the same as the wild-type strain at all three cell concentrations tested (Kaplan-Meier, d.f. = 7, $p > 0.05$; Figure 5.6g). The response to growth in the antifungal compounds 5-flucytosine and amphotericin B does not differ from the wild-type strain SC5314 (Figure 5.6i and j). In response to fluconazole, SC66, SC67 and SC68, the allele two knockout strains, have a similar response to the wild-type strain SC5314. However, SC90 and SC92, the allele one knockout strains, appear to have reduced recovery, with lower growth rates at 48 hours than SC5314, although this trend is not observed in the third isolate, SC69, which is also an allele one knockout strain (Figure 5.6h). Conversely, growth on fluconazole agar plates at 48 hours does not show a difference in growth for any of the heterozygous knockout strains (Figure 5.6m). As *ADH2* has previously been observed to exhibit increased protein expression under exposure to the related azole ketoconazole (Hoehamer *et al.*, 2010), it would be sensible to infer that the alleles may have a role in the response to fluconazole, however from the results gathered here, that role is not apparent.

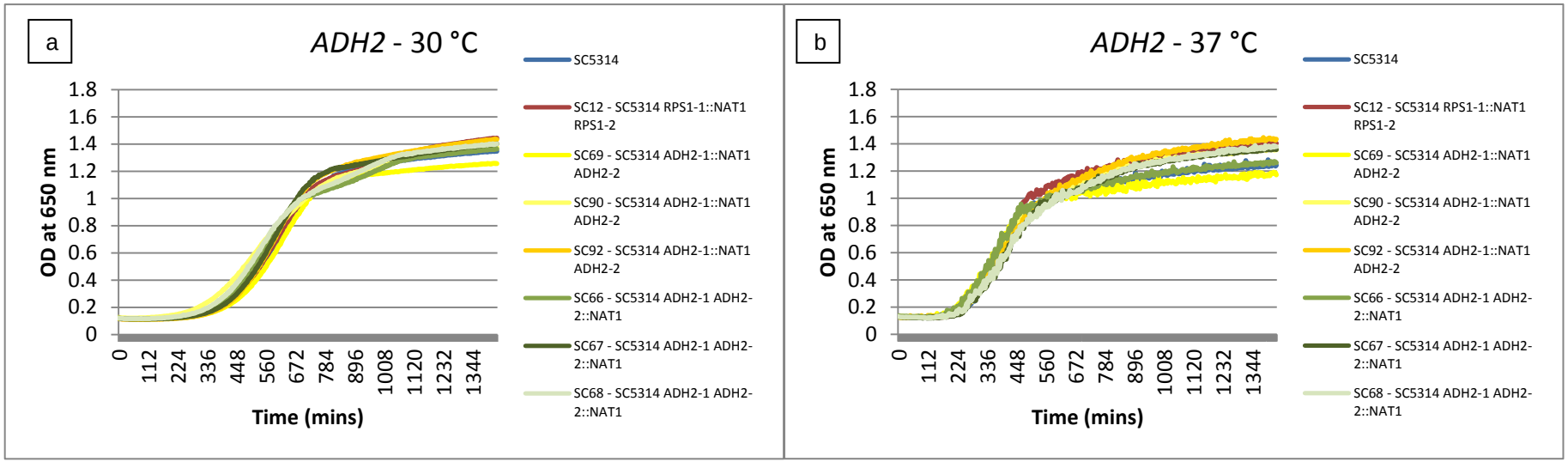
The *ADH2* gene has not been well characterised in *Candida* albicans, however the function of the closely related *ADH1* gene has been investigated in yeast species. *Saccharomyces cerevisiae* alcohol dehydrogenase knockout strains are unable to grow on ethanol as the sole carbon source, however the *C. albicans ADH1* gene can functionally compensate for this phenotype (Bertram *et al.*, 1996). In our investigation, all strains, including the wild-type strain SC5314, had reduced growth when ethanol was used as the sole carbon source as opposed to glucose, however comparison of the growth of the strains demonstrated no differences. This reduced rate of growth was observed when overnights were cultured in both glucose and ethanol prior to experimentation (Figures 5.6k and l). Additionally, Bertram *et al.* (1996) showed that *S. cerevisiae ADH* mutants were resistant to growth on the respiratory inhibitor antimycin A, however it can be seen in Figure 5.6m that all heterozygous knockout mutants of *ADH2* in *C. albicans* grow on antimycin A in a comparable manner to the wild-type strain SC5314, which is also resistant to antimycin A at a concentration of 1 µg/ml. Bertram *et al.* (1996) only observed sensitivity to growth on ethanol and resistance to antimycin A in *S. cerevisiae* knockout strains lacking all three alcohol dehydrogenase genes implying that a mechanism of functional redundancy occurs. Therefore it could hypothesised that in the *C. albicans ADH2* heterozygous knockout mutants, the remaining allele or the other alcohol dehydrogenase gene *ADH1* may be masking any functional consequences of the allele that is missing.

Additionally, as GO analysis of the genes identified with AEI from the Bruno *et al.* (2010) data-set revealed an over representation of genes involved in oxidation-reduction processes, the *ADH2* heterozygous knockout mutants were assayed for their growth under oxidative stress. Figure 5.6m shows that under all three conditions tested the knockout strains had comparable growth to the wild-type strain suggesting that the *ADH2* alleles do not have separate functions in response to oxidative stress. However, as was the case for growth in ethanol, this could be due to functional redundancy of other alcohol dehydrogenase genes.

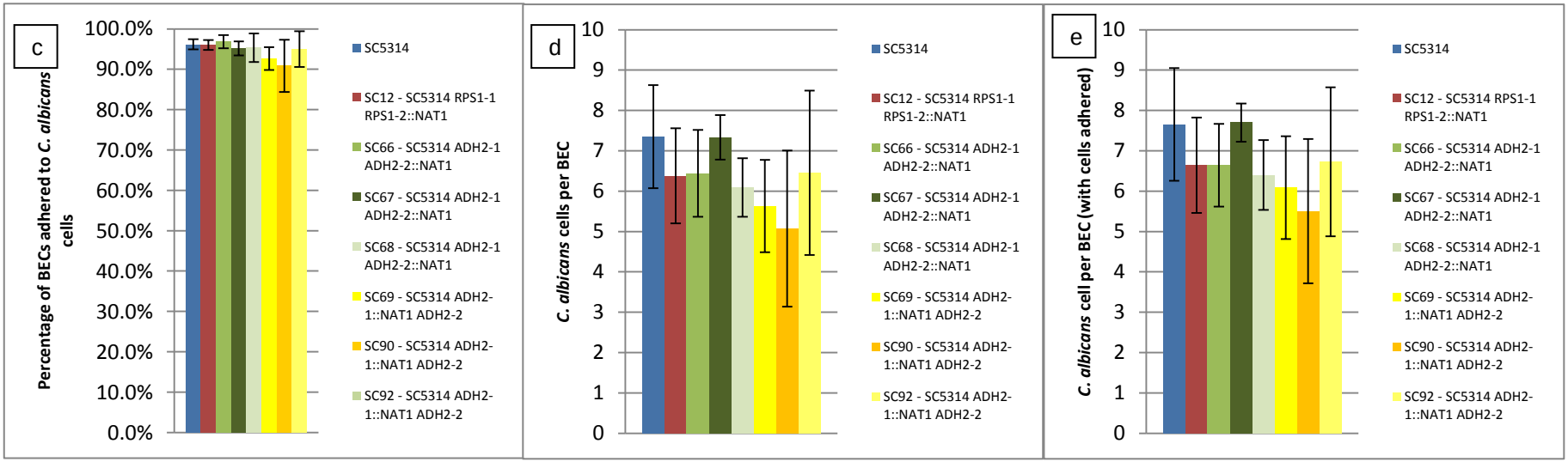
Table 5.12 Average generation times, times to maximum inflection and end-point optical densities of *ADH2* heterozygous knockout mutants at 30 °C and 37 °C (± one standard deviation)

Growth Curve (from Figure 5.6)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	115.93 ± 5.99	490.88 ± 102.58	1.25 ± 0.12
	SC12	118.90 ± 9.05	491.75 ± 132.56	1.24 ± 0.27
Allele 1 knockouts	SC69	119.51 ± 5.19	495.54 ± 76.21	1.18 ± 0.08
	SC90	123.43 ± 10.36	454.67 ± 163.52	1.23 ± 0.18
	SC92	114.45 ± 6.64	472.92 ± 65.54	1.29 ± 0.10
Allele 2 knockouts	SC66	114.13 ± 7.42	494.86 ± 158.93	1.19 ± 0.25
	SC67	116.45 ± 7.95	470.46 ± 69.03	1.27 ± 0.12
	SC68	115.78 ± 6.90	472.50 ± 162.12	1.24 ± 0.21
b) 37 °C	SC5314	91.40 ± 9.47	328.13 ± 38.81	1.16 ± 0.26
	SC12	90.12 ± 5.41	311.79 ± 23.29	1.30 ± 0.19
Allele 1 knockouts	SC69	82.63 ± 16.72	307.42 ± 58.50	1.10 ± 0.21
	SC90	96.02 ± 14.02	368.08 ± 93.34	1.27 ± 0.14
	SC92	95.21 ± 14.07	358.17 ± 120.43	1.30 ± 0.16
Allele 2 knockouts	SC66	79.75 ± 20.73	309.17 ± 37.01	1.18 ± 0.26
	SC67	91.68 ± 18.96	339.50 ± 116.15	1.25 ± 0.19
	SC68	82.61 ± 17.22	355.83 ± 104.70	1.26 ± 0.16

Figure 5.6 Phenotypic assays of *ADH2* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC69, SC90 and SC92 = knockout of “allele one” (yellow) and SC66, SC67 and 68 = knockout of “allele two” (green).

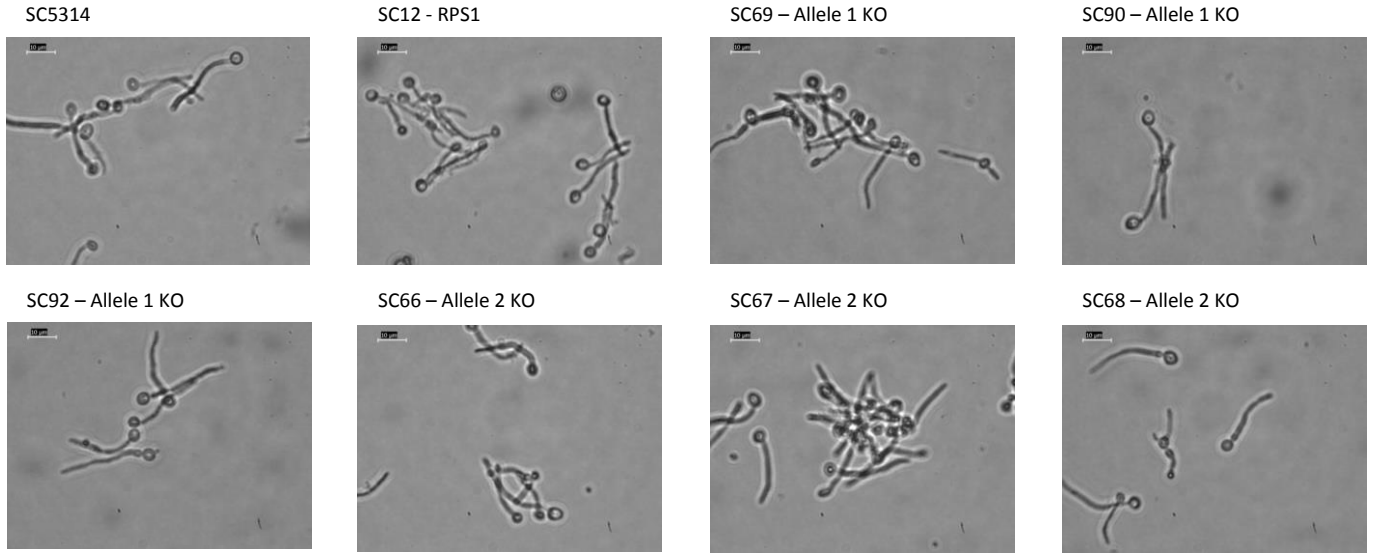


a) Growth rate at 30 °C. b) Growth rate at 37 °C.



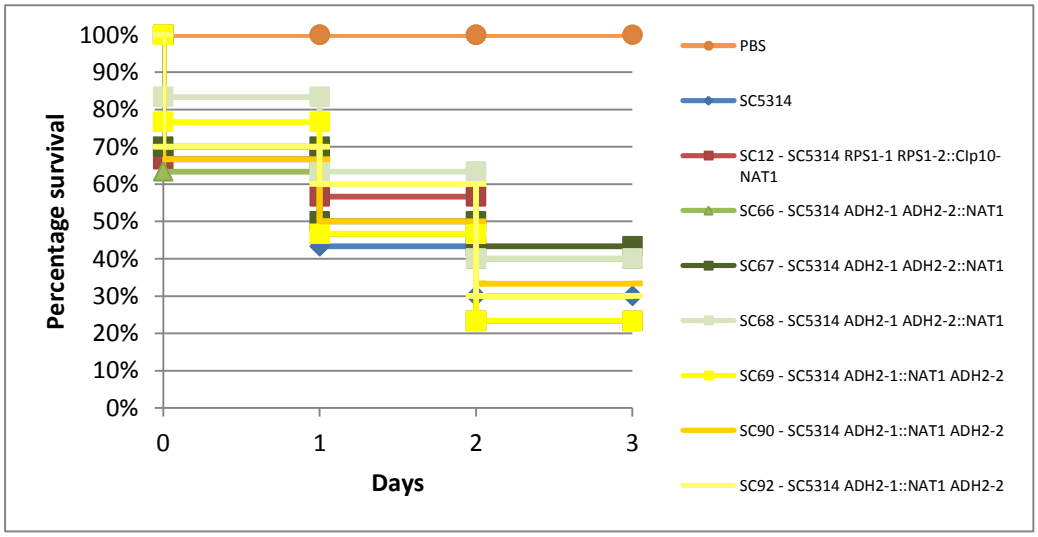
c) Percentage of BECs adhered to *C. albicans* cells. d) Number of *C. albicans* cells per BEC. e) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

f



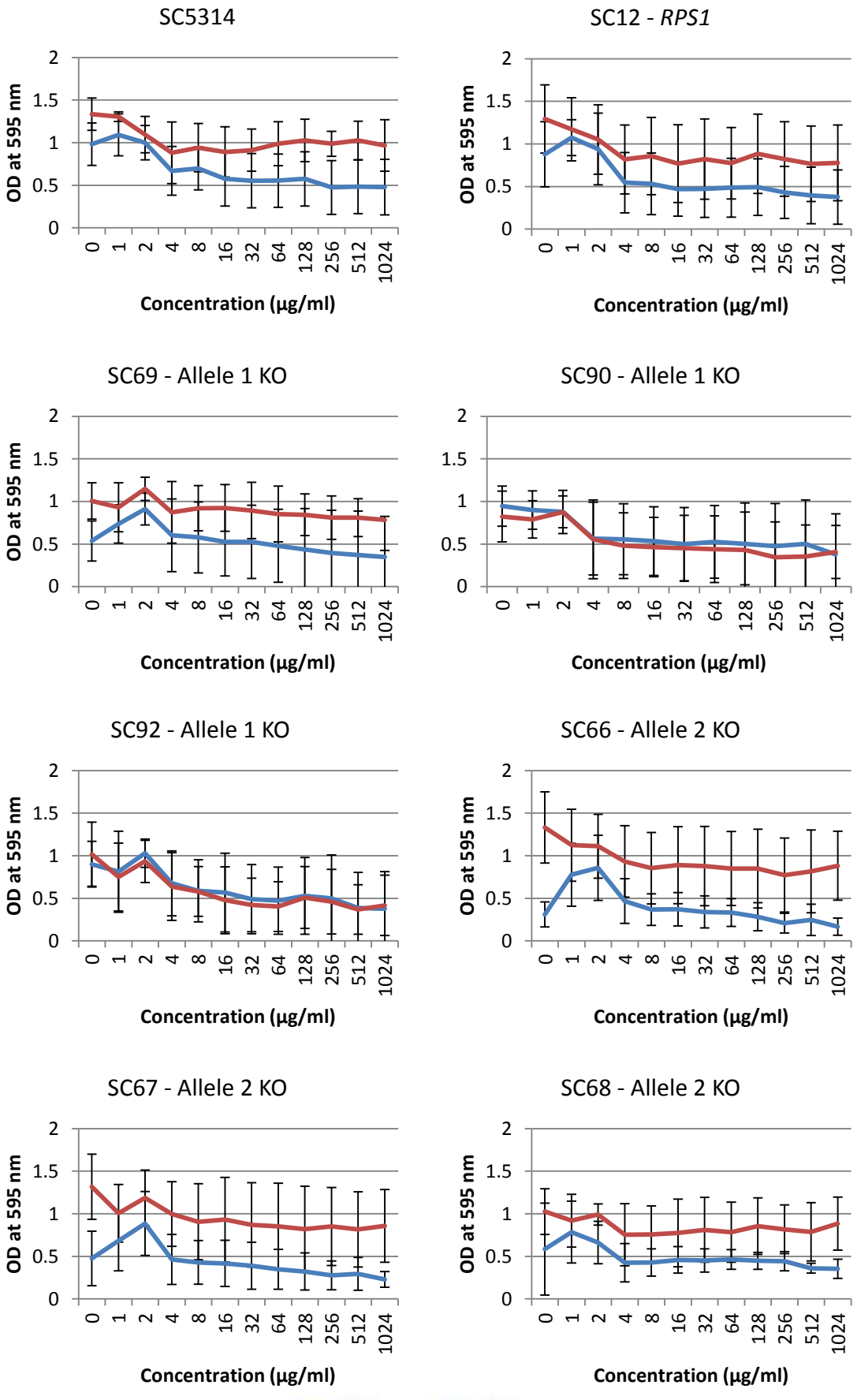
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

g

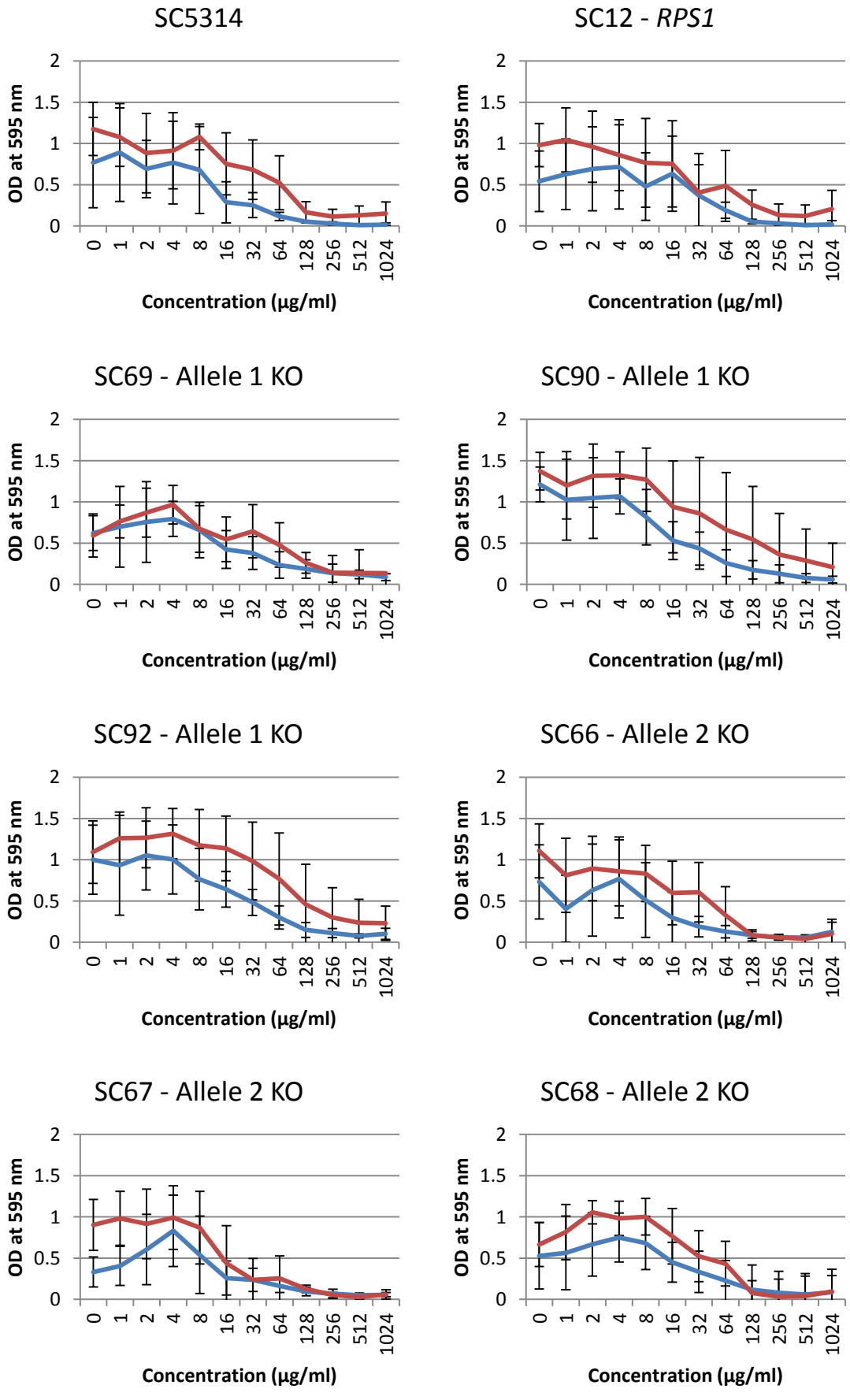


g) *Galleria mellonella* virulence assay at a cell concentration of 2×10^7 cells/ml.

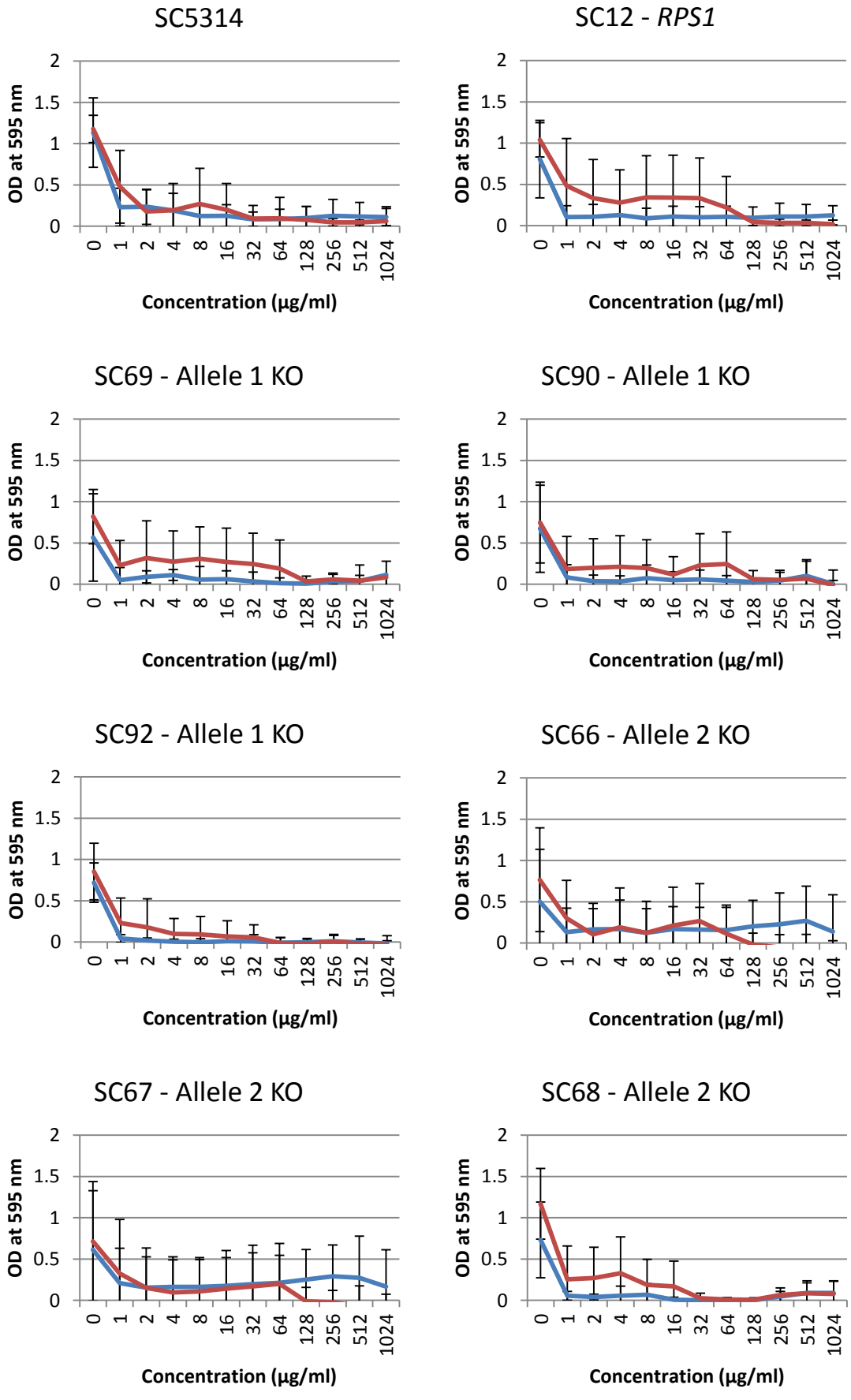
h) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



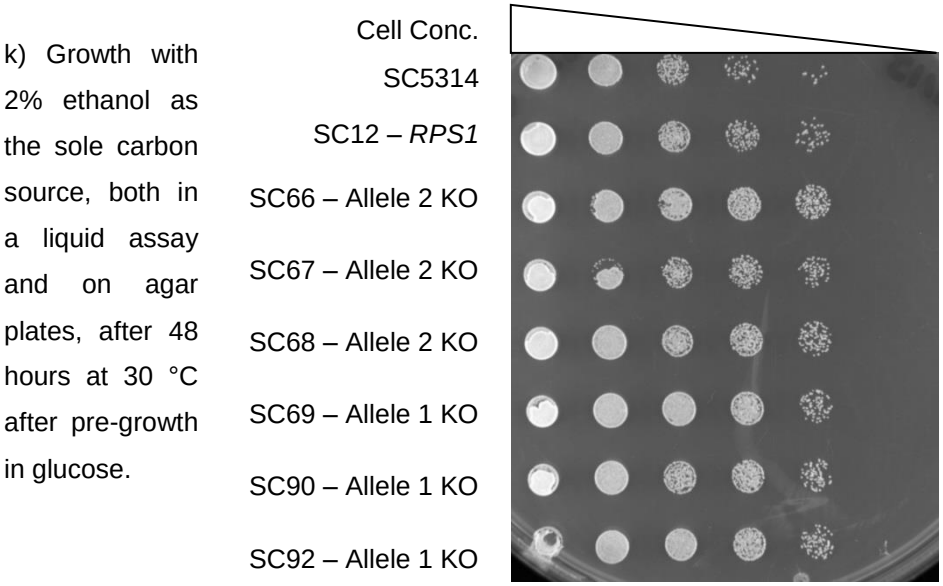
i) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.

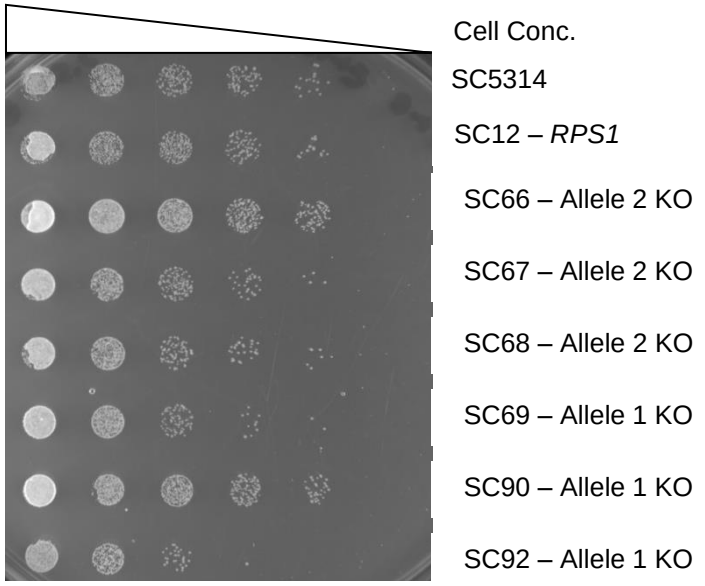


j) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.

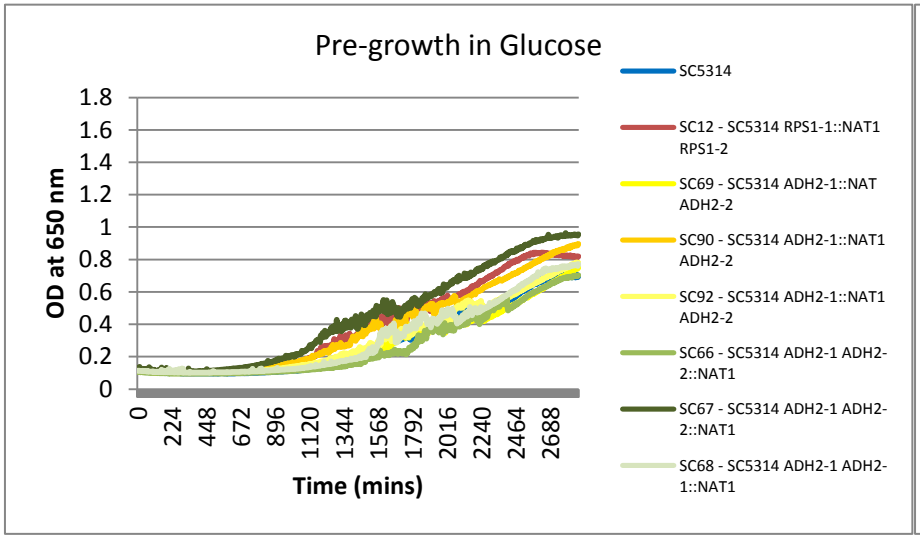


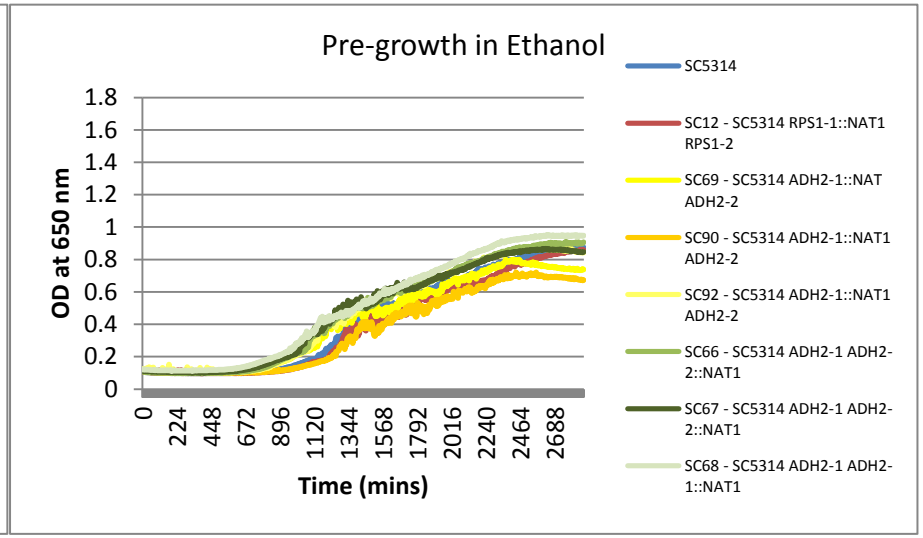
k) Growth with 2% ethanol as the sole carbon source, both in a liquid assay and on agar plates, after 48 hours at 30 °C after pre-growth in glucose.

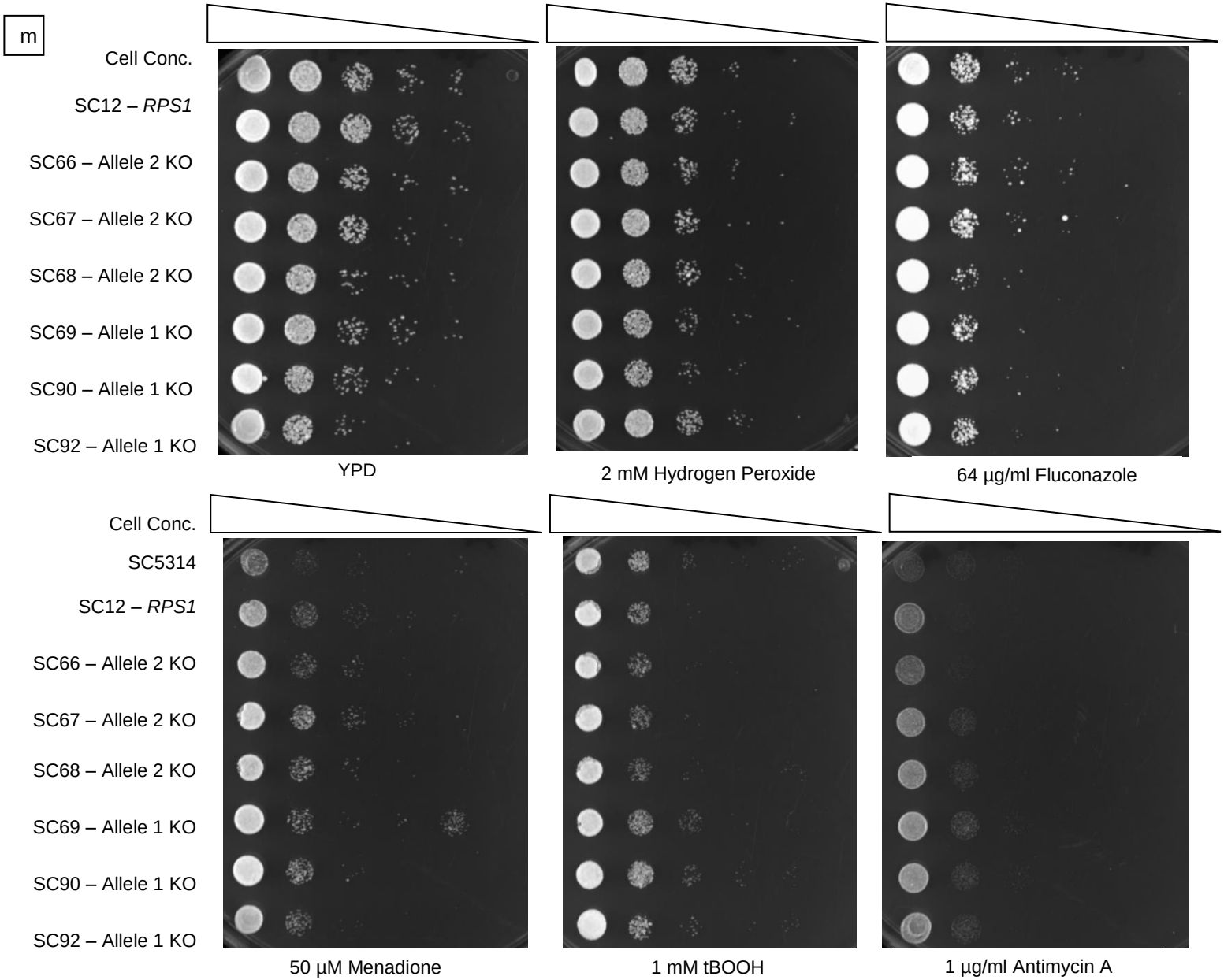




l) Growth with 2% ethanol as the sole carbon source, both in a liquid assay and on agar plates, after 48 hours at 30 °C after pre-growth in ethanol.







m) Growth under stress conditions – YPD agar with 2mM hydrogen peroxide, 50 µM menadione, 1 mM tBOOH and 1 µg/ml antimycin A at 24 hours. YPD + 64 µg/ml fluconazole agar plates at 48 hours YPD agar alone as a control at 24 hours. Cell concentrations range in tenfold dilutions from 1 x 10^7 to 1 x 10^3 cells/ml.

5.3.7.2 *GPX1* Phenotypic Screening

The *GPX1* gene in *Candida albicans* is currently uncharacterised but has been predicted to be a putative thiol peroxidase (The *Candida* genome database (Inglis *et al.*, 2012)). Both protein and gene expression levels have been shown to be up-regulated in response to hydrogen peroxide (Enjalbert *et al.*, 2006, Kusch *et al.*, 2007) and gene expression levels have been shown to be up-regulated in response to 2-amino-nonyl-6-methyoxyl-tetralin muriate, a drug with antifungal activity which causes high levels of endogenous reactive oxygen species (Liang *et al.*, 2011). Therefore, the growth of the heterozygous knockout mutants was assayed under three oxidative stress conditions, hydrogen peroxide, menadione and tBOOH (as described in section 2.14.4). Figure 5.7l shows that all strains grow in a similar manner to the wild-type strain SC5314, suggesting that the alleles do not solely have functions in the response to oxidative stress.

GPX1 has also been linked to the response to antifungal drug treatments, specifically azoles, Gpx1 protein expression levels are increased upon over expression of the drug transporter genes *MDR1, CDR1* and *CDR2* (Hoehamer *et al.*, 2009), *GPX1* has been shown to be differentially expressed in fluconazole susceptible and resistant strains (Garaizar *et al.*, 2006), and gene expression has been shown to be up-regulated by *TAC1* in azole resistant strains, the transcription factor which activates the *CDR* transporter genes (Liu *et al.*, 2007). Interestingly, gene expression is also increased upon treatment with milbemycin, an ABC drug transporter inhibitor (Silva *et al.*, 2013). Here all heterozygous knockout strains responded in a similar manner to the wild-type strain SC5314 when grown in the antifungal drugs 5-flucytosine (Figure 5.7j). Growth in amphotericin B shows an increased sensitivity at 48 hours for all strains (Figure 5.7k), however this is not supported by growth on agar plates containing 1 µg/ml amphotericin B (Figure 5.7l) suggesting that this observation is due to inter plate variability. In response to growth in fluconazole, all three isolates lacking allele one have a marginally reduced optical density across all drug concentrations at 48 hours when compared to the wild-type strain SC5314. On the other hand, SC81 and SC107, the heterozygous knockouts of allele 2, have a similar growth in fluconazole to SC5314 (Figure 5.7i). However growth at

48 hours on solid media containing 64 µg/ml of fluconazole does not support these results, with all strains growing comparably (Figure 5.7l)

The results of other general phenotypic assays showed few significant differences between the wild-type strain SC5314 and the knockouts of either allele. Growth at 30 °C (Figure 5.7a) shows that all strains have a similar time to maximum inflection. Although statistically, all three isolates of the allele one knockout have a significantly higher end-point optical density (ANOVA followed by Dunnett's test, $p < 0.05$; Table 5.13) upon observation of the graph in Figure 5.7a the biological significance of such a difference is not immediately apparent. The knockout strain of allele two, SC81, also has a significantly longer generation time (ANOVA followed by Dunnett's test, $p < 0.05$; Table 5.13). However, this difference is minimal and the second isolate lacking allele two does not have a significantly different generation time. Growth at 37 °C (Figure 5.7b) shows that all strains have a similar generation time. SC74, a knockout of allele one, and SC107, a knockout of allele two, both have statistically shorter times to maximum inflection (ANOVA followed by Dunnett's test, $p < 0.05$; Table 5.13), however Figure 5.7b again shows that this difference is minimal. However, SC81, a knockout of allele two, has a highly significant reduction in end-point optical density (ANOVA followed by Dunnett's test, $p < 0.05$; Table 5.13). As the other isolate of an allele two knockout, SC107, did not reproduce this difference in end-point optical density, a third isolate was constructed (SC108) and the growth curve at 37 °C was repeated. This strain showed no difference in end-point optical density (Figure 5.7c) and therefore it can be concluded that the growth differences of SC81 are likely to be due to secondary mutations elsewhere in the genome and not due to the loss of a copy of *GPX1.*

The ability to adhere to buccal epithelial cells did not differ significantly from the wild-type strain SC5314 for any of the knockouts analysed for any of the three measures tested (two sample t-test, $p < 0.05$; Figures 5.7d - f). The strains also did not differ in their ability to switch to hyphae (Figure 5.7g). Marginal differences were observed in the virulence of heterozygous knockout mutants in a *Galleria mellonella* infection model, therefore the total number of *Galleria* injected for each strain was increased to 50. Although strain SC74 (knockout of allele one) and SC107 (knockout of allele two) show a statistically lower

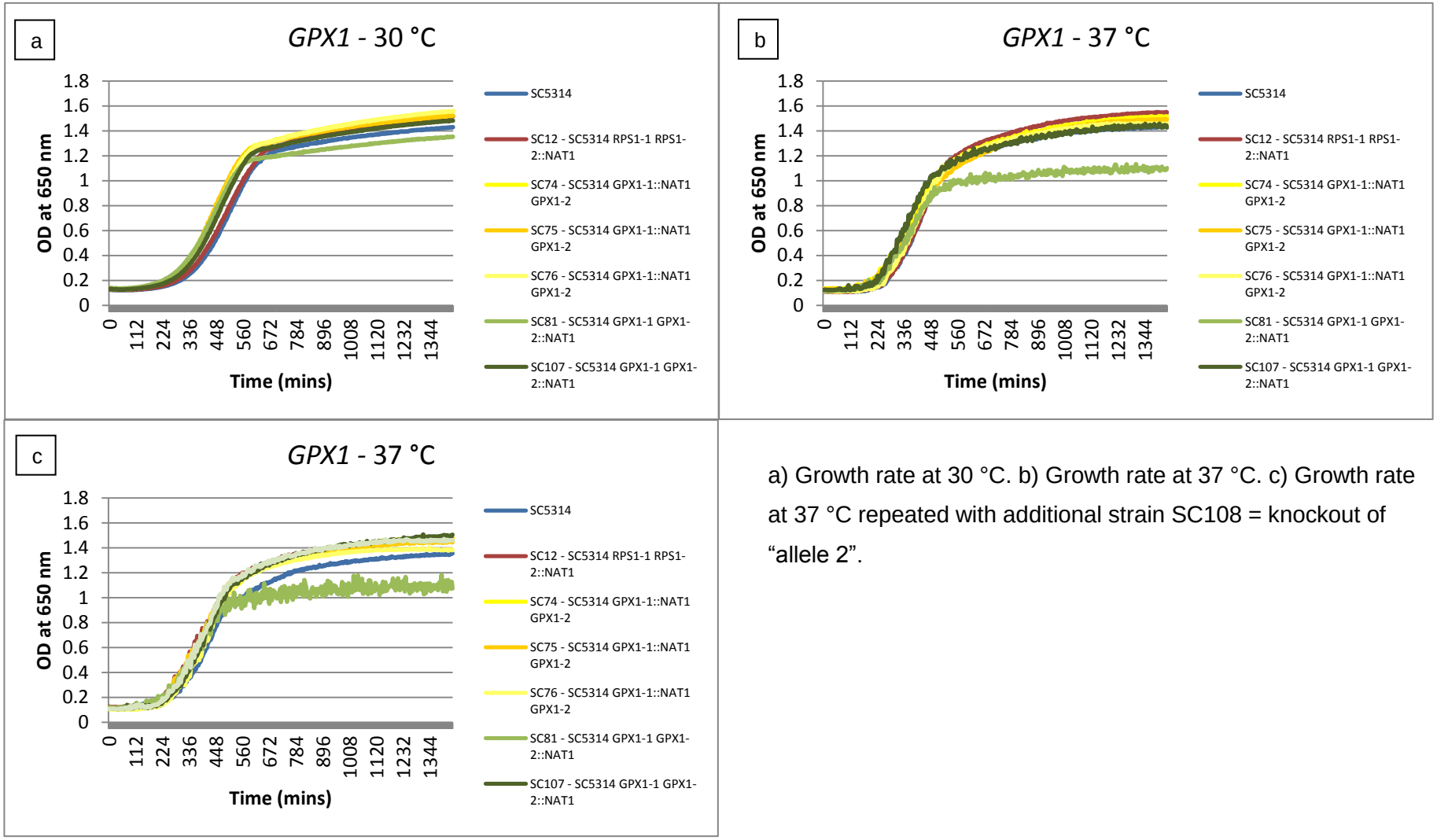
percentage survival than the wild-type strain SC5314 at 2 x 10^7 cells/ml (Kaplan-Meier, d.f. = 2, p = 0.001 and p = 0.039 respectively), this difference is not observed across other cell concentrations or across other isolates of the heterozygous knockout mutants, suggesting that the alleles themselves do not differ in their virulence. Any differences observed could be attributed to natural death of the *Galleria mellonella* used, which is reflected in the lower survival of the PBS control than normal in these particular assays.

Table 5.13 Average generation times, times to maximum inflection and end-point optical densities of *GPX1* heterozygous knockout mutants at 30 °C and 37 °C (± one standard deviation)

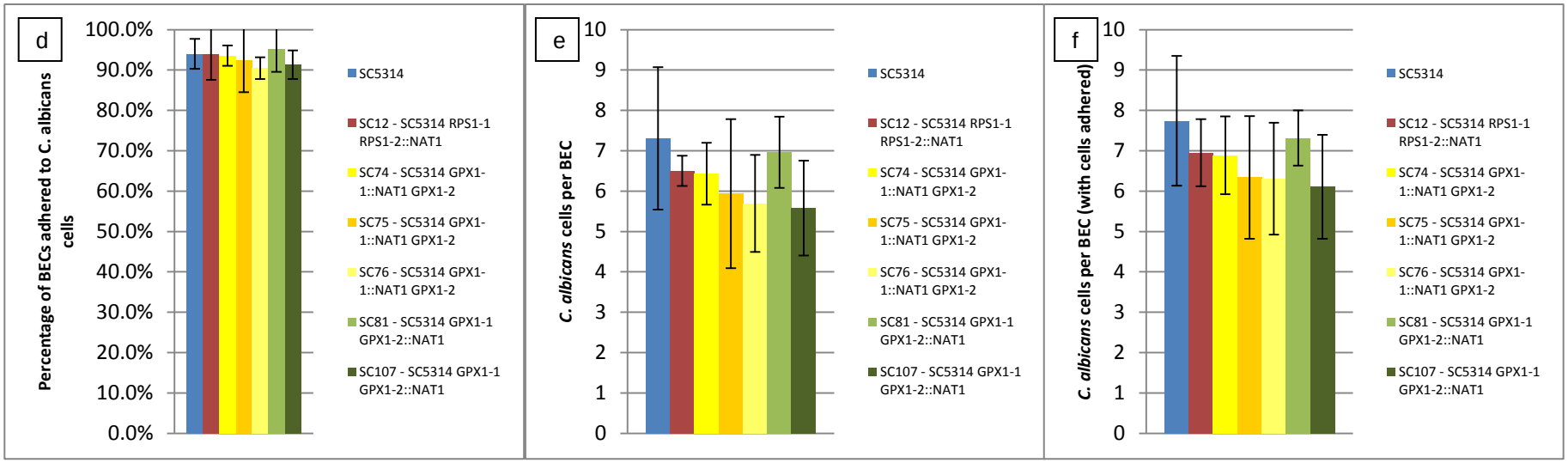
Growth Curve (from Figure 5.7)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) 30 °C	SC5314	120.43 ± 4.67	373.41 ± 36.52	1.33 ± 0.13
	SC12	124.22 ± 9.20	336.22 ± 42.55	1.41 ± 0.10
Allele 1 knockouts	SC74	118.67 ± 6.76	336.22 ± 53.30	1.43* ± 0.05
	SC75	127.79 ± 7.09	320.25 ± 30.67	1.42* ± 0.09
	SC76	123.23 ± 9.57	331.63 ± 38.24	1.44* ± 0.06
Allele 2 knockouts	SC81	131.75* ± 8.79	346.50 ± 83.18	1.26 ± 0.12
	SC107	123.98 ± 4.98	337.75 ± 25.59	1.39 ± 0.10
b) 37 °C	SC5314	67.82 ± 14.39	318.28 ± 26.79	1.37 ± 0.18
	SC12	66.94 ± 13.58	308.44 ± 22.57	1.46 ± 0.03
Allele 1 knockouts	SC74	66.52 ± 13.49	271.47* ± 30.72	1.42 ± 0.05
	SC75	68.56 ± 20.40	306.47 ± 105.57	1.41 ± 0.06
	SC76	66.73 ± 21.67	304.72 ± 35.08	1.40 ± 0.04
Allele 2 knockouts	SC81	78.44 ± 25.05	293.75 ± 45.03	1.06* ± 0.20
	SC107	59.97 ± 21.28	270.81* ± 21.83	1.37 ± 0.21
c) 37 °C Repeat	SC5314	85.22 ± 11.14	343.58 ± 82.73	1.28 ± 0.14
	SC12	70.83 ± 19.66	289.63 ± 32.56	1.40 ± 0.07
Allele 1 knockouts	SC74	78.17 ± 20.57	312.36 ± 41.15	1.40 ± 0.04
	SC75	78.28 ± 21.37	289.33 ± 28.76	1.38 ± 0.07
	SC76	77.03 ± 15.41	325.79 ± 28.65	1.35 ± 0.11
Allele 2 knockouts	SC81	106.35 ± 61.98	353.21 ± 101.58	1.04* ± 0.40
	SC107	77.43 ± 15.12	315.29 ± 45.99	1.42 ± 0.04
	SC108	79.98 ± 12.70	306.54 ± 37.05	1.40 ± 0.05

* Significantly different measurements from SC5314, identified by ANOVA followed by *post-hoc* analysis using a Dunnett's test, at $p < 0.05$, are annotated with an asterisk.

Figure 5.7 Phenotypic assays of *GPX1* heterozygous knockout mutants. SC5314 = wild-type strain (blue). SC12 = control strain with NAT cassette at *RPS1* locus (red). SC74, SC75 and SC76 = knockout of "allele one" (yellow) and SC81 and SC107 = knockout of "allele two" (green).

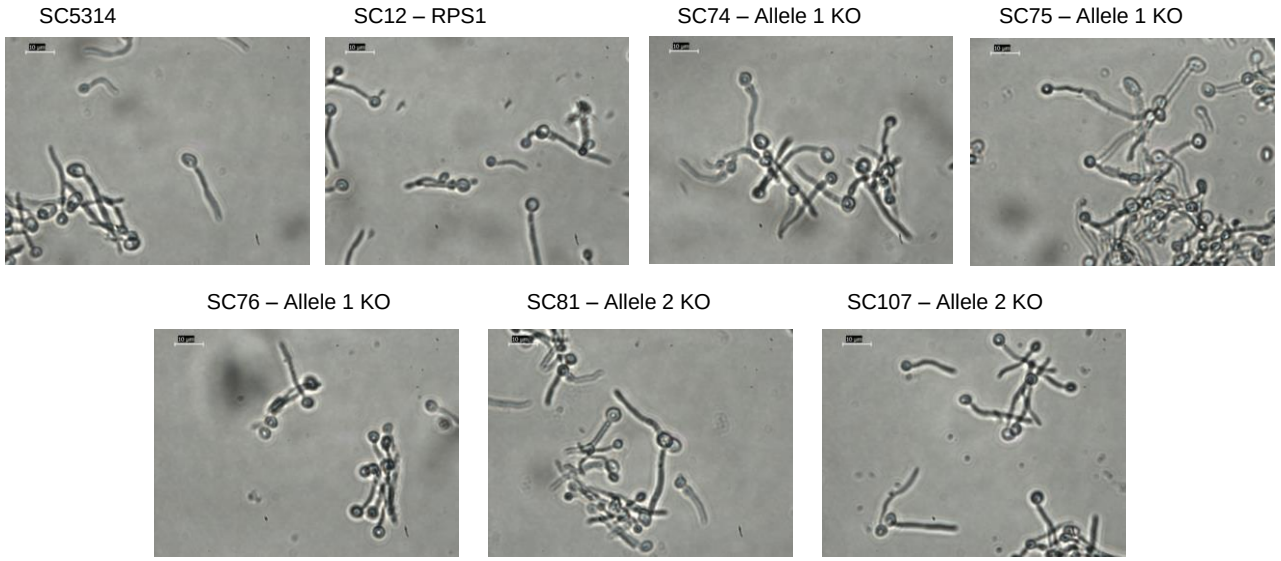


a) Growth rate at 30 °C. b) Growth rate at 37 °C. c) Growth rate at 37 °C repeated with additional strain SC108 = knockout of "allele 2".



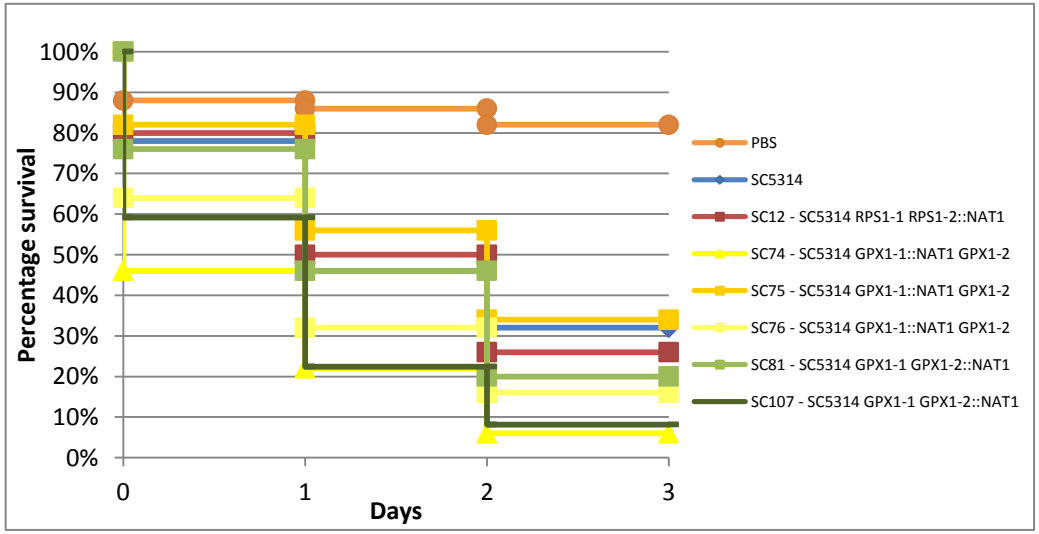
d) Percentage of BECs adhered to *C. albicans* cells. e) Number of *C. albicans* cells per BEC. f) Number of *C. albicans* cells per BECs (with cells adhered). Error bars = ± one standard deviation.

g



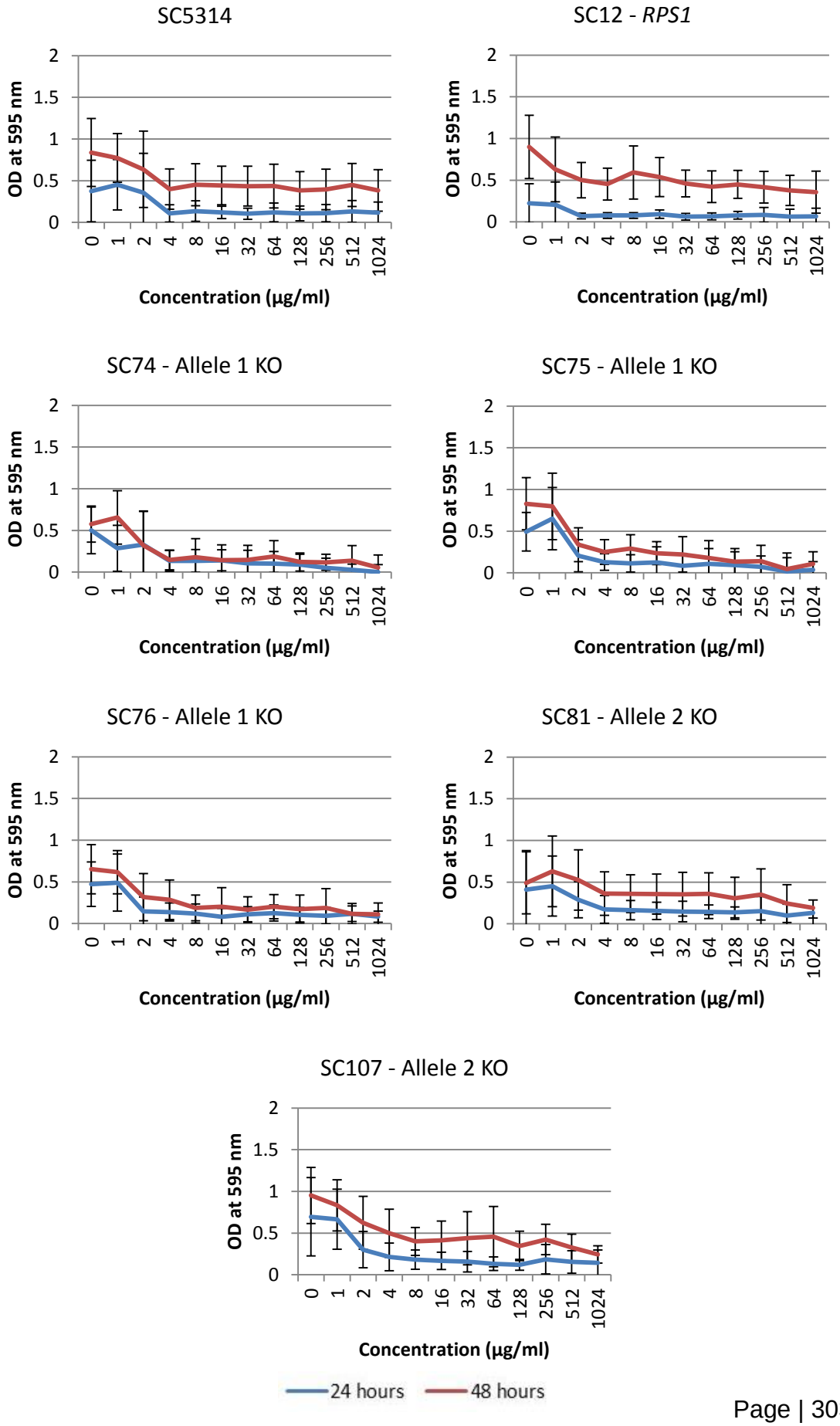
f) Induction of hyphae. Strains were exposed to 5% foetal calf serum and incubated at 37 °C. Figures shows cells at 120 minutes however samples were taken every 15 minutes. Scale bar = 10 µm.

h

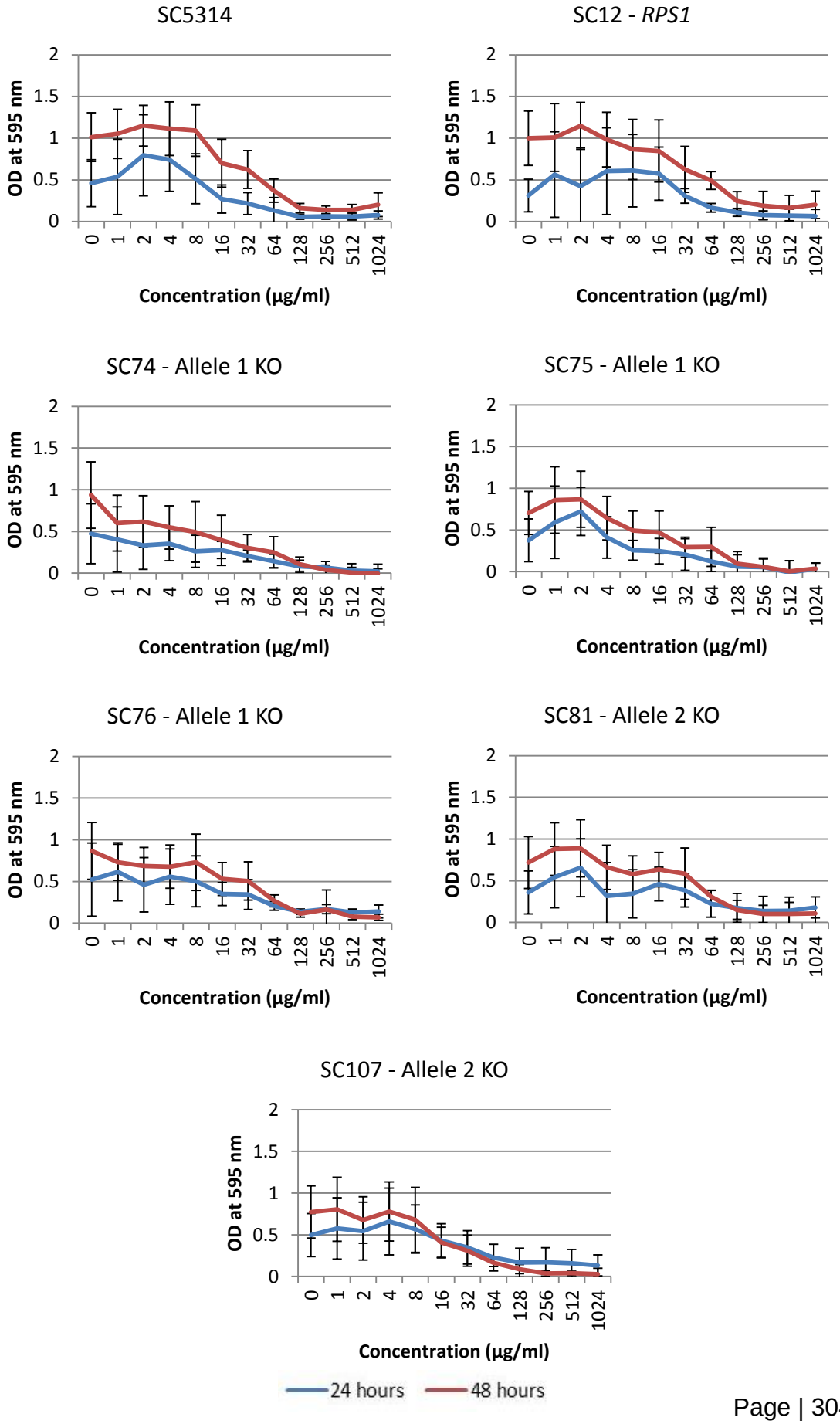


g) *Galleria mellonella* virulence assay at a cell concentration of 2 x 10^7 cells/ml.

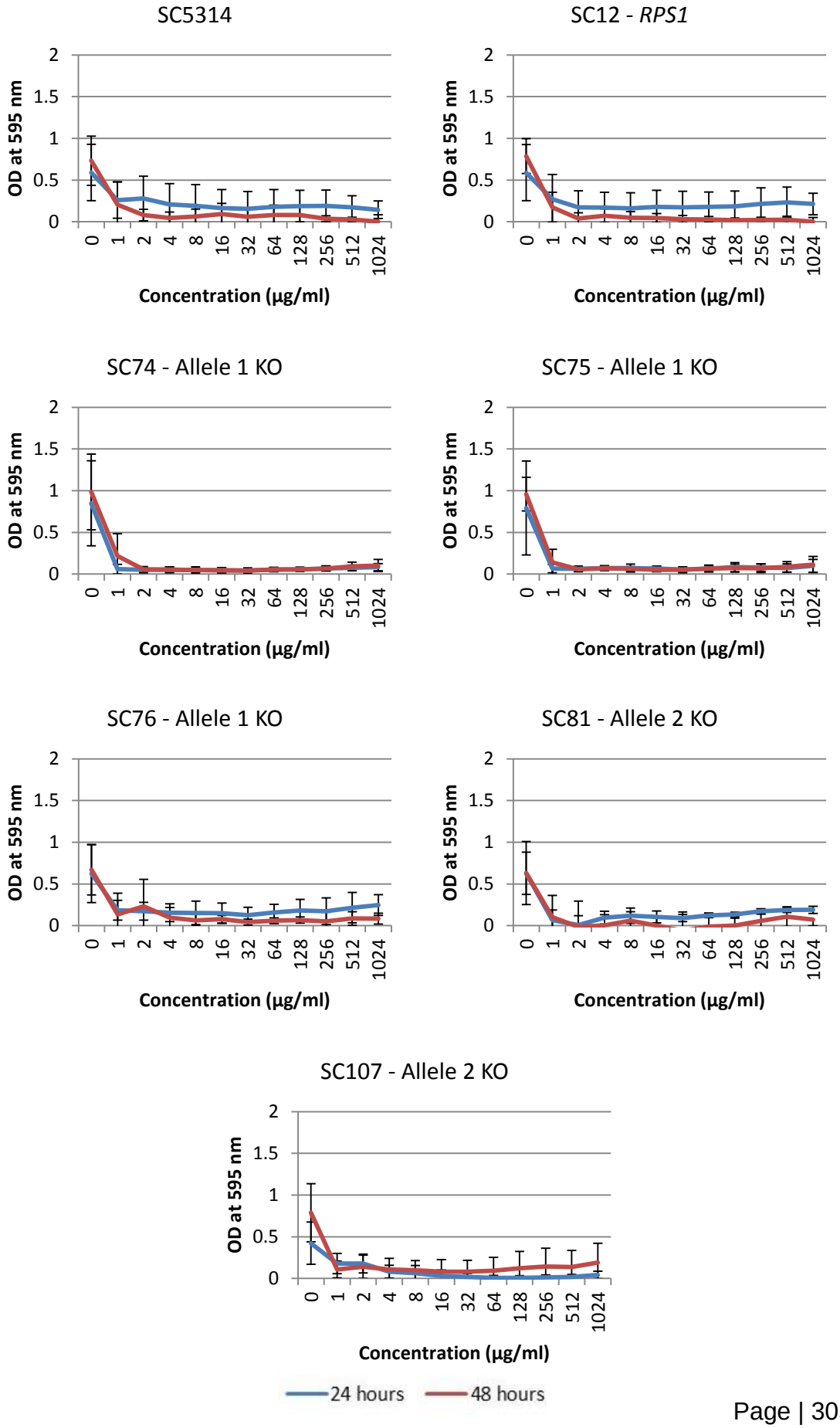
i) Growth in response to fluconazole. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.

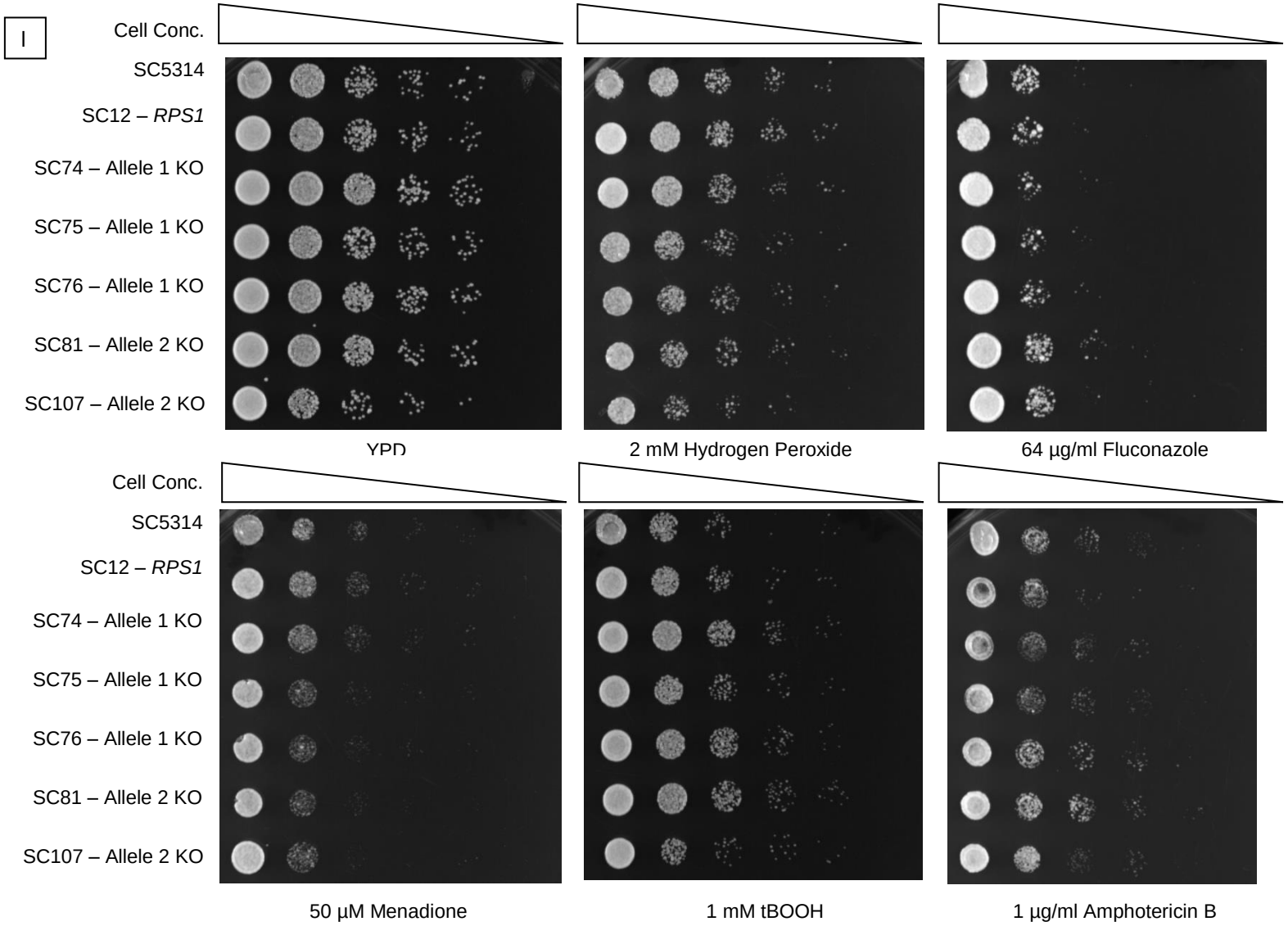


j) Growth in response to 5-flucytosine. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.



k) Growth in response to amphotericin b. Concentrations range from 0 – 1024 µg/ml. Blue = OD at 595 nm at 24 hours. Red = OD at 595 nm at 48 hours. Error bars = ± one standard deviation.





l) Growth under stress conditions – YPD agar with 2mM hydrogen peroxide, 50 µM menadione, 1 mM tBOOH and 1 µg/ml amphotericin B at 24 hours. YPD agar with 64 µg/ml fluconazole at 48 hours. YPD agar alone as a control. Cell concentrations range in tenfold dilutions from 1 x 10^7 to 1 x 10^3 cells/ml.

5.3.7.3 Phenotypic Screening Summary

Table 5.14 summarises the results for all of the phenotypic screening carried out in this chapter.

Table 5.14 Summary of phenotypic screening. Strains with significant results are highlighted with arrows indicating the direction of the results.

GENE	*ADH2*		*GPX1*	
ALLELE KO	1	*2*	*1*	*2*
Growth at 30 °C				
Generation Time				↑SC81
Endpoint Density			↑SC74	
			↑SC75	
			↑SC76	
Time to Maximum Inflection				
Growth at 37 °C				
Generation Time				
Endpoint Density				↓SC81
Time to Maximum Inflection				
Adhesion				
Hyphal Induction				
Virulence (at 2 x 10^7 cells/ml)			↓SC74	↓SC107
Sensitivity to Fluconazole (liquid)	↑SC90		↑SC74	
	↑SC92		↑SC75	
			↑SC76	
Sensitivity to Fluconazole (solid)				
Sensitivity to 5-Flucytosine				
Sensitivity to Amphotericin B				
Growth in Ethanol as Sole Carbon Source (liquid)	↓All including SC5314 & SC12			
Growth in Ethanol as Sole Carbon Source (solid)				
Growth on Antimycin A				
Growth on Oxidative Stress				

5.3.8 Homozygosity of *RPS7A* and orf19.5648

As was the case in chapter four for *RCK2,* only heterozygous knockout mutants of one allele were constructed for the genes *RPS7A* and orf19.5648, despite a total of eight strains being produced for *RPS7A* and 14 strains produced for orf19.5648. The probability that either allele is knocked out is equal, and is therefore 0.5. Therefore the probabilities that such a high number of a single knockout strain is made is significantly unlikely, as follows for *RPS7A* and orf19.5648:

$$\textit{RPS7A} - 0.5^{8} = 3.9063 \times 10^{-3}$$

$$\text{orf19.5648} - 0.5^{14} = 6.1035 \times 10^{-5}$$

This observation suggests that these genes are either homozygous or the allele that cannot be knocked out is essential. To confirm the heterozygosity of these genes conventional cloning techniques were used, as was the case for *RCK2.* This identifies if any SNPs are just a consequence of errors in the reference genome. The methodology used is described in section 2.8. To summarise, the gene sequences for both genes were amplified using colony PCR from the wild-type strain SC5314 with the assumption that both alleles would amplify with equal efficiency. These PCR fragments were ligated into a pGEM-T easy vector and transformed into *E. coli* cells. Only the sequence from one allele is ligated and transformed per single *E. coli* cell. Positive transformants were selected based on disruption of the *LacZ* operon and antibiotic resistance. Allele sequences were then amplified via PCR and sent for conventional Sanger sequencing, with a probability of 0.5 that the sequence is allele one and 0.5 that the sequence is allele two. In both cases, all seven positive transformations came back matching the same sequence; allele two for *RPS7A* and allele one for orf19.5648. Based on the probability of 0.5 that the sequence is either allele, the probability that all seven strains contained the same allele equates to:

$$0.5^{7} = 7.8125 \times 10^{-3}$$

With such a small probability of this occurring by chance, and all of the heterozygous knockout strains lacking the same allele, it is sensible to assume that both *RPS7A* and orf19.5648 are in fact homozygous genes with an identical sequence for both alleles. *RPS7A* alleles match the sequence of allele two (Appendix I Figure V) and the majority of orf19.5648 polymorphisms match the sequence of allele one (Appendix I Figure VI). Therefore, any differences in allele expression levels observed for these genes are due to reads aligning to erroneous SNPs in the reference genome producing false positive levels of AEI.

5.4 Discussion

This chapter has aimed to develop a new computational pipeline to identify AEI in *Candida albicans* using a diploid reference genome. From here, condition-specific levels of AEI were investigated using RNA sequencing data obtained from various sources. The functional consequence of AEI were further investigated using the results from this analysis and phenotypic screening of heterozygous knockout mutants.

5.4.1 Developing a Computational Pipeline to Identify AEI

The computational pipeline developed here has been used to identify allelic expression imbalance using a number of RNA sequencing data-sets as follows: re-analysis of the RNA sequencing data from chapter three where the wild-type strain SC5314 was grown in rich media, growth of the wild-type strain SC5314 under 11 different conditions as described by Bruno *et al.* (2010), and growth of SC5314 whilst being co-cultured with the oral bacterium *Streptococcus gordonii.* Variable numbers of genes were identified from analysis of each data-set with 21 genes identified from the data from chapter three, 27 genes from analysis of the data from Bruno *et al.* (2010), and 123 genes from analysis of the three co-culture conditions.

As discussed briefly in section 5.3.2, analysis of RNA sequenced using the older Illumina GAII platform identified an unexpectedly small number of genes with AEI. It can be assumed here that the number of reads produced from this platform is not high enough to surpass the level of stringent statistical tests used. This is evident from analysis of the co-culture data-sets, where a much larger number of genes were identified with significant AEI. This data-set was sequenced using the Illumina HiSeq2500 platform, which has previously been shown to produce a higher number of reads than the Genome Analyser (Minoche *et al.*, 2011). Evidence has been seen that supports this hypothesis, with greater sequencing depth leading to a greater statistical power for identification of differentially expressed genes (Zhang *et al.*, 2014). For future use of the computational pipeline, it may be beneficial to reduce the stringency of the statistical testing used to reduce the number of false negative results. However, identifying which statistical test provides the most accurate results would be reliant upon an allele specific qPCR system which, as discussed in

chapter three, is problematic for a number of reasons already discussed (sections 3.4.4). Alternatively, a simulated data-set could be constructed with known levels of allelic expression imbalance to test for the most appropriate statistical measures.

RPKM has previously been challenged as a poor normalisation method for RNA sequencing with evidence for bias against factors such as gene length and GC content (Bullard *et al.*, 2010a, Zheng *et al.*, 2011). Therefore it could be suggested that the statistical testing using the raw read counts and DESeq (Anders and Huber, 2010) may be a more appropriate method. Alternatively, a different normalisation technique could be applied to the data. TMM (trimmed mean of M) values have been shown to be a simple and effective method for calculating relative RNA levels which accounts for differences in total RNA amounts of samples (Robinson and Oshlack, 2010) which may be appropriate in this case as entirely unrelated RNA sequencing samples are being compared.

Surprisingly, re-analysis of the RNA sequencing data from chapter three did not identify AEI in the same genes, with just 14 genes identified by both computational pipelines. This may also be due to the high stringency of statistical testing used, causing false negative results from the re-analysis of the data-set. Additionally this difference in the numbers of genes identified could be explained by differences in the computational pipeline as the original pipeline aligned reads to both SNPs and INDELs. It was shown here that this caused a bias in the results with an over representation of high expression values from longer alleles and therefore reads were aligned to SNP regions only.

Analysis of differing growth conditions showed, surprisingly, that AEI does not appear to be changeable under different growth conditions with the same allele always showing higher allele expression for each gene. It can be hypothesised from this observation that expression of the favoured allele is maintained consistently across all conditions tested. However the fold changes in allele expression do differ on some occasions, indicating that the exact expression levels may alter dependent upon condition. This may indicate that AEI has a biological importance and is therefore unchanged; however if this was the case,

the same genes should always be identified under all conditions tested which was not observed. Therefore the level of noise in the RNA sequencing data may also be an influential factor in the results, accounting for the changing fold difference in expression. Due to this noise, a proportion of genes will be statistically likely to be identified with AEI as an artefact. The fact that the same gene is identified in some but not all conditions may be a consequence of bias skewing expression level measurements always in favour of one allele, such as GC content, as discussed in chapter three.

Conflicting evidence to the idea that AEI is consistent is seen from the analysis carried out in section 5.3.5, where genes with uneven changes in allele expression between growth conditions were identified. Here, it suggests that allele expression levels are variable. Again, this could be a consequence of noise within the RNA sequencing, and any changes are due to natural variability in the sequencing process. On the other hand, there could exist two separate types of genes with AEI – one set where AEI remains consistent regardless of growth condition, and one set where AEI changes in response to the environment. However some genes, such as *ADH2,* are present in both lists, reducing the likelihood of this hypothesis.

The computational pipeline itself could be altered to improve the accuracy of the results obtained. In alleles with more than one SNP identified, a check could be implemented to ensure all SNPs show the same direction of allelic expression bias, and remove them from analysis if this is not the case. This method has been adopted in a study of AEI in a "super-hybrid" rice species (Zhai *et al.*, 2013) and was also used by Muzzey *et al.* (2013) in their recent identification of AEI in *Candida albicans*. Muzzey *et al.* (2013) also developed the methodology to ensure that high numbers of reads at certain positions did not skew the measure of AEI by determining confidence intervals using bootstrapping (Muzzey *et al.*, 2013). In the follow-up investigation carried out by Muzzey *et al.* (2014) biases related to library preparation and sequencing processes were accounted for by comparing the coverage of allele specific regions to the coverage at non-specific regions, with the idea that systematic errors would increase the disparity between these measures (Muzzey *et al.*, 2014). Again,

use of a simulated data-set with known AEI levels could be used to determine the impact of these changes to the pipeline.

Here we have excluded alleles which differ in length as the earlier results showed an over representation of longer alleles with higher expression levels, with the assumption that this is due to the 3' bias in library preparation (Wilhelm *et al.*, 2008). Although Muzzey *et al.* (2013) also observed expression differences in alleles which differ in length in *C. albicans,* expression differences were proposed to be a consequence of nonsense-mediated decay (NMD) triggered by premature stop codons in the shorter allele. However the list of candidate genes which they investigated for NMD consisted of only 22 genes, of which only 73% (16/22) experienced allelic expression bias. Additionally, when these 16 genes are compared to the analysis here of the Bruno *et al.* (2010) YPD data-set, using the computational pipeline which includes INDELs, only one of these genes are identified. Therefore NMD cannot be conclusively proved as the causative factor of the difference in expression levels seen by alleles which differ in length in our re-analysis of the data from chapter three, where a total of 159 were found to favour the expression of the longer allele.

As with the initial identification of AEI in chapter three, this methodology still suffers from false positive results due to errors in the reference genome. *RCK2* was identified as having significant levels of AEI from analysis of all three RNA sequencing data-sets despite being shown in chapter four that *RCK2* is a homozygous gene with all SNPs being due to errors in the reference genome. This is also the case for *RPS7A* and orf19.5648, as described in section 5.3.9, which were identified as having significant AEI across a high number of conditions. For these genes, AEI is due to reads falsely aligning to SNPs in the reference genome which do not exist, producing false positive measurements of AEI. Although not practical within this study, it would be beneficial to confirm the heterozygosity of all genes with AEI to demonstrate the impact of errors in the reference genome upon this analysis. Another option to avoid these false positive results is to use the phased diploid reference genome of *C. albicans* which has recently been published (Muzzey *et al.*, 2013). The investigation used sequencing of the wild-type strain SC5314 alongside a number of strains homozygous for certain genomic regions to elucidate the full phasing and

identify a total of 69,688 SNPs (an increase of 69% from Assembly 19). Using a similar computational pipeline to here, this phased genome was used as a reference with data from Bruno *et al.* (2010) to measure allele-specific expression levels in comparison to identification of AEI using the Assembly 19 reference genome. The paper states that some genes were detected as false negative when the old Assembly 19 reference genome was used as opposed to the phased reference genome. Unfortunately, the phased reference genome is not yet available for download and therefore we are unable to elucidate the impact it would have upon the pipeline developed here.

5.4.2 Functional Consequences of AEI

Gene Ontology (GO) analysis of the genes identified from analysis of the various RNA sequencing data-sets suggests that the functions of the genes with AEI do not reflect the growth condition. However, when genes identified from all conditions were combined there was a significant over representation of genes involved in oxidation-reduction processes and metal binding and transport. This is supportive of our finding that AEI levels are consistently in favour of the same allele, suggesting that these genes are also involved in the same processes, regardless of the growth condition.

To further investigate if alleles with expression imbalance differ in function, heterozygous knockout strains were constructed for *ADH2* and *GPX1* and were screened for phenotypic differences under a number of different conditions. Results of all screens indicate that the alleles themselves are not functionally distinct, with no assay conclusively showing clear differences between the knockouts and the wild-type strain SC5314. This included screening on oxidative stress conditions, proposed due to the over representation of oxidation-reduction processes in the GO analysis. Combining these results with those found in chapter four suggest that allelic expression imbalance is not linked to distinct allele functions.

However, despite the over-whelming evidence suggesting AEI and allele function are not linked, this conclusion cannot be confirmed solely with the results from this analysis. Phenotypic screening may have missed the conditions which would identify functional differences. For example, the *ADH2*

gene has been identified to be over expressed in hypoxic growth conditions (Setiadi *et al.*, 2006), an observation of particular interest as genes with AEI are shown to have an over representation of genes involved in oxidation-reduction processes. Here it was not practical to test whether the heterozygous knockout mutants were differentially sensitive to growth in this condition but this is certainly an area for further investigation. A large scale phenotypic screen could be used to further increase the chances of finding a condition which demonstrates differences in allele function as also suggested in chapter four, such as the screen used by Homann *et al.* (2009) testing a transcription factor knockout library on 55 conditions (Homann *et al.*, 2009).

Functional redundancy, where the phenotypic consequences of the loss of a gene are compensated for by other closely related genes, has already been observed across the *ADH* genes in *S. cerevisiae* with mutants only displaying phenotypes when lacking all three alcohol dehydrogenase genes (Bertram *et al.*, 1996). Examples of functional redundancy have been reported in *C. albicans* genes such as the phosphatase gene *PTC6* (Yu *et al.*, 2010), the mannosyltransferase *MNN1* gene family (Bates *et al.*, 2013), and *ALS* genes *ALS2* and *ALS4* (Zhao *et al.*, 2005). Therefore, it could be hypothesised that any phenotypic effects presented in our heterozygous knockout mutants are being compensated for either by related genes or by the remaining allele itself.

A further explanation for the lack of evidence suggesting functional differences between the alleles could be found when looking at the protein levels. Recent research in *C. albicans* using ribosome profiling suggests that some cases of AEI at the transcriptional level are buffered and not present at the translational level (Muzzey *et al.*, 2014). This is known as a compensatory relationship between transcription and translation. However, the majority of alleles with transcriptional allelic bias showed a reinforcing relationship at the translational level (Muzzey *et al.*, 2014). Despite attempts to develop a method to monitor allele specific protein expression levels in chapter three, these techniques were shown to lack the sensitivity and precise quantification needed, and therefore we are unable to ascertain here that levels of AEI are present at the level of protein expression.

5.4.3 Conclusion

To conclude, although a computational pipeline was developed to interrogate RNA sequencing data for evidence of allelic expression imbalance, analysis suggests that AEI is maintained consistently regardless of the growth conditions tested. The functional consequence of this AEI is yet to be elucidated, with further phenotypic screening of heterozygous knockout mutants shedding no light on the matter.

Chapter 6: Allelic Expression Imbalance and *Candida albicans*

This body of work has attempted to detail the genome-wide occurrence of allelic expression imbalance in the pathogenic yeast *Candida albicans,* through the use of RNA sequencing and the development of a novel computational pipeline. The functional consequences of this phenomenon were then investigated through analysis of a number of RNA sequencing data-sets collected from *Candida albicans* grown under different conditions, and through targeted construction and phenotypic screening of heterozygous knockout strains. Investigations of the control mechanisms and sequence specific features driving the divergence in allele expression levels were also presented.

6.1 Allelic Expression Imbalance and *Candida albicans*

6.1.1 Identification of AEI

Allelic expression imbalance, or AEI, is the term given to an uneven level of allele expression from a single gene. This can be observed either as differential allele expression or as monoallelic gene expression. At the onset of this research project, AEI was yet to be identified on a genome-wide level in *Candida albicans* despite it being a diploid organism with a publically available annotated reference genome.

However, evidence for variability in allele expression levels has been reported in *C. albicans* on a gene-by-gene basis, suggesting that this phenomenon does occur under some circumstances. For example, the two alleles of the *SAP2* gene have been shown to be differentially regulated during the infection process due to differences in their promoter regions (Staib *et al.*, 2002). A similar story is observed for the drug-resistance gene *MDR1,* where differences in the promoter regions confer differences in allele expression levels (Bruzual and Kumamoto, 2011). And finally, even though the allele sequences of the chitin synthesis gene *CHS7* are homozygous, the promoter regions have been shown

to differ in length directly impacting upon allele expression levels (Sanz *et al.*, 2007).

From here, this project aimed to identify the genome-wide extent of allelic expression imbalance in *Candida albicans.* Chapter three details initial investigations using RNA sequencing data from the wild-type strain SC5314 grown under optimal laboratory conditions demonstrating that 233 genes have significant disparities in allele expression levels. Analysis of this gene set suggested no common functionality or chromosomal location to genes with AEI, leading to the conclusion that AEI occurs in seemingly unrelated genes. The functional consequences of this phenomenon for individual genes were investigated using heterozygous knockout mutant strains and the results of these investigations are further discussed in section 6.2.

As overall gene expression levels in *Candida albicans* have previously been shown to alter in response to the growth environment (Nantel *et al.*, 2002, Enjalbert *et al.*, 2003, Bensen *et al.*, 2004, Enjalbert *et al.*, 2006, Biswas *et al.*, 2007), in chapter five the condition specific patterns of AEI were also investigated. To achieve this, a novel computational pipeline was devised using the publically available diploid reference genome. RNA sequencing data published by Bruno *et al.* (2010) from the wild-type strain of *C. albicans* SC5314 grown under various conditions was then analysed, revealing that in general, AEI does not appear to alter in a condition-specific manner. Table 5.4 demonstrates this point clearly, showing that the majority of genes are identified in more than one growth condition, but with these conditions often being unrelated. For example, orf19.4212 was identified under growth in tissue culture media at pH 8 as well as growth in YPD under the no nitrosative stress control, but was not identified in tissue culture media at pH4, nor in other YPD control conditions. A second example shows that no genes were identified under growth in both oxidative stress conditions only, with all low oxidative stress genes also being identified under growth in YPD as a no oxidative stress control and other conditions. However a small number of genes do appear to have a condition specific response to AEI, orf19.4504 was identified in both tissue culture conditions only, orf19.4773 was identified under growth in nitrosative stress only, and orf19.6486, orf19.1763, orf19.2787, orf19.3365 and orf19.5145

were identified under high oxidative stress only. These few examples of genes do imply that there could be a very small set of condition specific responses that require AEI, but this needs further verification.

The appearance of a number of unexpected results has opened up the novel computational pipeline for criticism. A distinct lack of consistency is present when comparing the results from similar growth conditions. Growth under YPD alone was used as the control for comparison to growth under stresses by Bruno *et al.* (2010) and was therefore analysed on three separate occasions. On each occasion, different genes were identified as having AEI, implying that allele expression is varying even though the growth condition is unchanged. A number of explanations could account for this observation. Technical variability, whether as small differences in experimental protocol or during library preparation, could result in larger observations of change in expression levels. This has reportedly been the case for RNA sequencing data where coverage is low (McIntyre *et al.*, 2011), which is sometimes the case for the allele specific counts used here. Stochastic gene expression, where expression levels of genes in single cells are randomly distributed across a population, could also account for this variability. However, although stochastic expression levels have previously explained variability in single cell expression studies (Raj and van Oudenaarden, 2008), it has not been used to describe variability across entire cell populations. This result could also be a consequence of the computational pipeline. This is made evident when comparing the results from growth in YPD by Bruno *et al.* (2010), where just two genes are identified with AEI, to the re-analysis of the RNA sequencing data from chapter 3 also from growth in YPD, where 21 genes with AEI are identified. The pipeline devised here appears to be sensitive to both read length and replicate number, with the study by Bruno *et al.* (2010) using shorter 36 bp reads and just two biological replicates identifying a far smaller number of genes with AEI. The results of the co-infection study further compound this point with the identification of a much higher number of 123 genes from the use of longer 100 bp paired end reads and three biological replicates. To support these findings, an investigation into the most appropriate method for identification of differentially expressed genes has shown that increasing both replicate number and sequencing depth leads to an increase in statistical power and an increase in identification of differential

expression when using the existing software packages DESeq, Cuffdiff and edgeR (Zhang *et al.*, 2014).

The sensitivity to read length and replicate number is likely to be due in part to the high stringency of the statistical testing used by the computational pipeline. An increase in reads (and therefore allele counts) allows for more genes to pass all three statistical tests. This stringency can also explain why such a small number of genes with AEI are identified, as discussed in chapter 5. A number of statistical tests were chosen due to a lack of consistency in current investigations of AEI, with no apparent “best” statistical method. To overcome this, further validation of the computational methodology is needed. Due to the lack of success in developing a wet-lab method for validation of AEI, the most appropriate way forward could lie in the use of a simulated RNA sequencing data-set with known disparities in allelic expression imbalance which can be used to identify which statistical tests produce the most accurate results.

To further explain the point of inconsistencies in results from growth in similar experimental conditions, the RPKM normalisation technique used here has been previously highlighted as unsuitable for use when comparing RNA sequencing results across unrelated studies, as normalising for total read number calculates proportions of expression. If one gene is greatly increased in expression in a sample compared to others, the remaining genes will also appear to have reduced expression in this sample (Robinson and Oshlack, 2010). Therefore an alternative normalisation method, such as TMM as mentioned in section 5.4.1, may be more appropriate in this study. Additionally, RPKM has shown to be biased both by read length and GC content (Bullard *et al.*, 2010a, Zheng *et al.*, 2011). As the results of the analysis of structural factors in genes with AEI in chapter three show that both gene length and GC content have a significantly larger variance in genes with differentially expressed alleles, it is possible to suggest that any differences in expression are in fact due to the biases of the normalisation technique. Software which identifies differential expression through use of raw counts, such as DESeq (used here in chapter 5) and edgeR (Robinson *et al.*, 2010) may therefore be more appropriate for use with calculations of AEI. A recent comparison of software for identification of differentially expressed genes showed that edgeR out-performs DESeq and

Cuffdiff, a package based around RPKM normalisation, for detecting true positive results but may be subjected to a high number of false positive results (Zhang *et al.*, 2014).

Another observation of concern is the lack of consistency between the computational analyses of the same data-set carried out in chapter 3 and chapter 5, with just 14 genes identified both times. The differences in these results are discussed in chapter five, and are likely to be due to both the higher stringency of statistical testing as well as the removal of reads aligning to INDELs in the latter analysis. INDELs were removed in this analysis as reads aligning in these regions were leading to a distinct bias for identification of genes with AEI from alleles that differ in length. This bias due to differing allele length has not previously been reported in other studies of allelic expression imbalance, but is something that future investigations of AEI should take into consideration.

During the course of this investigation, evidence for genome-wide levels of allelic expression imbalance in *Candida albicans* was published by Muzzey *et al.* in 2013 and 2014. The initial investigation by Muzzey *et al.* (2013) focused around the development of a diploid reference genome with improved phasing information. The phased reference genome was then used for identification of allelic expression imbalance in combination with a similar computational pipeline to the one devised here. Unfortunately, as no statistical measures were applied to compare the expression of each allele, this study does not have a definitive list of genes with significant allelic expression imbalance against which our results can be compared. However, the methodology used by Muzzey *et al.* (2013) does indicate areas for improvement of the pipeline used here. Muzzey *et al.* (2013) included screening for the directionality of imbalance across SNPs of a single gene ensuring that all “SNP windows” favour expression of the same allele and included bootstrapping to produce a confidence interval in the fold change measurement to ensure that extreme counts at individual SNPs do not skew results. Implementing these steps into the pipeline constructed here may help to remove any false positive results. Muzzey *et al.* (2013) also highlighted that use of the improved phased reference genome is more sensitive for identification of AEI, using orf19.3556 as an example gene which is detected as

having AEI with the new reference but not with the old diploid reference genome published by Jones *et al.*, (2004). Differences in the sensitivity could be accounted for by the differences in SNP numbers between the two reference genomes. The new phased genome increases the SNP number from 54858 to 69688. Additionally, only 75% of the original SNPs could be corroborated by the new reference genome, with the statement that many locations were falsely identified as heterozygous in the original reference due to low coverage. This supports the findings found in this investigation where a number of genes identified with significant AEI were subsequently found to be homozygous through cloning and sequencing of the alleles. Use of the new phased reference genome alongside the pipeline constructed here will aid in improving the accuracy of the results obtained, as currently reads will be falsely aligned to SNPs that don't exist and will be missed from SNPs that are lacking from the original assembly. However, this reference genome is not yet publically available for use.

This study differs from previous investigations of allelic expression imbalance through the use of a diploid reference genome. Investigations in other species have taken the approach of using a haploid reference genome alongside variant calling. However, this method has been associated with a significant intrinsic bias, with reads tending to map to the reference allele as opposed to the "alternate" allele (Degner *et al.*, 2009, Stevenson *et al.*, 2013) and is also dependent upon accurate identification of SNPs. Use of a diploid reference genome here has aimed to overcome these issues. Similar approaches have been adopted by the software packages Allim (Pandey *et al.*, 2013), Alleleseq (Rozowsky *et al.*, 2011) and MMSEQ (Turro *et al.*, 2011), however these methods are often designed for use with human samples or with diploid offspring where full parental reference genomes are available. Use of the approach developed here, alongside aspects of the methodology from Muzzey *et al.* (2013), could allow for unbiased and accurate identification of allelic expression imbalance in any species with a diploid reference genome. Although that number of species is currently limited to humans, *C. albicans* and the giant panda, there are clear possibilities for use with manually constructed diploid references inferred from variant calling. Recent work into *Drosophila* show that although improvements can be made for use of a haploid reference genome,

using high quality SNP calling directly from the RNA sequencing data, small biases towards the reference genome still occurred (Quinn *et al.*, 2014). Therefore use of a diploid reference genome, or a similar approach, are likely to be the more reliable avenue for future investigations into allelic expression imbalance.

As well as issues surrounding statistical testing and false negative results, inherent bias caused by the sequencing process itself may also produce errors in identification of allelic expression imbalance, as discussed in section 3.4.2. Errors have been observed which are dependent upon the library preparation method, with cDNA fragmentation leading to a higher coverage of reads at the 3' end of transcripts (Wilhelm *et al.*, 2008) and RNA fragmentation methods causing depletion of coverage at either transcript end (Wang *et al.*, 2009). As cDNA fragmentation was used both in chapter 3 and by Bruno *et al.* (2010), this source of bias is a possible explanation for the results seen in chapter 5 where false positive results are obtained when mapping to both SNPs and INDELs. Alleles which have insertions in the 3' region are likely to have a much higher number of reads, causing a skew in the calculation of AEI towards that allele. This bias needs to be accounted for in future experiments. Removing reads mapping to INDELs, as has been done here, may lead to false negative results from genes which have only INDELs and no SNPs. Therefore, use of RNA fragmentation for the sequencing library preparation may be a more suitable option to overcome this issue.

GC rich regions have been shown to be more "sequenceable" by Illumina technologies than regions of low GC content (Bullard *et al.*, 2010b) with a 5% difference in GC content conferring as much as a 10% difference in reported expression level. As AEI identification is based around differences in allele sequences, GC content is likely to differ, even if only slightly. This observation was made in chapter three, where genes with AEI were seen to have a significantly larger difference in GC content than equally expressed genes. However, when this data is reanalysed, excluding all alleles which differ in length, i.e. which have INDELs, the difference in GC content is reduced from 0.80 to 0.27, a number far more comparable to the equally expressed alleles which have an average GC content difference of 0.24. Therefore GC content

bias may be less problematic with the novel computational pipeline which excludes INDELs. However, if a pipeline using INDEL information is developed in the future, GC content bias could be accounted for using a method which models expected expression change due to GC content and then corrects for this difference in AEI estimations, similar to that developed by Skelly *et al.* (2010) in a study of allele specific expression in hybrid diploid yeast.

Finally, systematic errors in the base calling during sequencing have also been shown to cause potential errors in identification of AEI, with certain base pair motifs frequently sequenced incorrectly (Meacham *et al.*, 2011). Currently, this is an unavoidable issue in computational analysis of AEI, but is something that all researchers in this field should be aware of.

6.1.2 Future Advancements in Identification of AEI

Advances in analysis and quantification of RNA are leading the way for possible future methods for improved identification of AEI. The most recent advance in sequencing technology is the advent of single-cell RNA sequencing methods, such as Smart-seq. Studies have suggested that although a gene may appear to have allelic expression imbalance at the level of the cell population, individual cells may deviate from this expression pattern (Levesque *et al.*, 2013), making single cell technologies important for identifying this difference. The Smart-seq system has been used to identify monoallelic gene expression in preimplantation embryos of mice, showing that allele selection appears to be random and dynamic (Deng *et al.*, 2014). However, using this methodology to identify levels of AEI is still problematic, as around 60% of all polyadenylated RNA species are lost in the preparation protocol.

Recently, a study has showed that modification of RNA FISH (fluorescence *in situ* hybridisation) can be used to detect single nucleotide differences and quantify allele expression at a single cell level. This paper developed oligonucleotide probes to enable them to be specific. In the past longer probes have been needed to have enough binding energy, but this then leads to mismatched binding over SNPs. In this study, a “toehold probe” was used. This has a 28 bp stretch of single stranded DNA including the SNP of interest, followed by a stretch of double stranded “mask” oligonucleotides. The single

stranded region is short enough to confer specificity, and then the mask region binds and increases the binding energy (Levesque *et al.*, 2013).

Improvements have also been seen in the statistical analysis of AEI using RNA sequencing data. A hierarchical Bayesian model has been shown, using a yeast hybrid model, to be more powerful and to produce results with more relevant biological inference than a standard binomial test assuming equal allele expression. (Skelly *et al.*, 2011). These results were demonstrated across two different sequencing platforms with high levels of reproducibility and specified false discovery rates. A distinct advantage of this method is that it accounts for variability in levels of AEI across SNPs within a single gene, which may be seen in cases of allele-specific splicing or alternative transcription start sites. A fundamental question raised in this study is that with enough sequencing depth and precision it may be possible to show that every single gene has some level of AEI, therefore what level of AEI is biologically significant?

As all methods investigating allelic expression imbalance are reliant upon known locations of SNPs, the quality of the reference genome used is the biggest challenge faced by researchers in this field. Improved variant calling to increase the reliability of SNPs is essential. Quinn *et al.* (2014) show that this is possible through use of SNP calling directly from the RNA sequencing data, but further improvements in this area are still needed. Better phasing of sequence information, as demonstrated by Muzzey *et al.* (2013), will also aid in improving the reliability of results particularly where it can be shown that all SNPs of a gene favour the expression of the same allele. With the recent advancements in long read sequencing from single molecule technologies, such as the PacBio and MinION, better phasing information for reference genomes should be achievable in the near future.

6.2 The Functional Impact of AEI and the Lack Thereof

6.2.1 Is AEI linked to Differences in Allele Function

A key hypothesis which was investigated in this piece of work was the theory that allelic expression imbalance is linked to differences in allele function. This hypothesis is supported by previous studies in genes in *C. albicans,* such as *MDR1* (Bruzual and Kumamoto, 2011) and *CHS7* (Sanz *et al.*, 2007), that link

divergent allele expression to function as discussed in section 1.8. Investigations of percentage protein identity in chapter three also further support this idea, showing that genes with AEI had a significantly lower percentage protein identity (Figure 3.3), suggesting significant divergence in sequence. However, genes with significant sequence differences are also more susceptible to the sequencing biases discussed in section 6.1.1. Therefore the differences in protein identity observed could actually be a consequence of an over-representation of genes with sequence differences during the identification of allelic expression imbalance.

Identification of over-represented or under-represented Gene Ontology terms in the gene sets generally suggests that there is no common functionality between genes with AEI. Interestingly, GO analysis of the significant genes identified from different RNA data-sets identified different significant functions and processes. In chapter three, no functions, processes or components were found to be over represented. However, re-analysis of the RNA data-set in chapter five revealed an over representation of genes involved in oxidation-reduction processes and activities, iron transport, and locations on the cell surface and plasma membrane (Table 5.6). Table 5.8 shows that differences in growth condition do not impact upon Gene Ontology patterns, with genes involved in metal ion binding and transport processes, oxidation and reduction processes and plasma membrane localisation identified across a number of conditions. Conversely, analysis of the genes identified in the co-culture RNA data-set show over-representation of functions including protein kinase, transferase and transaminase activities (Table 5.9). Overall this suggests that genes with AEI are not involved in a single biological process and functional consequences of differences in allele expression need to be investigated on a gene-by-gene basis. The lack of consistency between results from different data-sets again highlights the issues in the methodology as discussed in section 6.1.1.

The functional consequences of AEI in individual genes were therefore investigated through use of phenotypic screening of heterozygous knockout mutants. Genes were selected from both computational analyses performed (chapter three and chapter five). Although minor differences were seen for a number of assays in certain strains of some genes (summarised in Tables 4.8

and 5.13), taken together these results strongly suggest that allelic expression imbalance is not linked to functional differences in the alleles. Yet functional differences of alleles have previously been linked to allele expression in *Candida albicans* for a number of genes. The chitin synthesis gene *CHS7* has homozygous alleles with polymorphisms in the promoter regions. A heterozygous knockout strain containing just the allele with the shorter promoter had similar characteristics to the wild-type whereas a knockout strain with only the allele with the long promoter suffered from reduced chitin synthesis and moderate morphological differences during hyphal growth (Sanz *et al.*, 2007). A similar story is seen with the efflux transporter gene *MDR1,* where polymorphisms in upstream regions are linked to differences in allele expression. This has important clinical ramifications, since homozygosity of the allele with higher expression is linked to levels of antifungal resistance in fluconazole resistant strains (Bruzual and Kumamoto, 2011).

Although a wide range of both general and gene specific phenotypic assays were used here, it is possible that the conditions which indicate functional differences between alleles have been missed. For practical reasons, a number of gene specific assays were not carried out here which may have revealed disparities in allele function. For example, the *VPS1* gene has been shown to be involved in sorting of the vacuolar protein carboxypeptidase Y in *Saccharomyces cerevisiae* (Peters *et al.*, 2004, Bernardo *et al.*, 2008). A lack of availability of an antibody against the *C. albicans* carboxypeptidase Y protein means that the assay indicating defective sorting cannot be used with the heterozygous knockout mutants constructed here. Use of high throughput technologies, such as robotics which plate arrays of yeast strains, could also be used to intensively screen the heterozygous knockout strains on a large number of *in vitro* conditions over a short period in time, such as was used for the screening of the transcription factor library by Homann *et al.* (2009). A further avenue for investigation is use of *in vivo* assays such as growth in the presence of macrophages or neutrophils, or virulence assays using a murine model. However, regardless of the number of assays used, a lack of evidence supporting differences in allele function does not rule out the possibility entirely.

With a lack of a validation technique for allelic expression imbalance, such as allele-specific qPCR, or an unlimited amount of RNA sequencing data, it is currently not possible to determine to what extent allele expression levels are changing within the heterozygous knockout mutants. It is possible that when one allele is removed, the other alters expression to compensate for the loss or other genes that show functional redundancy compensate. Functional redundancy of genes has been seen in homozygous knockout strains such as for the adhesion genes *ALS2* and *ALS4,* where removal of *ALS2* showed increased expression of *ALS4* and vice versa through RT-PCR (Zhao *et al.*, 2005). Therefore it can be suggested that a compensatory increase in expression could also occur for alleles. If this is the case for the heterozygous knockout strains, this could be an explanation for the lack of evidence for functional divergence of alleles.

The results of large scale screening of heterozygous knockout mutants suggest that the product of a single allele can be functionally important, and removal of an allele is not always compensated for. Screening on 35 different inhibitory compounds by Xu *et al.* (2007) demonstrated that haploinsufficiency readily occurs and the lack of an allele can lead to increased sensitivity to a number of chemicals. However, it cannot be demonstrated from studies such as this whether the alleles are functionally distinct or if phenotypes of heterozygous knockouts are due to gene dosage effects. Surprisingly, when comparing the genes recorded as having heterozygous phenotypes on the *Candida* genome database (www.candidagenome.org) to the genes identified with AEI in chapter three, 31 were identified on both lists (Table 4.1), but a further 932 genes with heterozygous phenotypes did not have AEI. 202 genes with AEI were also not identified as having haploinsufficient phenotypes. The study by Hickman *et al.* (2013) presenting haploid strains of *C. albicans* also highlights the functional significance of single alleles. Haploid strains and homozygous diploids were observed to have reduced fitness suggesting that the presence of both alleles is important.

Although it cannot be concluded with certainty, the results of the study presented are surprisingly indicative of little linkage between allele function and expression level in *C. albicans*. This is also reflected in previous studies of AEI

in other species, which have readily identified the genome-wide extent of allele expression imbalance but have not made any inferences regarding the functional consequences. Imprinting, however, is an important exception to this with the functional impact of single allele expression, and often the loss of imprinting, commonly reported upon in higher eukaryotic species. For example loss of imprinting at the *H19/Igf2* locus resulting in biallelic expression has been seen in many cancer types (Feinberg, 1993), and a complete switch in monoallelic expression at this locus has been observed in oral squamous cell carcinoma (OSCC) (Tuch *et al.*, 2010a). In this investigation, a number of genes were identified as having monoallelic expression from the analysis of all data-sets. Two of these genes, *RPS7A* and orf19.5648, were shown to be homozygous in section 5.3.8, but heterozygous knockout mutants of one monoallelic gene, *SMI1,* were successfully constructed in chapter three showing that not all monoallelic genes are a consequence of errors in the reference genome. For future experiments, it may be worthwhile to make these monoallelic genes the focus of functional investigations in *Candida albicans* to identify processes similar to imprinting.

Lack of evidence for functional differences between alleles with expression imbalance could also be due to a lack of expression differences at the protein level. An experiment using single molecule fluorescence microscopy and fluorescence *in situ* hybridisation found that protein and mRNA levels did not significantly correlate for 137 genes in *E. coli* (Taniguchi *et al.*, 2010), and importantly ribosome profiling in *C. albicans* has also demonstrated that levels of AEI are not always present at the translational level (Muzzey *et al.*, 2014). Although attempts were made to validate the allele specific protein expression levels in chapter three, through the use of western blotting, the techniques were insufficiently quantitative to give a clear measure of imbalance. In future, developing a quantitative technique to measure absolute allele specific protein abundance would be advantageous to this work. A possible system is QconCAT technology, which uses internal standards formed from concatamers of tryptic peptides as controls during mass spectrometry analysis (Pratt *et al.*, 2006). However, this method is yet to be used to quantify protein products of individual alleles.

6.2.1 Why Did AEI Arise?

If allelic expression imbalance is not due to functional divergence between alleles, then why might it occur? One possible explanation relates to gene dosage effects. Tuch *et al.* (2010a) demonstrated in OSCC tissue that increasing allele copy numbers leads to an increase in allelic expression imbalance. It is possible to hypothesize that genes with AEI, especially monoallelic genes, may have finely tuned expression levels to prevent detrimental amounts of protein being present in the cell. Therefore, construction and phenotypic screening of over-expression strains in *C. albicans* may reveal a link between gene dosage effects and allelic expression imbalance. Attempts were begun in this area, with the aim to replace the upstream region of the monoallelic gene *RPS7A* (see section 5.3.6) with the upstream region of the highly expressed *ENO1* gene, however *RPS7A* was subsequently shown to in fact be a homozygous gene with polymorphisms being due to errors in the reference genome (see section 5.3.8). A further idea linked to gene dosage is that gene gain is a form of quantitative neofunctionalisation; where gain in gene dosage has no qualitative new function, but presence of an extra copy confers a fitness advantage (Scannell *et al.*, 2007). However if this was the case, phenotypic defects should have been observed in the heterozygous knockout strains due to a reduction in expression of one allele.

It is also possible that AEI in *Candida albicans* has no functional consequence and has arisen, as of yet, through unknown mechanisms. One possible mechanism is that during situations where selective pressures are reduced, loss-of-function mutations can occur and accumulate in the population via genetic drift. This is known as the “local neutrality hypothesis”. A heterozygous diploid individual therefore masks these loss-of-function alleles by retaining the functional copy (Zörgö *et al.*, 2012). However, functional differences between heterozygous knockout mutants would still be expected in this case. Then again, if areas of aneuploidy have arisen in the knockout mutants, extra allele copies could also be masking any phenotypes. Acquisition of genome changes have been observed in laboratory strains that have undergone molecular manipulations (Abbey *et al.*, 2011), and therefore using an assay such a comparative genome hybridisation to verify the copy number of chromosomes in the heterozygous knockout strains would aid in ruling out this scenario. A

final possible explanation for AEI returns to the idea of stochastic gene expression, as mentioned in section 6.1.1, where allele expression levels in cells are variable due to natural levels of “noise”.

6.2.2 The Medical Importance of AEI

Although functional differences of alleles have not been linked to AEI here in *Candida albicans*, examples of the impact of AEI upon human health have been reported previously, reminding us of the importance of gaining a better understanding of this phenomenon. As discussed in section 1.7, allelic expression imbalance has been implicated as a causative factor in a number of human cancers. AEI in the *BRCA1* gene have been shown to be increased in familial breast cancer patients (Chen *et al.*, 2008), the *H19/Igf2* locus has been associated with various types of cancer (Feinberg, 1993, Tuch *et al.*, 2010a), and colorectal cancer-specific AEI has been shown to occur in B cells (Lee *et al.*, 2013). Imprinting associated congenital disorders in humans such as Prader-Willi, Angelman, Beckwith-Wiedemann and Silver-Russell syndromes also occur.

Exploiting AEI for use in treatment of human disease is also being considered. Allele-specific gene silencing by RNAi has been investigated for use against the hepatitis B virus. The technique claims to have the potential to suppress the “disease-causing” allele whilst leaving the wild-type allele expression intact (Teng *et al.*, 2011). This work is still in the rudimentary stages of development, and in fact found that the allele-specific siRNA targeted the wrong allele through unknown mechanisms, but demonstrates the potential of allele specific treatments.

6.2.3 Future Avenues to Investigate the Functional Impact of AEI

If genes with AEI do have functional differences between the alleles, each allele will be under distinct evolutionary pressures. Generally, evolutionary pressures on proteins are quantified by measuring the ratio of substitution rates at non-synonymous and synonymous sites (dN/dS). This is calculated using divergent sequences where sequences under adaptive evolution produce dN/dS > 1 (Nielsen and Yang, 2003). However it has be found that dN/dS does not follow the same relationship in sequences sampled from a single population

(Kryazhimskiy and Plotkin, 2008) making it inappropriate, in our case, to measure the evolutionary pressures on separate alleles within a species. However, if an alternative measure of evolutionary pressure could be devised, it may be possible to infer if alleles are functional distinct due to the presence or lack of evolutionary pressures.

This study focused upon a single isolate of *Candida albicans,* the wild-type strain SC5314. Clinical isolates have been shown to differ in allele sequences and levels of heterozygosity. For example, across 60 isolates, 11 different alleles were found for the elongation factor 3 gene *EF3.* These alleles were found to be present in a total of 16 different combinations (Bretagne *et al.*, 1997). A clinical investigation of 204 isolates of *Candida albicans* obtained from HIV patients, non HIV patients and healthy individuals using restriction fragment length polymorphisms revealed 66 different genotypes, with samples from healthy individuals showing the highest level of heterogeneity (Xu *et al.*, 1999). Therefore it is possible to hypothesize that the genome-wide extent of allelic expression imbalance may differ between isolates. Ideally, repeating RNA sequencing on a large number of clinical isolates with differing characteristics, such as virulence or antifungal resistance, could link AEI to functionality, either showing that the genome-wide extent, or AEI in single genes, is commonly associated with certain traits. Additionally, a study such as this could also indicate if certain combinations of alleles are favoured at the population level or linked to differing phenotypes. Practically, if a validation technique for allelic expression imbalance, such as qPCR, can be achieved, using this on a subset of genes within clinical isolates may also reveal useful information.

6.3 What are the Control Mechanisms of AEI in *C. albicans?*

A key question that was touched upon in this study is what are the control mechanisms behind AEI? Chapter three details an analysis of the correlations between structural factors and allelic expression imbalance. Previous investigations have shown that factors such as GC content and length can impact upon gene expression levels (Coghlan and Wolfe, 2000, Goncalves *et al.*, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003, Urrutia and Hurst, 2003, Versteeg *et al.*, 2003). Here, gene length was the only factor identified which differed significantly between alleles with differential expression.

However, it was found that the longer allele had significantly higher expression levels (Figure 3.7). This observation goes against previous findings in *S. cerevisiae* which have suggested that shorter alleles have higher expression levels as they are more transcriptionally efficient (Coghlan and Wolfe, 2000, Jansen and Gerstein, 2000, Marín *et al.*, 2003). During the development of the computational pipeline in chapter five, the same observation was made with longer alleles showing higher expression levels. However, the majority of genes with AEI were seen to differ in length, and therefore this observation was deemed a bias in the computational pipeline likely to be due to uneven coverage of reads at the 3' end of transcripts (Wilhelm *et al.*, 2008).

A further observation made in chapter three is that genes with allelic expression imbalance have a significantly larger variance in structural factors than genes with equal allele expression. As opposed to these structural factors influencing expression levels, these variations indicate that sequencing bias, as discussed in section 6.1.1., could be impacting upon the process of identifying genes with AEI, and that these differences in structural factors should be accounted for.

Any study investigating the control of expression levels will need to consider variability in *cis*- factors, such as promoter regions, as a possible causative factor and their interaction with *trans*-acting factors such as transcription factors. In terms of AEI, a significant correlation has previously been observed between polymorphism density and levels of AEI on a genome-wide scale (Gagneur *et al.*, 2009). Single gene studies in *C. albicans* have also shown that variability in promoter regions are linked to allele expression differences in the genes *CHS7* (Sanz *et al.*, 2007) and *MDR1* (Bruzual and Kumamoto, 2011). Despite this, analysis of the upstream regions of genes with AEI, both in chapter three and chapter five, showed that only 75% of genes with AEI have polymorphisms in this region. As not all genes with AEI have upstream polymorphisms, *cis*-factors can be ruled out as the sole driving force behind differential allele expression. Additionally as a similar rate of polymorphism was also observed in the upstream regions of genes with equal allele expression and with calculations based upon the average rate of heterozygosity, these findings suggest that polymorphisms are actually selected against in upstream regions of all genes. To better quantify the exact influence of *cis*- factors, use of the

knockout library of transcription factors (Homann *et al.*, 2009), alongside RNA sequencing or allele-specific qPCR, could identify where *cis*- factors are influencing AEI and identify which transcription factors are linked to this process, possibly leading to elucidation of the signalling pathways driving these expression levels.

As discussed in section 1.6.1. methylation has been shown to be a causative factor in allelic expression imbalance, especially in imprinted genes such as *H19/Igf2* (Hou and Corces, 2011) and *Mest* (MacIsaac *et al.*, 2011). In *Candida albicans*, genome-wide analysis revealed that 150 genes have methylation patterns (Mishra *et al.*, 2011). Comparison of these genes to those identified with allelic expression imbalance in chapter three revealed that only six genes with AEI are methylated, suggesting that methylation is also not the driving force behind differences in allele expression. However, the study by Mishra *et al.* (2011) only identified methylation in terms of the haploid genome and did not identify allele-specific methylation patterns. Use of genome-wide bisulfite sequencing, such as that used by Lister *et al.* (2008) in *Arabidopsis thaliana,* alongside similar methodology to that used here could identify genome-wide allele-specific methylation patterns. Techniques such as this have previously been seen to be problematic as methylcytosines occur at a higher frequency than SNPs, leading to difficulties in assigning methylation to alleles. Nevertheless, with the advent of long read sequencing technologies, such as the PacBio and MinIon, efforts can be made to overcome these issues.

Sections 1.6.2 and 3.4.3.4 also detail alternative control mechanisms that could be investigated in future investigations into the causative factors behind allelic expression imbalance including chromosomal interactions, chromatin structure, untranslated regions and the influence of asynchronous replication. These factors have previously been linked to gene expression or allelic expression imbalance but are yet to be investigated in allelic expression imbalance in *Candida albicans.*

The focus in this study has been upon possible control mechanisms at the transcriptional stage, however levels of AEI may also be impacted upon by post-transcriptional mechanisms such as differing rates in RNA decay. As

discussed in section 3.4.3.4, RNA decay rates have been shown to be influenced by sequence motifs and structural factors such as AU rich elements (Vasudevan and Peltz, 2001) and gene length (Santiago *et al.*, 1986). Therefore the presence of polymorphisms in alleles could leads to differences in poly(A)-shortening rates or RNA secondary structure, in turn leading to differences in decay rates. Development of existing methods which measure RNA decay, such as use of 1,10 phenanthroline, to enable identification of allele specific rates is needed to identify the contribution of post-transcriptional factors upon levels of AEI.

6.4 Concluding Remarks

To summarise this body of work, RNA sequencing data has been used with a number of computational pipelines to identify the genome-wide extent of allelic expression imbalance in *Candida albicans.* The functional consequences of AEI were then investigated through phenotypic screening of heterozygous knockout mutants. Attempts at developing a validation system to support the results from the RNA sequencing were made and investigations into the possible control mechanisms driving the differences in expression levels were touched upon.

Results suggest that AEI is widespread in *Candida albicans,* with no clear link to a single biological process and no clear response to changes in growth conditions. However, errors due to false negative results and biases in the sequencing process should be considered during any interpretation of results. Phenotypic screening of heterozygous knockout mutants suggests that there is no link between allelic expression imbalance and divergent allele function in *Candida albicans;* nonetheless the possibility cannot be ruled out. Further high throughput phenotypic screening and the development of a validation technique which can quantify allele redundancy will help to confirm or reject this hypothesis. The control mechanisms behind allelic expression imbalance are still unclear, and further investigations are still needed in this area.

Together, this piece of work presents a novel investigation into the functional consequences of allelic expression imbalance in *Candida albicans.* Although no strong conclusions can be made regarding the functional importance, advances

into the understanding of gene expression mechanisms in this medically important pathogen have been made.

Appendix I

Table Ia. Expression data for genes with allelic expression imbalance

Average RPKM values represent the average normalised allele expression levels of three replicates. The fold difference in RPKM values for allele 1 and 2 is calculated. P values were calculated using Fisher's Exact Test. Colours show relativity of values with green representing the lowest value and red representing the highest value.

CA name[1]	SC ortholog[2]	Allele 1	Allele 2	Allele 1 Average RPKM	Allele 2 Average RPKM	Fold Diff	P-value	Total Tags (Haploid)
	TFP1	orf19.1680	orf19.9249	0.32	397.96	1258.03	0	5280
		orf19.3224	orf19.10734	0.40	890.88	2206.98	0	137
RPS7A	*RPS7A*	orf19.1700	orf19.9267	4.16	1241.96	298.33	0	14540
VMA11	*TFP3*	orf19.6538	orf19.13891	547.06	7.89	69.30	0	4850
RPL24A	*RPL24A*	orf19.3789	orf19.11269	693.76	4665.36	6.72	0	975
		orf19.5648	orf19.13093	740.85	25.95	28.55	0	5926
CHT2	*CTS1*	orf19.3895	orf19.11376	5204.49	1403.44	3.71	0	469
ATP1	*ATP1*	orf19.6854	orf19.14144	5231.12	0.69	7574.02	0	3040
MPP1	*MPP1*	orf19.1915	orf19.9471	0.39	296.35	752.15	1.3E-272	915
		orf19.5062	orf19.12528	0.56	185.96	331.87	2.2E-168	1337
		orf19.5128	orf19.12593	168.34	0.65	260.05	1.5E-144	1392
ARO8	*ARO8*	orf19.2098	orf19.9645	159.47	0.15	1097.28	3.8E-141	2050
POR1	*POR1*	orf19.1042	orf19.8644	648.59	225.68	2.87	2.5E-132	595
BAT21	*BAT2*	orf19.797	orf19.8416	389.36	85.89	4.53	4.4E-132	15265
	SCS2	orf19.1212	orf19.8800	145.77	0.80	182.66	1.5E-124	2684
VID21	*EAF1*	orf19.3077	orf19.10589	230.53	24.90	9.26	1.1E-121	641
	CYC8	orf19.4959	orf19.12424	143.02	3.74	38.24	1.8E-108	5924
IFF9	*MUC1*	orf19.465	orf19.8096	120.59	0.73	164.59	7.8E-103	1589
		orf19.1537	orf19.9111	24.24	183.54	7.57	1.21E-95	1976
MSN5	*MSN5*	orf19.2665	orf19.10182	0.17	100.81	586.10	9.48E-92	3500
ECM1	*ECM1*	orf19.5299	orf19.12758	252.02	62.31	4.04	1.24E-77	1971
PDX3	*PDX3*	orf19.550	orf19.8185	131.19	11.04	11.89	4.84E-77	4813
SSK2	*SSK2*	orf19.3775	orf19.11257	14.69	132.21	9.00	4.64E-75	1220
FAV3		orf19.1914	orf19.9470	0.12	78.90	648.51	7.16E-74	3436
FCY21	*FCY2*	orf19.1357	orf19.8937	28.74	160.07	5.57	2.01E-70	3670
	DMA1	orf19.1185	orf19.8776	162.26	27.08	5.99	7.27E-69	317
RCK2	*RCK2*	orf19.2268	orf19.9808	14.09	121.55	8.63	4.13E-68	33930
		orf19.233	orf19.7863	366.79	142.58	2.57	1.38E-63	182
CAR1	*CAR1*	orf19.3934	orf19.11416	401.62	166.11	2.42	1.22E-62	3877
RPS23A	*RPS23B| RPS23A*	orf19.6253	orf19.13632	75.16	2.89	26.05	9.28E-54	135

	NCL1	orf19.518	orf19.8149	63.26	0.54	116.29	7.81E-53	4983
ACH1	*ACH1*	orf19.3171	orf19.10681	136.91	26.33	5.20	9.72E-53	147
		orf19.3191	orf19.10702	28.52	129.70	4.55	1.97E-49	6211
BCY1	*BCY1*	orf19.2014	orf19.9565	79.65	218.15	2.74	1.16E-48	335
	PET127	orf19.2309	orf19.9845	7.07	77.94	11.02	1.42E-48	1697
		orf19.1266	orf19.8852	17.41	102.03	5.86	6.75E-47	735
	MED7	orf19.232	orf19.7862	81.95	8.10	10.12	4.55E-46	676
	MRPL11	orf19.3797	orf19.11278	116.77	268.87	2.30	1.13E-45	1264
	SDH4	orf19.4468	orf19.11949	162.97	333.89	2.05	5.08E-45	6937
MSS4	*MSS4*	orf19.3153	orf19.10663	14.50	89.93	6.20	2.62E-43	3232
TIF4631	*TIF4631*	orf19.3599	orf19.11082	226.99	87.54	2.59	2.16E-40	7226
	BBC1	orf19.2791	orf19.10309	93.51	17.16	5.45	3.75E-38	3179
DAO2		orf19.3365	orf19.10873	30.27	114.17	3.77	1.54E-37	78
PGA57	*HPF1*	orf19.4689	orf19.12158	3.29	53.36	16.20	3.59E-37	667
	NUP84	orf19.1298	orf19.8878	58.84	4.48	13.13	7.68E-37	2346
IDP2	*IDP2*	orf19.3733	orf19.11217	108.49	25.00	4.34	1.48E-36	1013
	SIP5	orf19.2458	orf19.9994	53.38	3.40	15.69	4.16E-35	20060
	YGR12W	orf19.5574	orf19.13020	57.87	4.87	11.88	1.87E-34	1179
RBT4	*PRY1*	orf19.6202	orf19.13583	3.71	50.96	13.73	2.6E-34	2261
	PAT1	orf19.3792	orf19.11271	278.52	130.90	2.13	4.38E-34	247
	YLL7C	orf19.5147	orf19.12613	13.32	72.02	5.41	7.36E-32	3541
POL5	*POL5*	orf19.5597	orf19.13042	6.78	55.67	8.21	2.15E-31	109
ALS1	*FLO9*	orf19.5741	orf19.13163	37.95	0.95	39.81	1.75E-29	180
VPS1	*VPS1*	orf19.1949	orf19.9505	147.92	55.06	2.69	1.39E-28	9898
		orf19.3644	orf19.11128	40.23	1.91	21.02	3.02E-28	35
	RPH1	orf19.2743	orf19.10257	25.29	89.67	3.55	4.52E-28	39699
	YLR137W	orf19.1557	orf19.9130	49.63	5.03	9.88	5.69E-28	14012
	YPR147C	orf19.4398	orf19.11876	18.86	76.49	4.06	4.34E-27	47
		orf19.1953	orf19.9508	30.66	94.26	3.07	3.07E-25	816
	WSC4	orf19.254	orf19.7886	88.64	179.98	2.03	1.55E-24	323
		orf19.3353	orf19.10861	65.90	147.65	2.24	1.7E-24	295
ITR1	*ITR2*	orf19.3526	orf19.11009	41.62	3.99	10.43	2.88E-24	3304
ILV6	*ILV6*	orf19.4650	orf19.12119	37.62	102.13	2.72	3.34E-23	265
		orf19.1383	orf19.8963	3.95	37.57	9.51	1.06E-22	3404
	ESF2	orf19.3161	orf19.10670	150.19	65.50	2.29	2.62E-22	18391
	PKH3	orf19.1196	orf19.8787	1.46	28.88	19.83	4.81E-22	44
SNF4	*SNF4*	orf19.5768	orf19.13191	63.02	136.36	2.16	2.33E-21	16777
	PNP1	orf19.317	orf19.7949	48.67	114.63	2.35	5.04E-21	621
		orf19.803	orf19.8421	98.26	34.78	2.83	6.05E-21	908
		orf19.4770	orf19.12233	3.53	33.95	9.62	1.59E-20	1226
	RLM1	orf19.5626	orf19.13071	1.65	27.82	16.86	6.53E-20	70
	DEF1	orf19.3773	orf19.11255	31.36	85.22	2.72	8.25E-20	6433
	RPO31	orf19.3103	orf19.10615	23.35	0.47	49.75	1.08E-19	104
	SVL3	orf19.1948	orf19.9504	52.77	11.25	4.69	1.39E-19	882
	OSH2	orf19.5095	orf19.12561	21.46	0.04	491.34	1.79E-19	1662
	YAP181	orf19.4184	orf19.11660	54.33	12.50	4.35	6.33E-19	227

		orf19.6894	orf19.14182	66.60	134.77	2.02	1.87E-18	10274
		orf19.1938	orf19.9493	0.26	20.77	81.23	2.72E-18	913
		orf19.4952	orf19.12417	31.11	3.79	8.21	7.42E-17	5075
WSC1	*SLG1*	orf19.5867	orf19.13289	3.58	28.97	8.09	1.28E-16	828
	BFA1	orf19.6080	orf19.13499	17.85	0.13	133.51	1.61E-16	1154
	CYK3	orf19.6242	orf19.13620	29.33	75.27	2.57	2.25E-16	2956
		orf19.2742	orf19.10256	35.50	84.51	2.38	2.51E-16	3105
TAC1	*HAL9*	orf19.3188	orf19.10698	9.60	41.66	4.34	2.56E-16	946
	NSG2	orf19.273	orf19.7905	6.70	35.30	5.27	4.52E-16	958
TBF1	*TBF1*	orf19.801	orf19.8420	45.73	97.22	2.13	2.53E-15	2354
OCA1	*OCA1*	orf19.1762	orf19.9331	25.31	2.31	10.94	2.94E-15	343
		orf19.4068	orf19.11551	20.48	59.13	2.89	3.04E-15	3979
EMC9	*NNF2*	orf19.1907	orf19.9463	17.85	0.41	43.83	4.45E-15	425
	EDC3	orf19.6858	orf19.14148	22.65	1.71	13.24	5.57E-15	2408
		orf19.2381	orf19.9917	32.18	5.37	6.00	6.89E-15	728
PWP1	*PWP1*	orf19.4640	orf19.12110	0.88	18.79	21.39	3.15E-14	885
		orf19.5624	orf19.13069	0.76	17.26	22.81	4.2E-14	7499
TPO3	*TPO2*	orf19.4737	orf19.12199	6.00	30.43	5.07	1.29E-13	62
	MRD1	orf19.1646	orf19.9215	1.35	18.61	13.80	2.62E-13	10859
	MPD1	orf19.3920	orf19.11402	37.51	8.87	4.23	4.2E-13	1026
	BUD17	orf19.3411	orf19.10914	50.53	15.95	3.17	4.75E-13	41
PIR1	*PIR1*	orf19.220	orf19.7851	25.18	62.30	2.47	5.84E-13	706
PGA45	*PLB2*	orf19.2451	orf19.9987	103.68	50.33	2.06	9.56E-13	7345
	CYK3	orf19.6240	orf19.13620	34.67	75.27	2.17	1.45E-12	1929
UGA1	*UGA1*	orf19.802	orf19.8421	9.07	34.78	3.83	1.86E-12	72
	PMU1	orf19.5103	orf19.12569	3.53	23.37	6.62	2.46E-12	173
	NBA1	orf19.4349	orf19.11826	66.45	26.28	2.53	2.58E-12	386
	MET7	orf19.4516	orf19.11991	21.02	2.48	8.49	3.38E-12	1953
	MUC1	orf19.2051	orf19.9599	15.84	46.48	2.93	3.55E-12	1908
	RKR1	orf19.1219	orf19.8806	0.45	13.52	30.20	5.34E-12	51450
APE3	*APE3*	orf19.3591	orf19.11073	37.45	9.99	3.75	9.12E-12	221
	PSD2	orf19.3954	orf19.11436	31.06	6.97	4.46	1.24E-11	947
ARP4	*ARP4*	orf19.5623	orf19.13069	1.60	17.26	10.78	2.09E-11	859
ECM25	*ECM25*	orf19.4958	orf19.12423	16.87	1.46	11.59	2.73E-11	2372
	HOS4	orf19.4728	orf19.12191	15.51	43.90	2.83	3.11E-11	3006
	NAF1	orf19.494	orf19.8124	25.22	58.26	2.31	4.5E-11	530
PUS4	*PUS4*	orf19.1954	orf19.9509	82.70	38.92	2.12	5.5E-11	1004
		orf19.3372	orf19.10880	15.97	43.77	2.74	9.07E-11	3178
	YOR246C	orf19.3352	orf19.10860	5.19	24.36	4.70	1.75E-10	1588
	YNL168C	orf19.2184	orf19.9730	20.60	50.13	2.43	2.2E-10	1498
	COX1	orf19.3167	orf19.10676	7.26	28.02	3.86	2.66E-10	4012
	AST2	orf19.3706	orf19.11190	46.99	17.50	2.69	4.85E-10	10803
HGT1	*HXT11*	orf19.4527	orf19.12002	36.98	11.58	3.19	6.57E-10	933
OYE23	*OYE2*	orf19.3433	orf19.10937	23.90	4.97	4.81	9.35E-10	4418
	AGA1	orf19.6556	orf19.13909	17.79	2.21	8.05	1.01E-09	394
	MOT3	orf19.2724	orf19.10239	75.94	36.92	2.06	1.19E-09	1288

VMA7	*VMA7*	orf19.806	orf19.8424	15.20	40.11	2.64	1.89E-09	1198
RFG1	*ROX1*	orf19.2823	orf19.10341	21.86	49.53	2.27	2.78E-09	1393
		orf19.6235	orf19.13615	76.19	37.98	2.01	2.97E-09	1212
		orf19.6351	orf19.13708	14.42	1.33	10.80	3.41E-09	6181
		orf19.2521	orf19.10057	0.93	12.48	13.45	6.76E-09	389
		orf19.1736	orf19.9304	29.98	9.05	3.31	8.43E-09	3666
	AVO1	orf19.5221	orf19.12688	2.44	15.72	6.43	9.02E-09	4225
	HOT13	orf19.6555	orf19.13908	1.68	13.26	7.88	3.01E-08	286
ATS1	*ATS1*	orf19.6399	orf19.13757	19.25	43.23	2.25	3.16E-08	771
		orf19.3448	orf19.10952	25.33	7.04	3.60	3.39E-08	4064
	YIL18W	orf19.246	orf19.7876	5.45	20.05	3.68	1.15E-07	195
		orf19.5843	orf19.13265	14.17	2.02	7.01	1.2E-07	617
CDC6	*CDC6*	orf19.5242	orf19.12707	9.45	26.73	2.83	1.92E-07	8162
	ESBP6	orf19.4337	orf19.11813	8.25	23.57	2.86	9.97E-07	2303
	SWI3	orf19.4488	orf19.11964	7.06	21.39	3.03	1.1E-06	2327
		orf19.3524	orf19.11006	20.03	40.83	2.04	1.19E-06	1028
	YMR86W	orf19.1246	orf19.8830	15.27	34.08	2.23	1.4E-06	523
HIT1	*HIT1*	orf19.2723	orf19.10238	45.50	21.71	2.10	1.85E-06	396
IFD6	*YPL88W*	orf19.1048	orf19.8650	16.48	4.02	4.10	2.37E-06	30548
	YOR251C	orf19.1356	orf19.8936	24.51	8.83	2.78	2.93E-06	517
		orf19.195	orf19.7825	7.77	0.44	17.61	3.18E-06	2128
	ERF2	orf19.4466	orf19.11946	6.36	19.27	3.03	3.37E-06	168
MTR1	*MTR1*	orf19.1119	orf19.8716	19.94	5.98	3.33	3.6E-06	4803
	ECM18	orf19.310	orf19.7943	16.77	4.46	3.76	3.63E-06	586
ERB1	*ERB1*	orf19.1047	orf19.8649	6.18	0.15	41.74	4.01E-06	487
	SEC59	orf19.261	orf19.7893	41.03	19.43	2.11	4.69E-06	2359
	IFH1	orf19.4282	orf19.11758	18.02	36.67	2.04	4.74E-06	450
	SPS19	orf19.3684	orf19.11168	7.38	20.72	2.81	5.27E-06	344
SSP96	*FMO1*	orf19.5145	orf19.12610	5.54	17.50	3.16	6.69E-06	146
	RPM2	orf19.48	orf19.7710	2.80	12.19	4.36	7.26E-06	7957

1. *Candida albicans* name according to the *Candida* genome database (www.candidagenome.org).
2. *Saccharomyces cerevisiae* ortholog as given by the *Candida* genome database (www.candidagenome.org).

Table Ib. Expression data for genes with monoallelic expression

Average RPKM values represent the average normalised allele expression levels of three replicates. The fold difference in RPKM values for allele 1 and 2 is calculated. P values were calculated using Fisher's Exact Test. Colours show relativity of values with green representing the lowest value and red representing the highest value.

CA name[1]	SC ortholog[2]	Allele 1	Allele 2	Allele 1 Average RPKM	Allele 2 Average RPKM	P-value	Total Tags (Haploid)
RPL2B	*RPL2B |RPL2A*	orf19.4632	orf19.12102	4584.12	0.00	0	39871
RIB3	*RIB3*	orf19.5228	orf19.12693	0.00	391.31	0	867
YDJ1	*YDJ1*	orf19.506	orf19.8136	0.00	1368.57	0	899
		orf19.1151	orf19.8744	0.00	186.15	5.3E-173	2041
		orf19.1152	orf19.8744	0.00	186.15	5.3E-173	2946
	ZRC1	orf19.1536	orf19.9111	0.00	183.54	7.8E-171	1976
	NRD1	orf19.581	orf19.8212	172.60	0.00	9.7E-153	12924
		orf19.1997	orf19.9548	0.00	149.85	1.2E-139	227
SES1	*SES1*	orf19.269	orf19.7901	0.00	149.50	2.5E-139	49
		orf19.3776	orf19.11257	0.00	132.21	2.8E-123	571
BMT6		orf19.5602	orf19.13045	134.90	0.00	1.2E-119	2806
		orf19.6389	orf19.13747	0.00	89.19	1.93E-83	37352
SGD1	*SGD1*	orf19.4363	orf19.11841	93.72	0.00	2.97E-83	4875
RPN4	*RPN4*	orf19.1069	orf19.8671	77.47	0.00	7.86E-69	1688
	BCP1	orf19.6346	orf19.13702	73.50	0.00	2.72E-65	935
HSP12	*HSP12*	orf19.3160	orf19.10669	71.26	0.00	1.61E-63	2334
	YLR14W	orf19.6630	orf19.13952	0.00	65.67	1.59E-61	416
	TES1	orf19.4122	orf19.11604	67.79	0.00	2.87E-60	1294
		orf19.1437	orf19.9011	67.22	0.00	5.68E-60	760
		orf19.1637	orf19.9205	57.19	0.00	4.55E-51	1988
	ALF1	orf19.2828	orf19.10346	55.11	0.00	5.55E-49	489
	PKR1	orf19.2378	orf19.9914	53.48	0.00	1.73E-47	918
		orf19.2053	orf19.9599	0.00	46.48	1.27E-43	524
		orf19.2731	orf19.10245	48.73	0.00	1.34E-43	311
	TVP38	orf19.5534	orf19.12980	45.62	0.00	5.9E-41	966
	PCP1	orf19.1643	orf19.9211	41.93	0.00	1.01E-37	3253
	NUP145	orf19.748	orf19.8368	41.03	0.00	7.67E-37	1111
SAC3	*SAC3*	orf19.1555	orf19.9129	0.00	35.66	9.57E-34	192
LTP1	*LTP1*	orf19.5104	orf19.12570	34.65	0.00	2.96E-31	1928
	GPI1	orf19.3996	orf19.11479	30.42	0.00	1.97E-27	860
		orf19.1230	orf19.8815	0.00	23.99	6.06E-23	2116
	YNL134C	orf19.2124	orf19.9672	0.00	23.34	2.51E-22	13000
		orf19.4749	orf19.12211	24.42	0.00	3.95E-22	1051
SMI1	*SMI1*	orf19.5058	orf19.12525	22.52	0.00	1.18E-20	7010
		orf19.2913	orf19.10430	0.00	20.99	3.63E-20	862

	BNA2	orf19.583	orf19.8215	19.25	0.00	1.06E-17	3873
		orf19.1440	orf19.9014	18.93	0.00	2.09E-17	600
	BET1	orf19.1386	orf19.8964	17.39	0.00	6.27E-16	695
	ALG3	orf19.1092	orf19.8693	0.00	15.22	6.4E-15	292
	NIC96	orf19.2002	orf19.9553	0.00	14.71	2.65E-14	1242
URA5	*URA5*	orf19.2555	orf19.10087	0.00	14.53	2.65E-14	296
	EXO84	orf19.135	orf19.7779	15.22	0.00	3.75E-14	1136
	SFL1	orf19.3969	orf19.11452	0.00	13.84	1.1E-13	190
		orf19.4332	orf19.11806	0.00	12.90	9.25E-13	11567
	BNA6	orf19.5054	orf19.12521	13.45	0.00	2.25E-12	1420
		orf19.2522	orf19.10057	0.00	12.48	3.83E-12	244
SNX4	*SNX4*	orf19.1990	orf19.9541	13.15	0.00	4.45E-12	456
		orf19.2127	orf19.9674	0.00	12.07	7.79E-12	129359
PHM7	*PHM7*	orf19.2170	orf19.9716	11.09	0.00	2.7E-10	10826
CHS6	*CHS6*	orf19.5155	orf19.12622	10.23	0.00	1.06E-09	2008
		orf19.4138	orf19.11613	0.00	9.36	2.29E-09	1034
PLD1	*SPO14*	orf19.1161	orf19.8753	9.45	0.00	8.29E-09	11203
		orf19.4469	orf19.11950	0.00	8.61	9.49E-09	578
		orf19.4470	orf19.11950	0.00	8.61	9.49E-09	22276
		orf19.3189	orf19.10699	9.10	0.00	1.65E-08	24
PPT1	*PPT1*	orf19.1673	orf19.9242	0.00	8.20	1.93E-08	96
	MUC1	orf19.1725	orf19.9293	8.23	0.00	6.49E-08	996
		orf19.5150	orf19.12615	0.00	7.77	8E-08	886
	RSM25	orf19.4751	orf19.12213	7.68	0.00	2.56E-07	1600
		orf19.4880	orf19.12344	0.00	7.03	3.31E-07	265
SSY1	*SSY1*	orf19.814	orf19.8434	0.00	6.54	6.74E-07	842
SEO1	*SEO1*	orf19.700	orf19.8319	0.00	6.55	6.74E-07	10617
	HIR2	orf19.4295	orf19.11771	0.00	6.16	2.79E-06	203
		orf19.1694	orf19.9261	0.00	5.87	2.79E-06	6544
	KAP14	orf19.3556	orf19.11039	5.86	0.00	7.97E-06	4151
AYR2	*AYR1*	orf19.5615	orf19.13059	5.91	0.00	7.97E-06	377
		orf19.479	orf19.8109	0.00	5.24	1.16E-05	3890
DAL4	*FUR4*	orf19.313	orf19.7944	5.22	0.00	3.16E-05	755
	ASG1	orf19.1497	orf19.9073	0.00	4.35	9.74E-05	3057
ECM21	*CSR2*	orf19.4887	orf19.12351	4.55	0.00	0.000125	281
		orf19.3079	orf19.10592	0.00	3.95	0.000198	4309
	ATG3	orf19.6020	orf19.13441	4.23	0.00	0.00025	99
STR2	*STR2*	orf19.1033	orf19.8635	4.39	0.00	0.00025	1243
RPN6	*RPN6*	orf19.1299	orf19.8879	0.00	3.55	0.000404	114
SUL2	*SUL1*	orf19.2738	orf19.10252	0.00	3.24	0.000821	1927
IFM1	*IFM1*	orf19.5167	orf19.12634	3.29	0.00	0.001979	2993
IFC1	*OPT2*	orf19.3746	orf19.11231	0.00	2.69	0.003401	3203
	BIR1	orf19.643	orf19.8257	0.00	2.86	0.003401	4452
	ECM18	orf19.3607	orf19.11090	0.00	2.19	0.00692	2260
GDH2	*GDH2*	orf19.2192	orf19.9738	0.00	1.62	0.028657	5470
MDN1	*MDN1*	orf19.4697	orf19.12167	1.07	0.00	0.25022	10695

1. *Candida albicans* name according to the *Candida* genome database (www.candidagenome.org).
2. *Saccharomyces cerevisiae* ortholog as given by the *Candida* genome database (www.candidagenome.org).

Table II. Expression data for genes with equally expressed alleles

210 genes with fold difference in allele expression closest to 1.00. Average RPKM values represent the average normalised allele expression levels of three replicates. The fold difference in RPKM values for allele 1 and 2 is calculated. P values were calculated using Fisher's Exact Test. Colours show relativity of values with green representing the lowest value and red representing the highest value.

CA name[1]	SC ortholog[2]	Allele 1	Allele 2	Allele 1 Average RPKM	Allele 2 Average RPKM	Fold Diff	P-value	Total Tags (Haploid)
GLN1	*GLN1*	orf19.646	orf19.8260	550.68	581.50	1.06	0.01	36819
	YKR16W	orf19.4396	orf19.11874	192.08	207.26	1.08	0.06	21919
EFT2	*EFT1| EFT2*	orf19.5788	orf19.13210	1555.37	1446.37	1.08	0.07	426666
SMD3	*SMD3*	orf19.4146	orf19.11622	240.43	251.47	1.05	0.13	1068
	MRPS28	orf19.2520	orf19.10056	239.63	249.63	1.04	0.15	8380
TRP4	*TRP4*	orf19.3099	orf19.10611	172.25	182.00	1.06	0.15	6815
	YKR7W	orf19.4246	orf19.11721	223.60	233.28	1.04	0.16	7962
	YKL27W	orf19.2115	orf19.9663	245.95	255.86	1.04	0.16	11984
	TOM6	orf19.1650	orf19.9219	922.65	923.36	1.00	0.19	3683
		orf19.5547	orf19.12993	767.38	768.02	1.00	0.23	4671
ADO1	*ADO1*	orf19.5591	orf19.13037	119.30	126.09	1.06	0.24	7578
	YER152C	orf19.1180	orf19.8771	128.14	133.29	1.04	0.30	2800
	ADH7	orf19.5517	orf19.12963	72.32	77.21	1.07	0.30	3183
SNF1	*SNF1*	orf19.1936	orf19.9491	47.72	51.22	1.07	0.35	6870
		orf19.1872	orf19.9428	1247.29	1230.02	1.01	0.38	19524
		orf19.1873	orf19.9429	766.30	721.79	1.06	0.39	5473
NAM2	*NAM2*	orf19.5705	orf19.13128	48.32	51.66	1.07	0.39	2766
	PEX7	orf19.89	orf19.7735	49.89	53.49	1.07	0.39	1485
SEC24	*SEC24*	orf19.4732	orf19.12194	100.53	104.32	1.04	0.40	22513
CTA8	*HSF1*	orf19.4775	orf19.12238	59.21	62.55	1.06	0.40	3864
HAS1	*HAS1*	orf19.3962	orf19.11444	66.37	69.25	1.04	0.43	12300
	DCP1	orf19.423	orf19.8053	84.62	87.58	1.04	0.45	1171
ARO1	*ARO1*	orf19.4704	orf19.12175	42.22	44.91	1.06	0.46	21105
SEC2	*SEC2*	orf19.5526	orf19.12972	42.81	45.41	1.06	0.46	738
SPT2	*SPT2*	orf19.422	orf19.8052	46.90	49.63	1.06	0.48	3438
	YML131W	orf19.3139	orf19.10651	31.98	34.30	1.07	0.48	1214
	CCT4	orf19.2720	orf19.10235	106.73	108.70	1.02	0.50	10315
	YLH47|MDM38	orf19.3321	orf19.10831	60.06	62.23	1.04	0.50	7940
	RFC5	orf19.2029	orf19.9577	44.15	46.24	1.05	0.50	1854
		orf19.2022	orf19.9571	346.53	344.69	1.01	0.52	8416
	MTL1	orf19.520	orf19.8151	33.14	34.98	1.06	0.53	1217
	VTI1	orf19.337	orf19.7970	26.76	28.69	1.07	0.53	471
	BUD2	orf19.2934	orf19.10451	113.28	114.94	1.01	0.54	866

ZUO1	*ZUO1*	orf19.2709	orf19.10224	205.33	205.37	1.00	0.55	24121
	YJR1C	orf19.3929	orf19.11411	45.56	47.58	1.04	0.55	1719
		orf19.3871	orf19.11352	29.53	31.40	1.06	0.56	160
NMD3	*NMD3*	orf19.706	orf19.8325	101.66	102.90	1.01	0.57	10177
	GCS1	orf19.3683	orf19.11167	19.87	21.37	1.08	0.59	1823
NAB3	*NAB3*	orf19.5530	orf19.12976	76.52	78.16	1.02	0.61	7643
ERG11	*ERG11*	orf19.922	orf19.8538	54.34	55.73	1.03	0.62	5129
	LAS17	orf19.9	orf19.7682	32.31	33.50	1.04	0.62	753
	RPO21	orf19.177	orf19.7810	0.49	0.52	1.06	0.62	290
	RSM26	orf19.3938	orf19.11420	105.15	105.62	1.00	0.63	2992
	LHP1	orf19.2795	orf19.10313	393.82	387.85	1.02	0.63	5124
DUT1	*DUT1*	orf19.3322	orf19.10832	322.33	304.37	1.06	0.63	5446
	YGR21W	orf19.805	orf19.8423	16.90	18.25	1.08	0.63	3325
CPP1	*MSG5*	orf19.4866	orf19.12330	26.27	27.79	1.06	0.64	1072
SEC12	*SED4*	orf19.3409	orf19.10912	80.77	81.48	1.01	0.65	6604
	KRI1	orf19.1609	orf19.9177	38.87	40.07	1.03	0.65	3586
PSP1	*YLR177W*	orf19.671	orf19.8288	20.96	21.95	1.05	0.66	2087
	CNE1	orf19.5300	orf19.12759	13.07	13.89	1.06	0.66	1266
	YOR52C	orf19.5813	orf19.13235	13.99	14.99	1.07	0.67	824
MNN2	*MNN2*	orf19.2347	orf19.9883	63.35	63.96	1.01	0.68	3620
	YEL43W	orf19.985	orf19.8600	23.51	24.77	1.05	0.68	2751
		orf19.1351	orf19.8931	52.11	52.57	1.01	0.69	686
	CTF4	orf19.6247	orf19.13625	24.96	26.02	1.04	0.69	2124
	SWD3	orf19.3457	orf19.10961	16.64	17.77	1.07	0.69	446
		orf19.4245	orf19.11720	40.75	41.36	1.01	0.70	720
		orf19.5598	orf19.13042	54.86	55.67	1.01	0.70	1372
	NAB3	orf19.1961	orf19.9516	96.82	96.68	1.00	0.71	2331
		orf19.4639	orf19.12108	29.06	29.74	1.02	0.71	719
	SEC16	orf19.4346	orf19.11823	30.67	31.30	1.02	0.71	8855
	DHR2	orf19.107	orf19.7754	59.60	55.19	1.08	0.71	2166
PDC2	*PDC2*	orf19.4863	orf19.12327	21.05	21.83	1.04	0.72	1616
	SGN1	orf19.1389	orf19.8967	718.73	686.77	1.05	0.72	8254
ADE13	*ADE13*	orf19.3870	orf19.11351	187.91	177.63	1.06	0.72	7437
	NPA3	orf19.6463	orf19.13821	50.44	50.75	1.01	0.73	2905
DOT4	*UBP1*	orf19.3370	orf19.10877	67.34	67.82	1.01	0.73	3494
REG1	*REG1*	orf19.2005	orf19.9556	56.22	56.65	1.01	0.74	6523
	PRE4	orf19.4230	orf19.11705	295.88	290.75	1.02	0.74	8162
		orf19.4860	orf19.12323	24.40	25.08	1.03	0.74	156
	YPR45C	orf19.6271	orf19.13650	12.99	13.58	1.04	0.74	2706
ENG1	*DSE4*	orf19.3066	orf19.10584	13.23	14.12	1.07	0.74	5695
GZF3	*GZF3*	orf19.2842	orf19.10361	75.52	70.68	1.07	0.74	3600
IPP1	*IPP1*	orf19.3590	orf19.11072	1348.28	1293.37	1.04	0.75	39645
	YKR43C	orf19.2202	orf19.9748	1.46	1.57	1.08	0.75	22
TOP1	*TOP1*	orf19.96	orf19.7742	62.47	62.44	1.00	0.76	3613
GAD1	*GAD1*	orf19.1153	orf19.8745	16.18	17.08	1.06	0.76	3926
		orf19.4250	orf19.11725	45.17	41.76	1.08	0.76	436

		orf19.6855	orf19.14145	30.32	30.63	1.01	0.77	2163
	TGL4	orf19.1504	orf19.9080	18.10	18.71	1.03	0.77	979
	CMK2	orf19.1754	orf19.9323	1.81	1.88	1.04	0.77	1034
		orf19.2547	orf19.10081	7.63	8.19	1.07	0.77	804
RMS1	*SET7*	orf19.2654	orf19.10177	35.47	35.67	1.01	0.78	1257
	YIL11W	orf19.4760	orf19.12224	18.45	19.16	1.04	0.78	663
	PEX25	orf19.5575	orf19.13021	8.08	8.57	1.06	0.78	515
	SGO1	orf19.3550	orf19.11034	36.93	34.48	1.07	0.78	1041
INT1	*BUD4*	orf19.4257	orf19.11733	36.81	34.08	1.08	0.78	6035
PIF1	*PIF1*	orf19.6133	orf19.13552	59.19	59.14	1.00	0.79	4636
		orf19.551	orf19.8186	20.15	20.28	1.01	0.79	3244
SMC1	*SMC1*	orf19.262	orf19.7895	8.64	9.22	1.07	0.79	762
	PRM5	orf19.3535	orf19.11019	39.31	36.61	1.07	0.79	3890
	YGR54W	orf19.2930	orf19.10447	9.18	8.47	1.08	0.79	7446
	POP1	orf19.2404	orf19.9941	66.03	65.71	1.00	0.80	1623
		orf19.5539	orf19.12985	2.48	2.55	1.03	0.80	160
		orf19.1130	orf19.8723	9.57	10.10	1.05	0.80	975
	DUG1	orf19.3915	orf19.11397	92.49	86.87	1.06	0.80	11959
CDC3	*CDC3*	orf19.1055	orf19.8657	101.11	95.26	1.06	0.80	5712
ASR1	*HPF1*	orf19.2344	orf19.9880	2.48	2.68	1.08	0.80	373
	YLR253W	orf19.4144	orf19.11620	25.76	26.02	1.01	0.81	921
SLD1		orf19.260	orf19.7892	26.63	27.05	1.02	0.81	812
	RIB2	orf19.2788	orf19.10304	11.25	11.56	1.03	0.81	273
		orf19.1841	orf19.9399	2.62	2.83	1.08	0.81	43
	YGR15C	orf19.5704	orf19.13127	28.47	28.83	1.01	0.82	1803
		orf19.4117	orf19.11598	3.14	3.30	1.05	0.82	251
	FAP1	orf19.3722	orf19.11206	54.01	50.61	1.07	0.82	6026
		orf19.4703	orf19.12172	15.04	15.43	1.03	0.83	659
RIM21	*RIM21*	orf19.3176	orf19.10686	13.53	14.16	1.05	0.83	950
KRR1	*KRR1*	orf19.661	orf19.8277	3.36	3.64	1.08	0.83	2295
	SQS1	orf19.2400	orf19.9936	16.98	17.38	1.02	0.84	1031
	RNT1	orf19.3796	orf19.11277	67.51	64.01	1.05	0.84	2540
ERG8	*ERG8*	orf19.4606	orf19.12076	146.71	139.64	1.05	0.84	4750
MUC1	*YBR18W*	orf19.4183	orf19.11659	15.90	14.77	1.08	0.84	2922
	YDR333C	orf19.1864	orf19.9420	17.09	15.77	1.08	0.84	3351
	COP1	orf19.1672	orf19.9241	41.85	39.22	1.07	0.85	7050
	JJJ1	orf19.2399	orf19.9935	22.61	22.53	1.00	0.86	549
ULP2	*ULP1*	orf19.4353	orf19.11831	20.82	20.59	1.01	0.86	510
RAD51	*RAD51*	orf19.3752	orf19.11236	50.50	50.07	1.01	0.86	3469
	OMS1	orf19.1300	orf19.8880	5.47	5.60	1.02	0.86	936
RPS1	*RPS1B*	orf19.3002	orf19.10520	4399.64	4259.88	1.03	0.86	135595
		orf19.1505	orf19.9081	5.48	5.78	1.05	0.86	334
	ELP6	orf19.4701	orf19.12171	5.50	5.78	1.05	0.86	81
PMI1	*PMI4*	orf19.1390	orf19.8968	204.01	195.03	1.05	0.86	7859
CDC2	*CDC2*	orf19.122	orf19.7769	5.56	5.85	1.05	0.87	2351
PGA49	*INO8*	orf19.4404	orf19.11882	5.76	6.05	1.05	0.87	221

MET13	*YMR295C*	orf19.2887	orf19.10405	24.88	23.47	1.06	0.87	590
	HOS4	orf19.3726	orf19.11210	7.44	7.52	1.01	0.88	882
MRF1	*ETR1*	orf19.1149	orf19.8742	7.83	7.90	1.01	0.88	397
	RNH1	orf19.5614	orf19.13057	29.77	29.47	1.01	0.88	1629
ALD5	*ALD5*	orf19.5806	orf19.13228	128.27	125.47	1.02	0.88	89708
ORC1	*ORC1*	orf19.3000	orf19.10518	7.07	7.40	1.05	0.88	3597
	CDC8	orf19.1137	orf19.8730	37.28	37.15	1.00	0.89	358
SSD1	*SSD1*	orf19.3959	orf19.11441	34.30	34.00	1.01	0.89	10334
	YMR134W	orf19.3804	orf19.11285	35.35	34.89	1.01	0.89	493
PRN3		orf19.2462	orf19.9999	8.99	9.17	1.02	0.89	351
FLO8	*SNF5*	orf19.1093	orf19.8695	8.41	8.63	1.03	0.89	7247
SHM1	*SHM1*	orf19.1342	orf19.8922	311.08	302.51	1.03	0.89	16510
	YIA6	orf19.1393	orf19.8971	36.91	35.14	1.05	0.89	1229
	SYN8	orf19.2411	orf19.9949	8.62	8.11	1.06	0.89	346
PRC3	*PRC1*	orf19.2474	orf19.10011	35.00	33.02	1.06	0.89	1513
	FYV7	orf19.4143	orf19.11619	37.31	35.25	1.06	0.89	822
GIN1	*MRC1*	orf19.658	orf19.8274	8.84	9.46	1.07	0.89	2110
	SNA4	orf19.3606	orf19.11089	9.97	10.15	1.02	0.90	523
HEM1	*HEM1*	orf19.2601	orf19.10132	43.77	42.97	1.02	0.90	6615
HSP14	*HPS14*	orf19.6387	orf19.13747	91.45	89.19	1.03	0.90	10432
DQD1		orf19.2283	orf19.9823	97.77	93.47	1.05	0.90	3210
		orf19.752	orf19.8372	51.02	50.03	1.02	0.91	939
		orf19.4195	orf19.11672	14.52	13.72	1.06	0.91	70
ROD1	*ROG3*	orf19.1509	orf19.9084	19.02	18.09	1.05	0.92	2089
TPK1	*TPK2*	orf19.4892	orf19.12357	17.77	16.70	1.06	0.92	897
	YLL23C	orf19.1054	orf19.8656	19.15	18.02	1.06	0.92	853
TPO5	*TPO5*	orf19.151	orf19.7792	19.11	18.97	1.01	0.93	764
FAV2	*WSC3*	orf19.1120	orf19.8718	21.27	20.98	1.01	0.93	172
	PUF2	orf19.921	orf19.8536	25.05	24.68	1.01	0.93	1161
		orf19.1625	orf19.9193	575.05	553.65	1.04	0.93	4211
CMP1	*CMP2*	orf19.6033	orf19.13454	20.97	19.52	1.07	0.93	2888
		orf19.4171	orf19.11647	29.00	28.73	1.01	0.94	1494
	IES2	orf19.3604	orf19.11087	29.18	28.61	1.02	0.94	945
PUP3	*PUP3*	orf19.1336	orf19.8916	117.28	114.28	1.03	0.94	2970
	YMR74C	orf19.713	orf19.8332	249.17	241.91	1.03	0.94	2087
	YKR7W	orf19.449	orf19.8079	27.17	25.83	1.05	0.94	610
	RPB2	orf19.3349	orf19.10857	31.28	29.71	1.05	0.94	18386
CLN3	*CLN3*	orf19.1960	orf19.9515	34.04	32.46	1.05	0.94	1377
	BNA3	orf19.597	orf19.8229	31.14	29.27	1.06	0.94	9570
		orf19.1556	orf19.9129	36.34	35.66	1.02	0.95	1701
RBF1	*DEF1*	orf19.5558	orf19.13004	36.90	36.23	1.02	0.95	4381
	YNL193W	orf19.686	orf19.8304	46.76	45.58	1.03	0.95	2045
	DAP1	orf19.1034	orf19.8636	46.80	44.68	1.05	0.95	463
	YDL119C	orf19.1804	orf19.9370	60.50	58.83	1.03	0.96	1258
	YPL247C	orf19.384	orf19.8014	60.04	57.31	1.05	0.96	3747
	MRPL23	orf19.3348	orf19.10856	383.83	371.61	1.03	0.97	4973

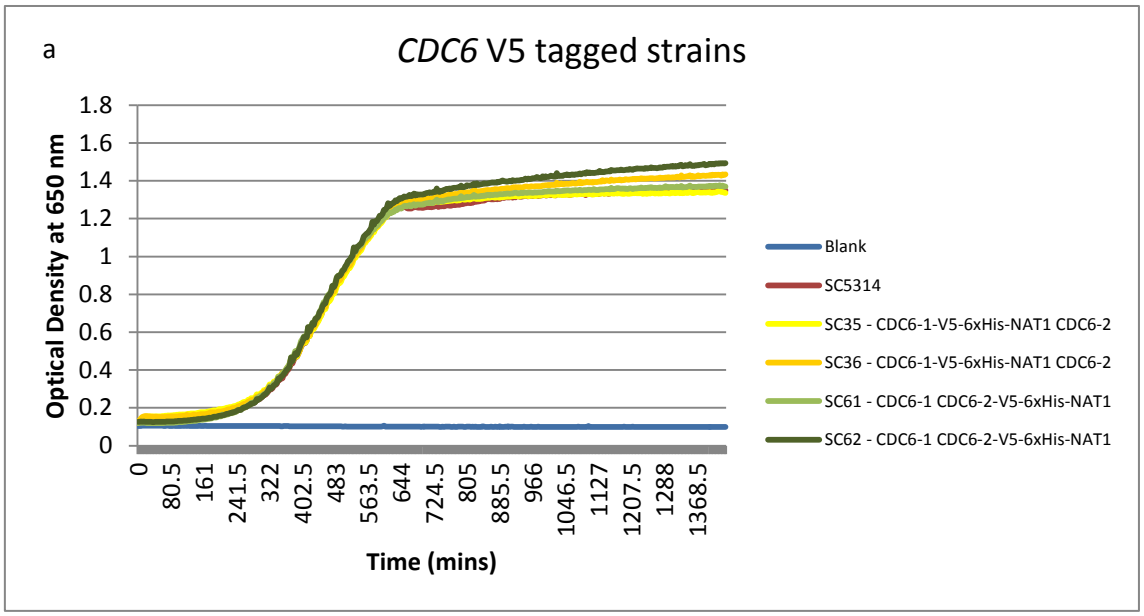
PHR2	*GAS1*	orf19.6081	orf19.13500	175.11	168.77	1.04	0.98	39120
IRA2	*IRA2*	orf19.5219	orf19.12686	0.63	0.64	1.00	1.00	1591
		orf19.3894	orf19.11375	5.95	5.98	1.00	1.00	390
	OTU1	orf19.2933	orf19.10450	6.69	6.70	1.00	1.00	315
		orf19.332	orf19.7954	10.27	10.31	1.00	1.00	4
		orf19.322	orf19.7954	0.99	0.98	1.01	1.00	4
	VMA5	orf19.2166	orf19.9712	2.02	2.00	1.01	1.00	330
	WSC4	orf19.6277	orf19.13656	2.20	2.17	1.01	1.00	84
GCS1	*GSH1*	orf19.5059	orf19.12526	4.53	4.57	1.01	1.00	1357
SFL1	*SFL1*	orf19.454	orf19.8085	8.01	8.08	1.01	1.00	6643
	CST26	orf19.137	orf19.7781	8.59	8.67	1.01	1.00	283
	YDR286C	orf19.319	orf19.7951	11.57	11.46	1.01	1.00	757
	DEF1	orf19.4643	orf19.12113	1.04	1.01	1.02	1.00	1549
		orf19.1301	orf19.8881	1.11	1.09	1.02	1.00	709
	JIP4	orf19.3213	orf19.10725	1.61	1.64	1.02	1.00	1180
		orf19.4783	orf19.12247	1.83	1.87	1.02	1.00	85
	SIN4	orf19.1343	orf19.8923	2.10	2.15	1.02	1.00	1521
	GEM1	orf19.6016	orf19.13437	8.34	8.20	1.02	1.00	594
	MSL1	orf19.4748	orf19.12210	10.62	10.38	1.02	1.00	50
		orf19.247	orf19.7878	15.07	14.77	1.02	1.00	272
	LIN1	orf19.2368	orf19.9904	2.64	2.55	1.03	1.00	427
	SEC23	orf19.1638	orf19.9206	5.66	5.50	1.03	1.00	481
ARP8	*ARP8*	orf19.3359	orf19.10867	8.15	7.88	1.03	1.00	2000
	CSR1	orf19.5711	orf19.13134	10.53	10.23	1.03	1.00	2009
	ENT2	orf19.6309	orf19.13686	13.23	12.83	1.03	1.00	2282
	CUS2	orf19.5767	orf19.13190	14.03	13.58	1.03	1.00	343
HEM14	*HEM14*	orf19.4747	orf19.12209	17.79	17.35	1.03	1.00	511
PEX3	*PEX3*	orf19.4426	orf19.11904	22.41	21.78	1.03	1.00	849
	RLF2	orf19.2739	orf19.10253	23.85	23.07	1.03	1.00	983
TCO89	*CBK1*	orf19.761	orf19.8381	24.61	23.96	1.03	1.00	2203
		orf19.1384	orf19.8963	38.79	37.57	1.03	1.00	1033
	FSH3	orf19.3921	orf19.11403	42.59	41.47	1.03	1.00	3612
BZZ1	*BZZ1*	orf19.1699	orf19.9266	51.41	49.80	1.03	1.00	2090
NOT3	*NOT3*	orf19.2012	orf19.9563	56.33	54.43	1.03	1.00	4269
	YCR9C	orf19.1394	orf19.8972	85.60	82.85	1.03	1.00	1042
FAB1	*FAB1*	orf19.1513	orf19.9088	0.34	0.35	1.04	1.00	1458
	SPT1	orf19.2361	orf19.9897	8.82	8.50	1.04	1.00	498
	COY1	orf19.841	orf19.8461	58.28	56.25	1.04	1.00	2942
	PPA1	orf19.4954	orf19.12419	101.30	97.70	1.04	1.00	2435
		orf19.419	orf19.8049	0.09	0.10	1.05	1.00	14
		orf19.1863	orf19.9419	0.30	0.28	1.05	1.00	6
ALG8	*ALG8*	orf19.1659	orf19.9228	0.69	0.72	1.05	1.00	380
IST2	*IST2*	orf19.2792	orf19.10310	0.76	0.72	1.05	1.00	1658
		orf19.2822	orf19.10340	0.84	0.80	1.05	1.00	26
	RRF1	orf19.477	orf19.8108	1.26	1.19	1.05	1.00	794
	PUT3	orf19.6203	orf19.13584	2.19	2.08	1.05	1.00	2511

	HRD1	orf19.719	orf19.8338	14.27	13.59	1.05	1.00	1579
	DEF1	orf19.1368	orf19.8948	2.93	2.75	1.06	1.00	196
	CTS2	orf19.4984	orf19.12451	4.60	4.34	1.06	1.00	435
VPS41	*VPS41*	orf19.4858	orf19.12321	10.27	9.66	1.06	1.00	696
	RRN1	orf19.718	orf19.8337	1.34	1.26	1.07	1.00	890
		orf19.5151	orf19.12617	1.89	2.03	1.07	1.00	907
TFC4	*TFC4*	orf19.274	orf19.7906	1.98	2.13	1.07	1.00	990
GPI13	*GPI13*	orf19.832	orf19.8452	2.23	2.39	1.07	1.00	761
		orf19.194	orf19.7824	2.52	2.69	1.07	1.00	133
		orf19.4742	orf19.12204	3.67	3.42	1.07	1.00	92
DAC1		orf19.2157	orf19.9704	11.79	11.04	1.07	1.00	383
	HST1	orf19.4761	orf19.12225	3.31	3.06	1.08	1.00	824
		orf19.1496	orf19.9073	4.71	4.35	1.08	1.00	298
	ZRG17	orf19.3769	orf19.11253	11.06	10.26	1.08	1.00	526
	ENT2	orf19.1444	orf19.9019	11.74	10.88	1.08	1.00	3776
MP65	*SCW1*	orf19.1779	orf19.9345	3885.73	4173.17	1.07	6.27E-17	1026

1. *Candida albicans* name according to the *Candida* genome database (www.candidagenome.org).
2. *Saccharomyces cerevisiae* ortholog as given by the *Candida* genome database (www.candidagenome.org).

Figure I Growth curves of V5 tagged strains at 30 °C

a) *CDC6* and b) *VPS1* tagged strains. Allele 1 strains are shown in yellow and allele 2 strains are shown in green.



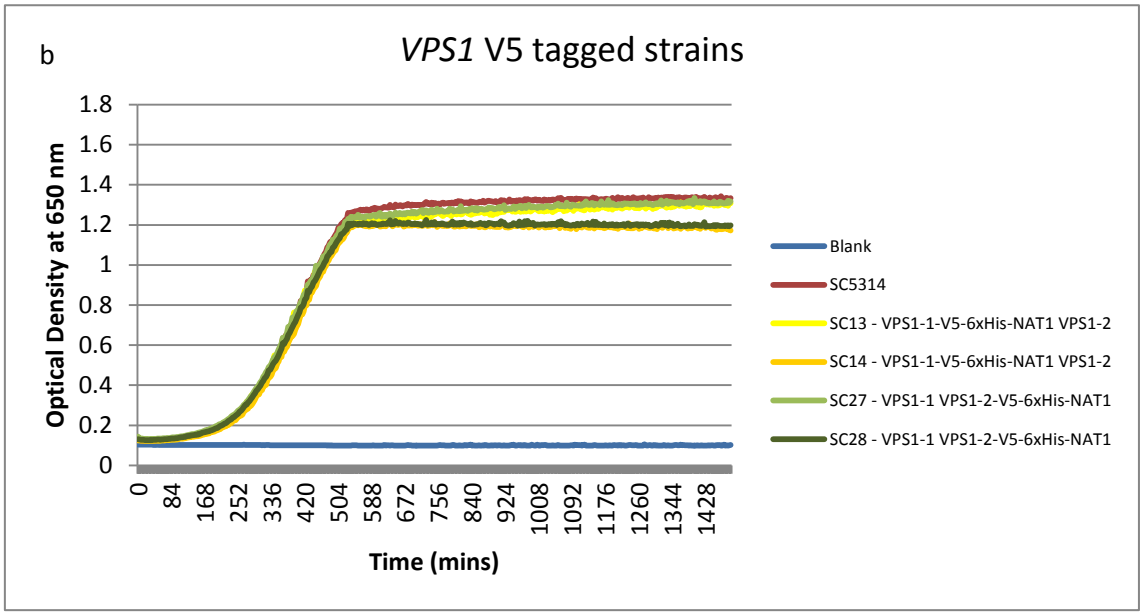


Table III Average generation times, times to maximum inflection and end-point optical densities of V5 tagged strains at 30 °C

(± one standard deviation)

Growth Curve (from Figure I)	Strain	Generation Time (mins)	Time to Maximum Inflection (mins)	End-Point Optical Density (OD at 650 nm)
a) *CDC6*	SC5314	97.37 ± 6.58	326.38 ± 7.21	1.32 ± 0.12
Allele 1	SC35	124.97 ± 53.47	311.50 ± 33.20	1.32 ± 0.05
	SC36	116.84 ± 39.05	317.63 ± 35.62	1.37 ± 0.17
Allele 2	SC61	93.03 ± 4.70	322.00 ± 0.00	1.34 ± 0.18
	SC62	94.10 ± 5.40	326.38 ± 13.21	1.41 ± 0.16
b) *VPS1*	SC5314	110.96 ± 6.11	217.88 ± 4.40	1.32 ± 0.19
Allele 1	SC13	112.85 ± 4.16	211.75 ± 2.02	1.28 ± 0.17
	SC14	111.44 ± 3.07	227.50* ± 4.04	1.20 ± 0.02
Allele 2	SC27	122.66* ± 13.36	203.88* ± 4.40	1.29 ± 0.15
	SC28	116.34 ± 4.66	213.50 ± 4.95	1.22 ± 0.01

* Significantly different measurements from SC5314, identified by ANOVA followed by *post-hoc* analysis using a Dunnett's test, at $p < 0.05$, are annotated with an asterisk.

Figure II Cell cycle distribution of wild-type strain SC5314 demonstrating cell cycle synchronisation by starvation

Samples taken every 15 minutes and analysed as described in section 4.2.3.3.

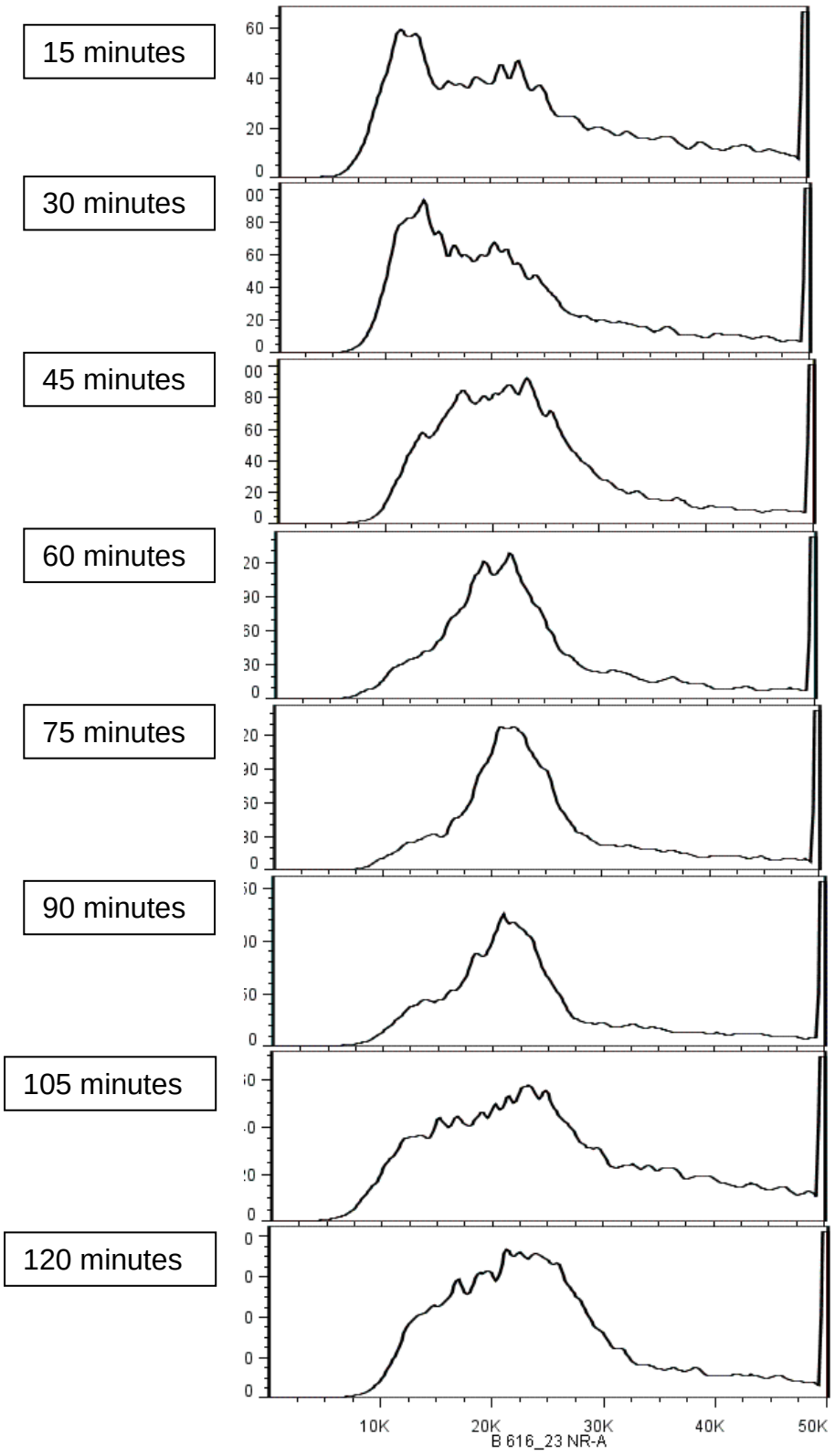


Figure III Corrected sequence of *VPS1* alleles

Erroneous SNPs in the reference genome are highlighted in red, all of which match allele one.

```
orf19.1949|VPS1   ATGGATGAGACATTGATTGCCACCATTAACAAATTACAAGATGCATTGGCTCCTTTAGGT
orf19.9505|VPS1   ATGGATGAGACATTGATTGCCACCATTAACAAATTACAAGATGCATTGGCTCCTTTAGGG
                  ************************************************************

orf19.1949|VPS1   GGAGGGTCATCCTCGCCAGTAGATTTGCCTCAAATCACTGTTGTTGGATCCCAATCCAGT
orf19.9505|VPS1   GGAGGATCATCCTCGCCAGTAGATTTGCCTCAAATCACTGTTGTTGGATCCCAATCCAGT
                  ***** ********************************************************

orf19.1949|VPS1   GGTAAATCGTCAGTATTGGAAAATGTTGTTGGTAGAGACTTTTTACCTAGAGGAACAGGT
orf19.9505|VPS1   GGTAAATCGTCAGTATTGGAAAATGTTGTTGGTAGAGACTTTTTACCTAGAGGAACAGGT
                  ************************************************************

orf19.1949|VPS1   ATTGTTACCAGAAGGCCCTTGGTTTTACAATTAATCAACAGAAGACCAAGCAAGGATTTG
orf19.9505|VPS1   ATTGTTACCAGAAGGCCCTTGGTTTTACAATTAATCAACAGAAGACCAAGCAAGGATTTG
                  ************************************************************

orf19.1949|VPS1   AAGAAAGCTAATGATTTGGTTGATGTTAATGCTTCAGAAAGCACAGGTGGTCAATCAGAA
orf19.9505|VPS1   AAGAAAGCTAATGATTTGGTTGATGTTAATGCTTCAGAAAGCACAGGTGGTCAATCAGAA
                  ************************************************************

orf19.1949|VPS1   AATAATGCTGATGAATGGGGTGAATTTTTGCATTTGCCAGGGAAAAAGTTTTTCAATTTT
orf19.9505|VPS1   AATAATGCTGATGAATGGGGTGAATTTTTGCATTTGCCGGGGAAAAAGTTTTTCAATTTT
                  ************************************ **********************

orf19.1949|VPS1   GAAGATATCAGAAACGAAATTGTTAGAGAAACTGATGCCAAAACAGGTAAGAATTTGGGT
orf19.9505|VPS1   GAAGATATCAGAAACGAAATTGTTAGAGAAACTGATGCCAAAACAGGTAAGAATTTGGGT
                  ************************************************************

orf19.1949|VPS1   ATTTCACCAGTGCCAATCAATTTGAGAATTTACTCTCCTCACGTTTTAACGTTAACTTTA
orf19.9505|VPS1   ATTTCACCAGTGCCAATCAATTTGAGAATTTACTCTCCTCACGTTTTAACGTTAACTTTA
                  ************************************************************

orf19.1949|VPS1   GTTGATTTACCAGGGTTGACAAAAGTCCCCGTTGGTGATCAGCCCAAAGATATTGAAAGG
orf19.9505|VPS1   GTTGATTTACCAGGGTTGACAAAAGTCCCCGTTGGTGATCAGCCCAAAGATATTGAAAGA
                  ************************************************************

orf19.1949|VPS1   CAAATCAAAGATATGATTATGAAATTTATTTCCAAGCCTAACGCCATTATCTTGTCTGTC
orf19.9505|VPS1   CAAATCAAAGATATGATTATGAAATTTATTTCTAAGCCTAACGCCATTATCTTGTCTGTC
                  ******************************** **************************

orf19.1949|VPS1   AACGCTGCTAATACGGATTTGGCTAATTCAGATGGGTTGAAATTAGCAAGAGAAGTTGAC
orf19.9505|VPS1   AACGCTGCTAATACAGATTTGGCCAATTCAGATGGGTTGAAATTAGCAAGAGAAGTTGAC
                  ************** ******** ***********************************

orf19.1949|VPS1   CCTGAAGGTGCAAGAACAATTGGTGTTTTAACCAAAGTGGATTTAATGGATCAAGGTACT
orf19.9505|VPS1   CCTGAAGGTGCAAGAACAATTGGTGTTTTAACCAAAGTGGATTTAATGGATCAAGGTACT
                  ************************************************************

orf19.1949|VPS1   GATGTTATTGACATCTTGGCTGGACGTGTCATCCCATTGAGATTTGGTTATGTTCCAGTG
orf19.9505|VPS1   GATGTTATTGACATCTTGGCTGGACGTGTCATCCCATTGAGATTTGGTTATGTTCCAGTG
                  ************************************************************

orf19.1949|VPS1   ATAAACAGAGGTCAAAAGGATATCGAAGCTAAGAAAACTATCAGGGACGCATTGAAAGAT
orf19.9505|VPS1   ATAAACAGAGGTCAAAAGGATATCGAAGCTAAGAAAACTATCAGGGACGCATTGAAAGAT
                  ************************************************************

orf19.1949|VPS1   GAAAGAAACTTTTTTGAAAATCACCCATCATACAGAGCCAAAGCCCAATTCTGTGGTACG
orf19.9505|VPS1   GAAAGAAACTTTTTTGAAAATCACCCATCATACAGAGCCAAAGCCCAATTCTGTGGTACT
                  ************************************************************
```

```
orf19.1949|VPS1   CCTTACTTGGCCAAGAAATTGAATGGTATTTTGTTGCACCACATCAAGAGTACTTTACCT
orf19.9505|VPS1   CCTTACTTGGCCAAGAAATTGAATGGTATTTTGGTGCACCACATCAAGAGTACTTTACCT
                  ********************************* ************************

orf19.1949|VPS1   GACATCAAGATGAGAATCGAACATTCATTGAAGAAATACCAACAGGAATTATCAATGCTT
orf19.9505|VPS1   GACATCAAGATGAGAATTGAGCATTCATTGAAGAAGTATCACCAAGAATTATCAATGCTT
                  ***************** ** ************** ** ** ** ***************

orf19.1949|VPS1   GACCAGAAATGGCCGAATCTCCTGCATCAATTGCATTGAGTATGATCACTAATTTCTCC
orf19.9505|VPS1   GACCTGAAATGCCTGAGCCTCCTGCATCAATTGCATTGAGTATGACCACTAATTTCTCC
                  **** ****** * **  ************************** *************

orf19.1949|VPS1   AAAGATTACACTGGCATCTTAGATGGTGAATCCAAAGAATTGAGCTCACAAGAATTGAGT
orf19.9505|VPS1   AAAGATTACACTGGCATCTTAGATGGTGAATCCAAAGAATTGAGCTCACAAGAATTGAGT
                  ************************************************************

orf19.1949|VPS1   GGTGGTGCCCGTATTTCCTTTGTGTTTCATGAAATTTTCAAGAATGGGGTTAATGCCATT
orf19.9505|VPS1   GGTGGTGCCCGTATTTCCTTTGTGTTTCATGAAATTTTCAAGAATGGGGTTAATGCCATT
                  ************************************************************

orf19.1949|VPS1   GATCCATTTGATCAAATTAAAGATGCTGATATTAGAACTATTATGCATAATACCTCTGGG
orf19.9505|VPS1   GATCCATTTGATCAAATTAAAGATGCTGATATTAGAACTATTATGCATAATACCTCTGGG
                  ************************************************************

orf19.1949|VPS1   TCGGCACCCTCGTTGTTTGTCGGTACCCAAGCTTTCGAGGTGTTGGTAAGACAACAAATC
orf19.9505|VPS1   TCGGCACCCTCGTTGTTTGTCGGTACCCAAGCTTTCGAGGTGTTGGTAAGACAACAAATC
                  ************************************************************

orf19.1949|VPS1   AAAAGATTGGAAGAACCTTCTATCAGATGTATCAATTTAATTTTCGATGAGTTAGTCAGA
orf19.9505|VPS1   AAAAGATTGGAAGAACCTTCTATCAGATGTATCAATTTAATTTTCGATGAGTTAGTCAGA
                  ************************************************************

orf19.1949|VPS1   ATTTTATCACAAATTATTAGTCAACCACAATATTCAAGATACCCCGGTTTGAAAGAGCAA
orf19.9505|VPS1   ATTTTATCACAAATTATTAGTCAACCACAATATTCAAGATACCCCGGTTTGAAAGAGCAA
                  ************************************************************

orf19.1949|VPS1   TTGTCTCAGAATTTCATTTTATACTTGAGAGATTTGTTGATTCCAACCACTGAGTTTGTC
orf19.9505|VPS1   TTGTCTCAGAATTTCATTTTATACTTGAGAGATTTGTTGATTCCAACCACTGAGTTTGTC
                  ************************************************************

orf19.1949|VPS1   AATGATATAATTCAAGCTGAGGAGACATATGTTAACACTGCTCATCCAGATTTGTTGAAG
orf19.9505|VPS1   AATGATATAATTCAAGCTGAGGAGACATATGTTAACACTGCTCATCCAGATTTGTTGAAG
                  ************************************************************

orf19.1949|VPS1   GGGACACAAGCAATGTCTATTGTGGAAGAGAAGTTCCATCCAAAGCCACAAGTTGCTGTT
orf19.9505|VPS1   GGGACACAAGCAATGTCTATTGTGGAAGAGAAGTTCCATCCAAAGCCACAAGTTGCTGTT
                  ************************************************************

orf19.1949|VPS1   GATCCTAAGACTGGTAAACCATTGCCGCCAAGTCAACAACCAGCACAAGCCACATCACCT
orf19.9505|VPS1   GATCCTAAGACTGGTAAACCATTGCCGCCAAGTCAACAACCAGCACAAGCCACATCACCT
                  ************************************************************

orf19.1949|VPS1   AAACCAGAAGATGGGTCATCTAATGGATTCTTTGGTGGATTCTTTTCTAGCAAAAACAAA
orf19.9505|VPS1   AAACCAGAAGATGGGTCATCTAATGGATTCTTTGGTGGATTCTTTTCTAGCAAAAACAAA
                  ************************************************************

orf19.1949|VPS1   AAGAGATTACAACAAATGGAAGCCCCACCTCCAGTATTGAGAGCCACAGGTACTATGAGT
orf19.9505|VPS1   AAGAGATTACAACAAATGGAAGCCCCACCTCCAGTATTGAGAGCCACAGGTACTATGAGT
                  ************************************************************

orf19.1949|VPS1   GAAAGAGAAACTATGGAAACCGAAGTTATCAAATTATTGATTTCTTCATACTATAATATT
orf19.9505|VPS1   GAAAGAGAAACTATGGAAACCGAAGTTATCAAATTATTGATTTCTTCATACTATAATATT
                  ************************************************************

orf19.1949|VPS1   GTTAAGCGTACTGTTGGTGATGTTGTTCCTAAAGCTATTATGTTGAAATTGATCAACAAA
orf19.9505|VPS1   GTTAAGCGTACTGTTGGTGATGTTGTTCCTAAAGCTATTATGTTGAAATTGATCAACAAA
                  ************************************************************
```

```
orf19.1949|VPS1   TCCAAGGATGAGATCCAAAAGACTTTATTGGAAAAGTTGTACAGCAGTCCAGACTTGGAT
orf19.9505|VPS1   TCCAAGGATGAGATCCAAAAGACTTTATTGGAAAAGTTGTACAGCAGTCCAGACTTGGAT
                  ************************************************************

orf19.1949|VPS1   GATTTGGTTAAGGAAAATGAGCTTACTGTTCAAAAGAGAAAGGAATGTGTTAGAATGGTT
orf19.9505|VPS1   GATTTGGTTAAGGAAAATGAGCTTACTGTTCAAAAGAGAAAGGAATGTGTTAGAATGGTT
                  ************************************************************

orf19.1949|VPS1   GAGGTGTTGAGAAATGCTAGTGAAATTGTTTCTAGTGTTTAG
orf19.9505|VPS1   GAGGTGTTGAGAAATGCTAGTGAAATTGTTTCTAGTGTTTAG
                  ***************************************
```

Figure IV Corrected sequence of *RCK2* alleles

Erroneous SNPs in the reference genome are highlighted in red, all of which match allele two.

```
orf19.2268|RCK2   ATGTTTGAGAATCTCAAAGCTTTTATTCGACATGGGAAGCAAGCCAATGATATGAAAAGA
orf19.9808|RCK2   ATGTTTGAGAATCTCAAAGCTTTTATTCGACATGGGAAGCAAGCCAATGATATGAAAAGA
                  ************************************************************

orf19.2268|RCK2   AAGCAACAACAGCAGCCACAGCAATATCAACAACCATTTAGTACTGCTACTGCCAATGAA
orf19.9808|RCK2   AAGCAACAACAGCAGCCACAGCAATATCAACAACCATTTAGTACTGCTACTGCCAATGAA
                  ************************************************************

orf19.2268|RCK2   AATCCATTTCAACAAGCTTCCAACGAAACTCCAGACAGTATCAATGTTATTACCCCCAAC
orf19.9808|RCK2   AATCCATTTCAACAAGCTTCCAACGAAACTCCAGACAGTATCAATGTTATTACCCCCAAC
                  ************************************************************

orf19.2268|RCK2   GATATCATAAATGAATACCAACAACCAGATCAAGAACCACAACAATACTATCCCCAACAA
orf19.9808|RCK2   GATATCATAAATGAATACCAACAACCAGATCAAGAACCACAACAATACTATCCCCAACAA
                  ************************************************************

orf19.2268|RCK2   CAACAACAACAACAAGACCCATATCAACAGGAAACCCAATTCCAGCAACAGCAACAAGGA
orf19.9808|RCK2   CAACAACAACAACAAGACCCATATCAACAGGAAACCCAATTCCAGCAACAGCAACAAGGA
                  ************************************************************

orf19.2268|RCK2   GTGTATACCAACTATAATCAATCCGATGTTACCCTCAATGACAAAAACGCAGATTACAAT
orf19.9808|RCK2   GTGTATACCAACTATAATCAATCCGATGTTACCCTCAATGACAAAAACGCAGATTACAAT
                  ************************************************************

orf19.2268|RCK2   AGAGTAGCGCTGCAACTTGTTGAAGAAGAGAATGAACAGAGAAAAAAATCTGTCAAATAT
orf19.9808|RCK2   AGAGTAGCGCTGCAACTTGTTGAAGAAGAGAATGAACAGAGAAAAAAATCTGTCAAATAT
                  ************************************************************

orf19.2268|RCK2   CCAAACTTGGAAAATTATCAAATATTAGACCAAATGGGTGAAGGTGCTTTTTCCGTTGTT
orf19.9808|RCK2   CCAAACTTGGAAAATTATCAAATATTAGACCAAATGGGTGAAGGTGCTTTTTCCGTTGTT
                  ************************************************************

orf19.2268|RCK2   TATAAAGCCAAACACTTGTCGACTGGCAAAGAAGTTGCCGTCAAGATTTTGCGCAAGTTT
orf19.9808|RCK2   TATAAAGCCAAACACTTGTCGACTGGCAAAGAAGTTGCCGTCAAGATTTTGCGCAAGTTT
                  ************************************************************

orf19.2268|RCK2   CAAATGGACCAAGCTCAGAAACAGGCCGTACTAAAAGAAGTTACTATTATGAGGCAGTTG
orf19.9808|RCK2   CAAATGGACCAAGCTCAGAAACAGGCCGTACTAAAAGAAGTTACTATTATGAGGCAGTTG
                  ************************************************************

orf19.2268|RCK2   GACCACCCAAATATTGTTAGATTTATTAAATTTATCGACTCCCCAACATACTATTATATT
orf19.9808|RCK2   GACCACCCAAATATTGTTAGATTTATTAAATTTATCGACTCCCCAACATACTATTATATT
                  ************************************************************

orf19.2268|RCK2   GTCCAAGAATTAGTTCCTGGTGGTGAAATCTTCACTATGATTGTGAAGTATACTTATCTT
orf19.9808|RCK2   GTCCAAGAATTAGTTCCTGGTGGTGAAATCTTCACTATGATTGTGAAGTATACTTATCTT
                  ************************************************************

orf19.2268|RCK2   TCTGAAGATTTATCACGTTGGGTGATTACTCAAATTGCTCATGCAATAAGATATTTACAT
orf19.9808|RCK2   TCTGAAGATTTATCACGTTGGGTGATTACTCAAATTGCTCATGCAATAAGATATTTACAT
                  ************************************************************

orf19.2268|RCK2   GAAGAGGTTGGTATTGTCCACCGTGACATTAAGCCAGAAAATTTATTGTATGTACCTATT
orf19.9808|RCK2   GAAGAGGTTGGTATTGTCCACCGTGACATTAAGCCAGAAAATTTATTGTATGTACCTATT
                  ************************************************************

orf19.2268|RCK2   GACTTGAAGCCAAGTGCCAATCCTATATCGAAATTGAGAAAATCCGATGACCCAAACACT
orf19.9808|RCK2   GACTTGAAGCCAAGTGCCAATCCTATATCGAAATTGAGAAAATCCGATGACCCAAACACT
                  ************************************************************
```

```
orf19.2268|RCK2   AAATTAGATGAAGGTGAGTTTGTGAATGGGGTTGGAGGTGGTGGAATTGGGACAGTTAAA
orf19.9808|RCK2   AAATTAGATGAAGGTGAGTTTGTGAATGGGGTTGGAGGTGGTGGAATTGGGACAGTTAAA
                  ************************************************************

orf19.2268|RCK2   TTAGCAGATTTTGGATTATCGAAACAAATATGGGAACATAACACCAAAACACCCTGTGGT
orf19.9808|RCK2   TTAGCAGATTTTGGATTATCGAAACAAATATGGGAACATAACACCAAAACACCCTGTGGT
                  ************************************************************
orf19.2268|RCK2   ACAGTTGGGTATACTGCTCCAGAAATTGTTCGTGATGAGCGCTATTCAAAAGAAGTTGAC
orf19.9808|RCK2   ACAGTTGGGTATACTGCTCCAGAAATTGTTCGTGATGAGCGCTATTCAAAAGAAGTTGAC
                  ************************************************************

orf19.2268|RCK2   ATGTGGGCGTTAGGATGTGTATTGTATACATTGTTATGTGGATTTCCACCGTTTTACGAT
orf19.9808|RCK2   ATGTGGGCGTTAGGATGTGTATTGTATACATTGTTATGTGGATTTCCACCGTTTTACGAT
                  ************************************************************

orf19.2268|RCK2   GAAAGAATCGAAACATTGACTGAAAAAGTTGCCAAAGGTGAATTTACATTTTTGAAACCA
orf19.9808|RCK2   GAAAGAATCGAAACATTGACTGAAAAAGTTGCCAAAGGTGAATTTACATTTTTGAAACCA
                  ************************************************************

orf19.2268|RCK2   TGGTGGGACGAAATAAGTGACGGAGCCAAGAATTGTGTTGGTAGGTTGTTGACTGTGGAC
orf19.9808|RCK2   TGGTGGGACGAAATAAGTGACGGAGCCAAGAATTGTGTTGGTAGGTTGTTGACTGTGGAC
                  ************************************************************

orf19.2268|RCK2   CCAAAAAAGAGGTACACAATTGACGAGTTTTTGCAAGACCCTTGGATGCAAAAAACTTCT
orf19.9808|RCK2   CCAAAAAAGAGGTACACAATTGACGAGTTTTTGCAAGACCCTTGGATGCAAAAAACTTCT
                  ************************************************************

orf19.2268|RCK2   CTTAGTCAGCAACCACAGATTCCAATACCTGTTACTAACCAATACCCACCAGCTACAAAA
orf19.9808|RCK2   CTTAGTCAGCAACCACAGATTCCAATACCTGTTACTAACCAATACCCACCAGCTACAAAA
                  ************************************************************

orf19.2268|RCK2   GTTGCTCATCCTATACAAGTTGCCAATAATAGATACTCCAAGAAGTTTAGATCTACCAAT
orf19.9808|RCK2   GTTGCTCATCCTATACAAGTTGCCAATAATAGATACTCCAAGAAGTTTAGATCTACCAAT
                  ************************************************************

orf19.2268|RCK2   TCTGATTTATATTCTCCTGCAGCTGTTGCTTGGCGCGTTGCCTTTGATATATCTACAGCC
orf19.9808|RCK2   TCTGATTTATATTCTCCTGCAGCTGTTGCTTTGCGTGATGCCTTTGATATATCTACAGCT
                  ******************************** *** * ********************

orf19.2268|RCK2   GACCGCCGTATGGGGGATGGGGCTGCTTTGCAAACTAAAAAGCAAGCCCCAATTGCCGGT
orf19.9808|RCK2   GTTCACCGTATGGGTGAAGAAGCTGCTTTGCAAACTAAAAAGCAAGCCCCAATTGAAGGT
                  *  * ********* ** *  ********************************  ***

orf19.2268|RCK2   TTGATTGAAGAAGAAGAAGAAGAGCACGAAGAAACAGTTACTAAGGACGGCAGAGTTGTA
orf19.9808|RCK2   TTGATTGAAGAAGAAGAAGAAGAGCACGAAGAAACAGTTACTAAGGATGGCAGAGTTGTA
                  ********************************************** ************

orf19.2268|RCK2   CAGGATACCTACAACCAAGTCCCAAGAACTCATCGTCATAGACATCATTTGAAAAATAAC
orf19.9808|RCK2   CAGGATACCTACAACCAAGTCCTAAGAACTCATCGTCATAGACATCATTTGAAAAATAAC
                  ********************** ************************************

orf19.2268|RCK2   AACAATCCAAACGCTTTTGATTTGAATCTTGGAGGTGCATCGATAATAGAACGGAGAAAG
orf19.9808|RCK2   AACAATCCAAACGCTTTTGATTTGAATCTTGGAGGTGCATCGATAATAGAACGGAGAAAG
                  ************************************************************

orf19.2268|RCK2   AACAAACAGATTCCTATTCAAAGCAGCTAG
orf19.9808|RCK2   AACAAACAGATTCCTATTCAAAGCAGCTAG
                  ******************************
```

Table IV ORFs with significant difference between the fold difference in expression of alleles in more than two conditions

Condition Comparison	High Oxidative vs. Low Oxidative	YPD vs. Low Oxidative	YPD vs. High Oxidative	Low Oxidative vs. No Oxidative	High Oxidative vs. No Oxidative	YPD vs. No Nitrosative	Nitrosative vs. No Nitrosative	YPD vs. M199 pH 8	M199 pH 4 vs. M199 pH 8	Congo Red vs. No Congo Red	YPD vs. Nitrosative	YPD vs. M199 pH 4	YPD vs. serum	YPD vs. No Oxidative	YPD vs. Congo Red	YPD vs. No Congo Red
allele1																
orf19.7732	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1
orf19.1265	0	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
orf19.1945	0	1	1	1	1	1	0	1	0	0	0	0	1	0	1	1
orf19.5863	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	1
orf19.4135	0	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1
orf19.12579	0	1	0	0	0	1	0	1	0	0	1	1	1	1	1	1
orf19.10016	0	0	0	1	1	0	1	0	1	0	1	0	1	1	1	1
orf19.11836	1	1	0	0	0	1	0	1	1	0	1	0	0	1	1	0
orf19.9253	0	1	1	1	0	1	1	1	0	0	0	1	0	0	1	0
orf19.131	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1
orf19.3643	0	0	0	1	1	0	0	0	0	0	1	1	0	1	1	1
orf19.877	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
orf19.6350	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1
orf19.2382	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	1
orf19.11233	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
orf19.5602	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
orf19.9178	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
orf19.9142	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
orf19.8485	0	0	0	1	1	1	0	0	0	1	0	0	0	1	1	0
orf19.2413	0	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1
orf19.4706	1	1	0	1	0	0	0	0	1	1	0	1	0	0	0	0
orf19.3395	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	1
orf19.12697	0	0	0	0	0	0	1	0	1	0	1	1	0	0	1	1
orf19.136	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	1
orf19.5548	0	1	0	0	0	1	0	0	0	0	0	0	1	1	1	1
orf19.9419	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
orf19.1122	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
orf19.723	0	0	0	0	0	1	0	1	1	0	0	0	1	1	1	0
orf19.11750	0	0	1	0	0	1	0	1	0	0	0	1	0	1	1	0
orf19.2445	0	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0
orf19.4635	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0

orf19.12995	1	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0
orf19.11980	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.11687	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0
orf19.13026	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0
orf19.9331	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0
orf19.1479	0	1	1	0	0	1	0	1	1	0	0	0	0	0	0	0
orf19.2124	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
orf19.4250	1	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0
orf19.12237	1	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0
orf19.1048	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
orf19.2787	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0
orf19.1763	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0
orf19.11071	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	1
orf19.13747	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1
orf19.13150	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1
orf19.8307	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	1
orf19.2841	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1
orf19.1990	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1
orf19.10252	0	0	0	0	0	1	1	0	0	0	0	0	1	1	0	1
orf19.6143	0	0	0	0	0	0	0	1	0	0	0	1	1	0	1	1
orf19.14149	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
orf19.4067	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
orf19.2789	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1	1
orf19.11220	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0
orf19.1930	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1
orf19.13192	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1
orf19.13076	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0
orf19.12331	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0
orf19.5859	0	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0
orf19.2018	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0
orf19.8395	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0
orf19.12344	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0
orf19.12240	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0
orf19.5302	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
orf19.3080	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0
orf19.9488	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1
orf19.11762	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0
orf19.9930	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
orf19.7889	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1
orf19.5231	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1
orf19.4527	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
orf19.10974	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.8714	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0
orf19.11816	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1
orf19.7954	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
orf19.3727	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0

orf19.12265	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.13289	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0
orf19.13448	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
orf19.13062	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
orf19.13024	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0
orf19.1736	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
orf19.14132	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
orf19.4983	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
orf19.3788	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
orf19.1862	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
orf19.1557	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
orf19.793	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
orf19.4678	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0
orf19.4212	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
orf19.12346	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.3803	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0
orf19.8421	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.10682	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.1133	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.11827	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.5535	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.3869	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.11912	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
orf19.11559	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0
orf19.8882	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0
orf19.4169	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
orf19.5520	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
orf19.12024	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
orf19.10206	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
orf19.8949	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0
orf19.193	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0
orf19.1148	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1
orf19.4972	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
orf19.11726	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
orf19.10952	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
orf19.13909	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0
orf19.9554	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0
orf19.13616	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0
orf19.125	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0
orf19.4063	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
orf19.6487	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0
orf19.532	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0
orf19.97	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.113	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.11957	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.9088	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

orf19.251	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.9403	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
orf19.132	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.8878	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.2751	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.12170	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
orf19.13175	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0
orf19.1915	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0
orf19.4737	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0
orf19.9825	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
orf19.11943	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
orf19.5095	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
orf19.1933	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
orf19.9715	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
orf19.8649	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0
orf19.13034	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0
orf19.9115	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0
orf19.8652	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0
orf19.11844	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
orf19.10368	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
orf19.11256	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
orf19.13163	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
orf19.1529	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
orf19.3940	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
orf19.3894	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
orf19.11189	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
orf19.1373	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
orf19.4690	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
orf19.13840	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
orf19.10182	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.11469	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.290	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.258	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.12529	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
orf19.4117	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
orf19.10147	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
orf19.5616	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
orf19.4691	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
orf19.12355	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
orf19.7718	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
orf19.10881	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
orf19.4607	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
orf19.4901	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
orf19.4689	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
orf19.6923	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
orf19.3460	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0

orf19.8707	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
orf19.805	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.3158	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.10399	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
orf19.2468	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
orf19.11253	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
orf19.3526	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
orf19.12007	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
orf19.454	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
orf19.1258	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
orf19.8753	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
orf19.7705	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
orf19.11244	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
orf19.1532	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
orf19.10860	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
orf19.1570	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
orf19.728	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
orf19.13019	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.13065	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.2781	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.317	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.3781	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.3801	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.8768	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.6464	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
orf19.8737	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.11139	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.2907	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
orf19.10676	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
orf19.3623	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
orf19.13908	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
orf19.465	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
orf19.10004	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.10973	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.12171	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.1351	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.14177	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.8319	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.9012	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.8389	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
orf19.10584	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
orf19.11402	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
orf19.12400	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
orf19.13250	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
orf19.2381	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
orf19.7861	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0

orf19.6272	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
orf19.11254	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
orf19.1139	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
orf19.124	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
orf19.2746	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
orf19.1308	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
orf19.2928	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
orf19.3535	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
orf19.13615	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.7895	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.4470	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.8267	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.2006	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
orf19.9382	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
orf19.9312	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.11219	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
orf19.10259	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
orf19.2649	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
orf19.4770	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
orf19.5606	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
orf19.7803	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
orf19.12337	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
orf19.9891	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
orf19.4438	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
orf19.10196	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
orf19.11110	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
orf19.1356	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
orf19.477	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
orf19.12970	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
orf19.5518	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
orf19.7688	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
orf19.1120	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
orf19.9748	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
orf19.10310	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.9706	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.6014	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.3170	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.153	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.13614	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.12432	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
orf19.10657	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0

Table V Genes with significant disparities in fold difference of allele levels over a significant number of condition comparisons

P value calculated as described in section 5.2.6.

Gene Name	Allele 1	Allele 2	Number of Comparisons where Gene is Significant	P value
GPX1	orf19.87	orf19.7732	12	3.76E-17
	orf19.1265	orf19.8850	10	2.33E-14
ADH2	orf19.5113	orf19.12579	9	4.00E-13
UGA4	orf19.2479	orf19.10016	9	4.70E-13
	orf19.5863	orf19.13285	9	6.17E-13
AUR1	orf19.1945	orf19.9500	9	7.36E-13
PRC2	orf19.4135	orf19.11612	9	8.57E-13
	orf19.131	orf19.7777	8	6.13E-12
	orf19.4358	orf19.11836	8	9.26E-12
	orf19.1684	orf19.9253	8	1.57E-11
	orf19.877	orf19.8496	7	1.38E-10
	orf19.6350	orf19.13707	7	2.02E-10
	orf19.2382	orf19.9918	7	2.34E-10
	orf19.3643	orf19.11126	7	3.07E-10
IFC1	orf19.3746	orf19.11233	7	3.48E-10
BCR1	orf19.723	orf19.8342	6	3.54E-09
	orf19.4706	orf19.12177	6	3.60E-09
CSI2	orf19.5232	orf19.12697	6	3.66E-09
	orf19.1863	orf19.9419	6	3.83E-09
	orf19.1122	orf19.8720	6	3.83E-09
LYS14	orf19.5548	orf19.12994	6	4.15E-09
	orf19.2413	orf19.9951	6	4.89E-09
	orf19.136	orf19.7780	6	5.30E-09
RAD32	orf19.866	orf19.8485	6	5.34E-09
	orf19.3395	orf19.10898	6	6.25E-09
PUT1	orf19.4274	orf19.11750	6	6.29E-09
UTP22	orf19.1569	orf19.9142	6	7.05E-09
	orf19.1610	orf19.9178	6	8.65E-09
BMT6	orf19.5602	orf19.13045	6	8.65E-09
NIP1	orf19.4635	orf19.12105	5	7.60E-08
	orf19.6143	orf19.13562	5	7.87E-08
FET3	orf19.4211	orf19.11687	5	7.89E-08
SNX4	orf19.1990	orf19.9541	5	7.94E-08
SUL2	orf19.2738	orf19.10252	5	8.03E-08
	orf19.6859	orf19.14149	5	8.20E-08
	orf19.2445	orf19.9981	5	8.44E-08
PGM2	orf19.2841	orf19.10359	5	8.53E-08

TEL1	orf19.5580	orf19.13026	5	8.62E-08
HSP104	orf19.6387	orf19.13747	5	8.83E-08
	orf19.5728	orf19.13150	5	8.83E-08
	orf19.2789	orf19.10305	5	9.76E-08
FGR18	orf19.4067	orf19.11550	5	1.02E-07
	orf19.5549	orf19.12995	5	1.14E-07
PLB1	orf19.689	orf19.8307	5	1.16E-07
	orf19.1479	orf19.9054	5	1.25E-07
	orf19.2124	orf19.9672	5	1.31E-07
SPO11	orf19.3589	orf19.11071	5	1.34E-07
	orf19.4250	orf19.11725	5	1.37E-07
AOX2	orf19.4773	orf19.12237	5	1.37E-07
OCA1	orf19.1762	orf19.9331	5	1.54E-07
IFR1	orf19.1763	orf19.9332	5	1.77E-07
PRY1	orf19.2787	orf19.10303	5	1.94E-07
IFD6	orf19.1048	orf19.8650	5	1.96E-07
	orf19.3735	orf19.11220	4	1.63E-06
CFL5	orf19.1930	orf19.9486	4	1.65E-06
OPT8	orf19.5770	orf19.13192	4	1.65E-06
PGA31	orf19.5302	orf19.12761	4	1.80E-06
	orf19.2394	orf19.9930	4	1.89E-06
	orf19.4286	orf19.11762	4	1.94E-06
	orf19.2018	orf19.9568	4	1.97E-06
	orf19.5631	orf19.13076	4	2.03E-06
	orf19.1117	orf19.8714	4	2.16E-06
	orf19.773	orf19.8395	4	2.18E-06
	orf19.4341	orf19.11816	4	2.25E-06
	orf19.257	orf19.7889	4	2.36E-06
	orf19.5231	orf19.12696	4	2.36E-06
SWE1	orf19.4867	orf19.12331	4	2.41E-06
	orf19.3470	orf19.10974	4	2.42E-06
DAL8	orf19.5859	orf19.13281	4	2.61E-06
	Orf19.4880	orf19.12344	4	2.66E-06
CFL4	orf19.1932	orf19.9488	4	2.69E-06
LYS143	orf19.4776	orf19.12240	4	2.69E-06
	orf19.3080	orf19.10592	4	3.08E-06
	orf19.322	orf19.7954	4	3.08E-06
HGT1	orf19.4527	orf19.12002	4	3.17E-06
PHO112	orf19.3727	orf19.11211	4	4.40E-06
FTH1	orf19.4802	orf19.12265	4	4.40E-06
	orf19.1050	orf19.8652	3	3.15E-05
	orf19.1302	orf19.8882	3	3.22E-05
	orf19.6027	orf19.13448	3	3.25E-05
	orf19.5617	orf19.13062	3	3.25E-05
WSC1	orf19.5867	orf19.13289	3	3.42E-05
	orf19.5752	orf19.13175	3	3.55E-05

HNM1	orf19.2003	orf19.9554	3	3.58E-05
	orf19.4700	orf19.12170	3	3.67E-05
	orf19.6236	orf19.13616	3	3.72E-05
	orf19.3448	orf19.10952	3	3.76E-05
	orf19.2691	orf19.10206	3	3.77E-05
	orf19.4972	orf19.12437	3	3.81E-05
	orf19.4078	orf19.11559	3	3.99E-05
FET99	orf19.4212	orf19.11689	3	4.09E-05
	orf19.4983	orf19.12450	3	4.19E-05
	orf19.4169	orf19.11645	3	4.25E-05
CAK1	orf19.793	orf19.8412	3	4.35E-05
ZCF22	orf19.4251	orf19.11726	3	4.42E-05
	orf19.1541	orf19.9115	3	4.45E-05
	orf19.1862	orf19.9418	3	4.50E-05
	orf19.2285	orf19.9825	3	4.53E-05
	orf19.4463	orf19.11943	3	4.53E-05
ERB1	orf19.1047	orf19.8649	3	4.62E-05
	orf19.5587	orf19.13034	3	4.62E-05
SPC34	orf19.3788	orf19.11268	3	4.69E-05
MPP10	orf19.1915	orf19.9471	3	4.71E-05
	orf19.1148	orf19.8740	3	4.76E-05
	orf19.2169	orf19.9715	3	4.81E-05
	orf19.5095	orf19.12561	3	4.95E-05
	orf19.1933	orf19.9489	3	4.95E-05
	orf19.1369	orf19.8949	3	4.98E-05
TPO3	orf19.4737	orf19.12199	3	5.09E-05
ASG7	orf19.5520	orf19.12966	3	5.10E-05
	orf19.193	orf19.7823	3	5.26E-05
FGR38	orf19.4549	orf19.12024	3	5.38E-05
	orf19.4883	orf19.12346	3	5.70E-05
NUP84	orf19.1298	orf19.8878	3	5.75E-05
	orf19.2751	orf19.10265	3	5.75E-05
	orf19.132	orf19.7778	3	5.75E-05
MNN22	orf19.3803	orf19.11284	3	5.87E-05
CPH1	orf19.4433	orf19.11912	3	6.02E-05
	orf19.1844	orf19.9403	3	6.03E-05
	orf19.4678	orf19.12147	3	6.31E-05
	orf19.6487	orf19.14137	3	6.35E-05
RBR2	orf19.532	orf19.8165	3	6.35E-05
	orf19.1736	orf19.9304	3	6.40E-05
TUS1	orf19.6842	orf19.14132	3	6.46E-05
	orf19.1557	orf19.9130	3	6.61E-05
	orf19.6556	orf19.13909	3	6.62E-05
	orf19.5535	orf19.12981	3	6.66E-05
	orf19.3869	orf19.11350	3	6.66E-05
	orf19.4349	orf19.11827	3	6.66E-05

MSB1	orf19.1133	orf19.8726	3	6.66E-05
	orf19.3172	orf19.10682	3	6.66E-05
EBP1	orf19.125	orf19.7772	3	6.79E-05
GPT1	orf19.4063	orf19.11546	3	6.84E-05
	orf19.4504	orf19.11980	5	7.15E-05
	orf19.5578	orf19.13024	3	7.20E-05
FAB1	orf19.1513	orf19.9088	3	8.33E-05
GLX3	orf19.251	orf19.7882	3	8.33E-05
	orf19.4476	orf19.11957	3	8.33E-05
CIP1	orf19.113	orf19.7761	3	8.33E-05
CAN1	orf19.97	orf19.7744	3	8.93E-05

Figure V Corrected sequence of *RPS7A* alleles

Erroneous SNPs in the reference genome are highlighted in red, all of which match allele two.

```
orf19.1700|RPS7A  ATGTCCTCTAAGATCTTATCAGAAAACCCAACTGAATTAGAATTAAAAGTTGCTCAAGCT
orf19.9267|RPS7A  ATGTCCTCTAAGATCTTATCAGAAAACCCAACTGAATTAGAATTAAAAGTTGCTCAAGCT
                  ************************************************************

orf19.1700|RPS7A  TTCGTTGATTTGGAATCTCAAGCTGATTTAAAAGCTGAATTGAGACCATTACAATTCAAA
orf19.9267|RPS7A  TTCGTTGATTTGGAATCTCAAGCTGATTTAAAAGCTGAATTGAGACCATTACAATTCAAA
                  ************************************************************

orf19.1700|RPS7A  TCTATCAAAGAAATTGATGTTAATGGAGGTAAAAAAGCTTTAGCTGTTTTCGTTCCACCA
orf19.9267|RPS7A  TCTATCAAAGAAATTGATGTTAATGGAGGTAAAAAAGCTTTAGCTGTTTTCGTTCCACCA
                  ************************************************************

orf19.1700|RPS7A  CCAAGTTTACAAGCTTACAGAAAAGTTCAAACTAGATTAACTAGAGAATTAGAAAAAAAA
orf19.9267|RPS7A  CCAAGTTTACAAGCTTACAGAAAAGTTCAAACTAGATTAACTAGAGAATTAGAAAAAAAA
                  ************************************************************

orf19.1700|RPS7A  TTCCCAGATAGACATGTTGTCTTTTTAGCTGAAAGAAGAATCTTACCAAAACCAGCTAGA
orf19.9267|RPS7A  TTCCCAGATAGACATGTTGTCTTTTTAGCTGAAAGAAGAATCTTACCAAAACCAGCTAGA
                  ************************************************************

orf19.1700|RPS7A  AAAGCTAGAAAACAGCAAAAAAGACCAAGATCAAGAACTTTGACTGCTGTTCATGATAAA
orf19.9267|RPS7A  AAAGCTAGAAAACAACAAAAAAGACCAAGATCAAGAACTTTGACTGCTGTTCATGATAAA
                  ************** ************************************************

orf19.1700|RPS7A  ATTTTGGAAGATTTAGTTTTCCCAACTGAAATCATTGGTAAAAGAGTTAGATACTTGGTT
orf19.9267|RPS7A  ATTTTGGAAGATTTAGTTTTCCCAACTGAAATCATTGGTAAAAGAGTTAGATACTTGGTT
                  ************************************************************

orf19.1700|RPS7A  GGTGGTAACAAAATCCAAAAAGTCTTGTTGGATTCTAAAGATTCAACTGCTGTTGATTAC
orf19.9267|RPS7A  GGTGGTAACAAAATCCAAAAAGTCTTGTTGGATTCTAAAGATTCAACTGCTGTTGATTAC
                  ************************************************************

orf19.1700|RPS7A  AAATTGGATTCTTTCCAACAATTGTACTCAAAATTGACTGGTAAACAAGTTGTTTTTGAA
orf19.9267|RPS7A  AAATTGGATTCTTTCCAACAATTGTACTCAAAATTGACTGGTAAACAAGTTGTTTTTGAA
                  ************************************************************

orf19.1700|RPS7A  ATCCCAGGTGAATCTCATTAG
orf19.9267|RPS7A  ATCCCAGGTGAATCTCATTAG
                  *********************
```

Figure VI Corrected sequence of orf19.5648 alleles

Erroneous SNPs in the reference genome are highlighted; blue corresponds to SNPS which match allele one and red corresponds to SNPS which match allele two.

```
orf19.5648       ATGATAAACAGTGGTAATGGTTGTTGCTGTTGTTTTTTTCTTGCTGGTCGCCTTTCATAT
orf19.13093      ATGATAAACAGTGGTAATGGTTGTTGCTGTTGTTTTTTTCTTGCTGGTCGCCTTTCATAT
                 ************************************************************

orf19.5648       TTTTTTTTCAGATGTTGGCAAAAATGGACTGAAAAAAAAATCGAAAAAAAAGTTGAAAGC
orf19.13093      TTTTTTTTCAGATGTTGGCAAAAATGGACTGAAAAAAAAATCGAAAAAAAAGTTGAAAGC
                 ************************************************************

orf19.5648       TGCGGTCGTGTCTCATGTCCAAATCAAGCGTTATTGGAATTTTGGCTTATATTACATGAA
orf19.13093      TGCGGTCGTGTCTCATGCCCAAATCAAGCGTTATTGGAATTTTGGCTTATATTACATGAA
                 ***************** ****************************************

orf19.5648       AATTCTGAGAAGTTTCTCCATTCACCTTTTTCTTACAAGTACCGAAATATGTTTAGAATT
orf19.13093      AATTCTGAGAAGTTTCTCCATTCACCTTTTTCTTACAAGTACCGAAATATGTTTAGAATT
                 ************************************************************

orf19.5648       GTTGCCAGAGCCCCTAGGATACTCCCATATCGTCGATTTACCACTACCCCTAGCTTGAGG
orf19.13093      GTTGCCAGAGCCCCTAGGATACTCCCATATCGTCGATTTACCACTACCCCTAGCTTGAGG
                 ************************************************************

orf19.5648       TTTTTTGACAAAGGGCCTACTGCTGAAGAACAGGCCCAGGCATTGGAAAAGGTCAGTAAA
orf19.13093      TTTTTTGACAAAGGGCCTACTGCTGAAGAACAGGCCCAGGCATTGGAAAAGGTCAGTAAA
                 ************************************************************

orf19.5648       GTGGTTGCCGAAAACCCAGAGTTGTACAAGTTGATGGTTGAATTTAAACAGTTACTTGAA
orf19.13093      GTGGTTGCCGAAAACCCAGAGTTGTACAAGTTGATGGTTGAATTCAAACAGTTACTTGAC
                 ****************************************** **************

orf19.5648       AAGAAAGGATTTGAAACCGGGGCAAAACCATCTATGACTCAAATGTTTAAATTGTTGGCC
orf19.13093      CAGAAAGGGTTTGAATCATGGGCAATACCATCTATGACTCAAATGTTTAAATTGTTGGCC
                  ******* ****** *  ****** ********************************

orf19.5648       GACAAAGATATCAGAGAACATGGTGCCAAATTCAAACACTTTTTGGAAACTACAGACACA
orf19.13093      GACAAAGATATCAGAGAACATCGTGCCAAATTCAAACACTTGTTGGTGACTACAGACACA
                 ********************* ****************** ****  ************

orf19.5648       GGACTCACTCAAAATGAGATCGCAACTGTAAGTGGTGCATT-TTTATTCAAAAATAAAGA
orf19.13093      GGACTCACTCAAGATGAGATCGCAACTGTATATGGTGCATTATTTATTCAGACATGA---
                 ************ *****************  ********* ******** * ** *

orf19.5648       TATTAAATAG
orf19.13093      ----------
```

Appendix II – Perl Scripts Written and Used

represents a comment

II.I Script to identify frequency of CUG codons within an open reading frame

```
#!usr/bin/perl/
# CUG_script.pl
use strict; use warnings;

#This programme takes a file containing the list of genes of interest and the
fasta file with all sequences.
#It outputs a file containing just the sequences of the genes of interest.

die "not enough arguments\n" unless @ARGV == 3;
#This bit makes an array of the genes of interest from file specified in
command line-this needs to be a text file with one gene per line.

open (GENELIST, "<$ARGV[0]") or die "error opening genenames for reading\n";
my @genes = ();                          #declares empty array
while (my $gene = <GENELIST>) {
       chomp $gene;
       push (@genes, $gene);             #adds genes to the array, one at a
time
}
close GENELIST;
my $length = @genes;
print "Amount = $length\n";              # checking how many genes have been
                                         inputted

my $i = 0;
my $found = 0;
while ($i < $length) {
       open(IN,  "<$ARGV[1]")  or  die  "error  opening  sequence  file  for
reading\n";
       open(OUT, ">>$ARGV[2].txt") or die "error creating file for reading";

              while (my $line = <IN>) {  #goes through line by line
              if ($line =~ /$genes[$i] /) {    #matches  genename  from  array
@genes
```

```
                    my $check =0;       #reset to 0
                    my $found ++;       #confirms match
                    my $seq = <IN>;     #makes $seq the new seq, overwrites
                                        previous.
                    while ($check == 0) {
                          my $nextline = <IN>;              #takes next line
                          if ($nextline =~ /^[ATGC]/){$seq = $seq.$nextline}
                                                            #adds it to utr
                          else {$check +=1}          #stops it from taking
                                                     nextline if not a
                                                     sequence line
                    }
                    chomp $seq;
                    $seq =~ s/\s//g;
                    $seq =~ s/\n//g;    #trying to get rid of spaces or
                                        newlines in utr sequence

                    my $count = 0;
                    for (my $i = 0; $i < length($seq) - 3 + 1; $i+= 3){
#sliding window
                          my $codon = substr($seq, $i, 3); #extracts 3 bp
                          if ($codon eq "CTG")              {$count++} #if CUG
+ 1
                    }
                    print OUT "$genes[$i]\t$count\n";
              }
       }
       close IN;
       close OUT;
$i ++;
print "$i\n";
}
```

II.II Script to identify all SNP locations in the genome

```
#!usr/bin/perl/
#match_allele_sequence.pl
use strict; use warnings;

die "Insufficient Input Files\n" unless @ARGV == 2; #inputs - 1 = allele list.
2 = sequence file.

open(ALLELE, "<$ARGV[0]"); #open allele list
while(my $line = <ALLELE>){
	open (SEQ, "<$ARGV[1]"); #open sequence file
	open (OUT, ">>seq.txt");
	chomp $line;
	my @allele = split("\t",$line); #split allele list into array
	my $allele1 = $allele[0];
	my $allele2 = $allele[1];
	my $seqcount = 0;
	while (my $seq = <SEQ>){ #runs through sequence file
		if ($seq =~ m/$allele1 /){ #if lines matches allele one
			chomp $seq;
			print OUT "$seq\n"; #print line name, new line
			my $check = 0;
			my $seq2 = <SEQ>;
			while ($check == 0) {
				my $nextline = <SEQ>;	#takes next line
				if ($nextline =~ /^[ATGC]/){$seq2 = $seq2.$nextline}
								#adds it to sequence
				else {$check +=1}	#stops it from taking
								nextline if not a
								sequence line
			}
			chomp $seq2;
			$seqcount ++;
			print OUT "$seq2\n";		#prints sequence in
single line
		}
		elsif ($seq =~ m/$allele2 /){		#as above with allele 2
			chomp $seq;
			print OUT "$seq\n";
			my $check = 0;
			my $seq2 = <SEQ>;
			while ($check == 0) {
				my $nextline = <SEQ>;	#takes next line
				if ($nextline =~ /^[ATGC]/){$seq2 = $seq2.$nextline}
								#adds it to sequence
```

```
                        else {$check +=1}          #stops it from taking
                                                   nextline if not a
                                                   sequence line
                }
                chomp $seq2;
                $seqcount ++;
                print OUT "$seq2\n";
        }
    }
    close SEQ;
    close OUT;

    if ($seqcount == 2){} #if loop to move on if the sequences have not been
    found for both alleles
    else {
        system("rm seq.txt");
        next;
    }

    #print "Allele 1 $allele1\nAllele 2 $allele2\n";

    system("muscle -in seq.txt -out align.txt"); #run muscle using sequence
                                                 file, outputting alignment
    system("rm seq.txt"); # removes sequence file ready for next one.

    use Bio::SeqIO; #use of Bio::SeqIO to open file align.txt and split into
                    ID, sequence and description
    my $inseq = Bio::SeqIO->new('-file' => "align.txt",'-format' => 'fasta'
) ;

    my $seq_obj1 = $inseq->next_seq;
    my $id1 = $seq_obj1->id ;
    my $aligned_seq1 = $seq_obj1->seq ;
    my $desc1 = $seq_obj1->description ;

    my $seq_obj2 = $inseq->next_seq;
    my $id2 = $seq_obj2->id ;
    my $aligned_seq2 = $seq_obj2->seq ;
    my $desc2 = $seq_obj2->description ;

    #print   "ID   =   $id1\nSEQ   =$aligned_seq1\nID   =   $id2\nSEQ   =
$aligned_seq2\n";

    ### Iterate through the alignment and check for differences
    my $pos_in_seq1 = 1;
    my $pos_in_seq2 = 1;
    foreach my $i (1 .. length($aligned_seq1)) {
```

```
            my $seq1char = substr($aligned_seq1, ($i-1), 1);
            my $seq2char = substr($aligned_seq2, ($i-1), 1);

            if ($seq1char eq $seq2char) { ### they are the same
                  $pos_in_seq1++;
                  $pos_in_seq2++;
            }
            elsif ($seq1char =~ m/^[a-z]$/i and $seq2char =~ m/^[a-z]$/i) { #
SNP
                  print "$id1\t$pos_in_seq1\n$id2\t$pos_in_seq2\n";
                  $pos_in_seq1++;
                  $pos_in_seq2++;
                  ## record this as SNP position
                  }
            elsif ($seq1char =~ m/^[a-z]$/i and $seq2char =~ m/^\-$/) { # A
                                                                  deletion   in
                                                                  sequence 2
                  #print "$id1\t$pos_in_seq1\n";
                  $pos_in_seq1++;
                  #$snps_positions_in_seq1{$pos_in_seq1} = 1;
            }
            elsif ($seq2char =~ m/^[a-z]$/i and $seq1char =~ m/^\-$/) { # A
                                                                  deletion   in
                                                                  sequence 1
                        #print "$id2\t$pos_in_seq2\n";
                  $pos_in_seq2++;
                  #$snps_positions_in_seq2{$pos_in_seq2};
                   }
            else {
                  die "This should never happen!";
            }
      }
      system("rm align.txt"); # removes alignment file ready for next one.
}

close ALLELE;
```

II.III Script to filter mpileup file based upon SNP locations

Written by Paul O'Neill

```
#!/usr/bin/perl

# takes input as contig \t pos and loads that value into specifies database

use strict;
use Getopt::Long;
use IO::File;
use PileupFunctions;

my $usage = qq/
   Usage is:
   findCandidateLOHRegions.pl
              -mp|mp_file <mp_file> location of mp_file
              -ref|reference <ref.fasta>   # location of reference file (to map
                                           sequence names to ids)
              -snp|snp_file <file containing SNPs.
/;


# process user options
my ($mp_file, $snp_file, $reference);
GetOptions (
       'ref=s'             => \$reference,
       'mp=s'        => \$mp_file,
       'snp=s'             => \$snp_file
);

       my $snp_fh = IO::File->new();
       open ($snp_fh, $snp_file) || die "Could not open file".$mp_file.": $!";

       # initialise feed for pileup
       my $pf = new PileupFunctions($mp_file, $reference);

       while (my $line = readline($snp_fh)) {
              chomp $line;
              my ($seq, $pos) = split ( /\t/, $line);
              warn sprintf "%s\t%s\n", $seq, $pos;

           if ( $pf->advance_to($seq, $pos)) {
                    $pf->count_frequencies;
                    printf
                    "%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\n", (
                           $seq, $pos,
```

```
                        $pf->cov,
                        $pf->allele1, $pf->a1_fwd_cov, $pf->a1_rev_cov, $pf-
>a1_fwd_qual, $pf->a1_rev_qual,
                        $pf->allele2, $pf->a2_fwd_cov, $pf->a2_rev_cov, $pf-
>a2_fwd_qual, $pf->a2_rev_qual);
            } else {
                  printf ("%s\t%s not found.\n", $seq, $pos);
            }

      }
```

II.IV Script to total reads aligned to each allele

```
#!usr/bin/perl/
#allele_total.pl
use strict; use warnings;

die "Insufficient input files\n" unless @ARGV == 1; #input = shortened mpileup

open(IN, "<$ARGV[0]");
my %allele_total; #empty hash
while(my $line = <IN>){
	chomp $line;
	my @info = split("\t",$line);
	if(exists $allele_total{$info[0]} and $info[1] =~ /not found/){ #if gene
                                                                    is already in hash
		next;
	}
	elsif(exists $allele_total{$info[0]} and $info[1] =~ /\d/){
		$allele_total{$info[0]} += $info[2]; #add on the read count
	}
	elsif($info[1] =~ /not found/){
		$allele_total{$info[0]} = 0;
	}
	else{
		$allele_total{$info[0]} = $info[2];
	}
}
close IN;

#print hash
open(OUT, ">>allele_total.txt");
foreach my $key (keys %allele_total){
	print OUT "$key\t$allele_total{$key}\n";
}
close OUT;
```

II.V Script to match allele counts/RPKM values for each gene

```
#!usr/bin/perl/
#match_pairs.pl
use strict; use warnings;

die "Insufficient Input files\n" unless @ARGV == 2; #Input 1 = allele pairs
Input 2 = allele_counts

open(PAIRS, "<$ARGV[0]");
while(my $line = <PAIRS>){
	chomp $line;
	my @pairs = split("\t",$line);
	my @pairedcount = ();
	open(COUNTS, "<$ARGV[1]");
	open(OUT, ">>out_file.txt");
	while(my $line2 = <COUNTS>){
		chomp $line2;
		my @counts = split("\t",$line2);
		if ($pairs[0] eq $counts[0]){
			push(@pairedcount, $counts[0]);
			push(@pairedcount, $counts[3]);
		}
		elsif ($pairs[1] eq $counts[0]){
			push(@pairedcount, $counts[0]);
			push(@pairedcount, $counts[3]);
		}
	}
	if (@pairedcount == 4){
	my $pairedcount = join("\t",@pairedcount);
	print OUT "$pairedcount\n";
	}
	else{
	next;
	}
	close COUNTS;
	close OUT;
}
close PAIRS;
```

II.VI Script to calculate RPKM values

```
#!usr/bin/perl/
#rpkm.pl
use strict; use warnings;

die "Insufficient input files\n" unless @ARGV == 2; #Input 1 = allele count.
                                                    Input 2 = lengths.

open (COUNTS, "<$ARGV[0]");
while(my $line = <COUNTS>){
       chomp $line;
       my @counts = split("\t",$line);
       open (LENGTH, "<$ARGV[1]");
       while (my $line2 = <LENGTH>){
              chomp $line2;
              my @lengths = split("\t",$line2);
              open (OUT, ">>rpkm.txt");
              if ($counts[0] eq $lengths[0]){
                     my $top = $counts[1]/($lengths[1]/1000);
                     my $reads = total number of reads aligned/1000000;
                     my $rpkm = $top / $reads;
                     printf  OUT  "%s\t%d\t%d\t%e\n",  $counts[0],  $counts[1],
                     $lengths[1], $rpkm,;
                     #print OUT "$rpkm\n";
                     }
       close OUT;
       }
}

close COUNTS;
```

References

ABBEY, D., HICKMAN, M., GRESHAM, D. & BERMAN, J. 2011. High-resolution SNP/CGH microarrays reveal the accumulation of loss of heterozygosity in commonly used *Candida albicans* strains. *G3: Genes, Genomes, Genetics,* 1**,** 523-530.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol.,* 11**,** R106.

ANDRULIS, E. D., NEIMAN, A. M., ZAPPULLA, D. C. & STERNGLANZ, R. 1998. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature,* 394**,** 592-595.

ARONESTY, E. 2011. ea-utils: "Command-line tools for processing biological sequencing data". Durham, NC: Expression Analysis.

ARTIERI, C. G. & FRASER, H. B. 2014. Evolution at two levels of gene expression in yeast. *Genome Research*, 24, 411-421.

BASSO, L. R., BARTISS, A., MAO, Y., GAST, C. E., COELHO, P. S. R., SNYDER, M. & WONG, B. 2010. Transformation of *Candida albicans* with a synthetic hygromycin B resistance gene. *Yeast,* 27**,** 1039-1048.

BATEMAN, E. & PAULE, M. R. 1988. Promoter occulsion during ribosomal-RNA transcription. *Cell,* 54**,** 985-992.

BATES, S., HALL, R., CHEETHAM, J., NETEA, M., MACCALLUM, D., BROWN, A., ODDS, F. & GOW, N. 2013. Role of the *Candida albicans MNN1* gene family in cell wall structure and virulence. *BMC Research Notes,* 6**,** 294.

BECKER, J. M., KAUFFMAN, S. J., HAUSER, M., HUANG, L., LIN, M., SILLAOTS, S., JIANG, B., XU, D. & ROEMER, T. 2010. Pathway analysis of *Candida albicans* survival and virulence determinants in a murine infection model. *Proceedings of the National Academy of Sciences,* 107**,** 22044-22049.

BENNETT, R. J. & JOHNSON, A. D. 2003. Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *The EMBO Journal,* 22**,** 2505-2515.

BENNETZEN, J. L. & HALL, B. D. 1982. Codon selection in yeast. *Journal of Biological Chemistry,* 257**,** 3026-3031.

BENSEN, E. S., MARTIN, S. J., LI, M., BERMAN, J. & DAVIS, D. A. 2004. Transcriptional profiling in *Candida albicans* reveals new adaptive responses to extracellular pH and functions for Rim101p. *Molecular Microbiology,* 54**,** 1335-1351.

BERGER, B., PENG, J. & SINGH, M. 2013. Computational solutions for omics data. *Nat Rev Genet,* 15**,** 333-346.

BERMAN, J. 2006. Morphogenesis and cell cycle progression in *Candida albicans*. *Current Opinion in Microbiology,* 9**,** 595-601.

BERMAN, J. & SUDBERY, P. 2002. *Candida albicans*: A molecular revolution built on lessons from budding yeast. *Nat Rev Genet,* 3**,** 918-932.

BERNARDO, S. M., KHALIQUE, Z., KOT, J., JONES, J. K. & LEE, S. A. 2008. *Candida albicans VPS1* contributes to protease secretion, filamentation, and biofilm formation. *Fungal Genetics and Biology,* 45**,** 861-877.

BERTRAM, G., SWOBODA, R. K., GOODAY, G. W., GOW, N. A. R. & BROWN, A. J. P. 1996. Structure and regulation of the *Candida albicans ADH1* gene encoding an immunogenic alcohol dehydrogenase. *Yeast,* 12**,** 115-127.

BISWAS, S., VAN DIJCK, P. & DATTA, A. 2007. Environmental sensing and signal transduction pathways regulating morphopathogenic determinants of *Candida albicans*. *Microbiology and Molecular Biology Reviews,* 71**,** 348-376.

BLIGNAUT, E., PUJOL, C., LOCKHART, S., JOLY, S. & SOLL, D. R. 2002. Ca3 fingerprinting of *Candida albicans* isolates from human immunodeficiency virus-positive and healthy individuals reveals a new clade in South Africa. *Journal of Clinical Microbiology,* 40**,** 826-836.

BLOOM, J., KHAN, Z., KRUGLYAK, L., SINGH, M. & CAUDY, A. 2009. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics,* 10**,** 221.

BOORE, J. L. 1999. Animal mitochondrial genomes. *Nucleic Acids Research,* 27**,** 1767-1780.

BORST, P. & CHAVES, I. 1999. Mono-allelic expression of genes in simple eukaryotes. *Trends in Genetics,* 15**,** 95-96.

BOUMIL, R. M. & LEE, J. T. 2001. Forty years of decoding the silence in X-chromosome inactivation. *Human Molecular Genetics,* 10**,** 2225-2232.

BRAND, A., MACCALLUM, D. M., BROWN, A. J. P., GOW, N. A. R. & ODDS, F. C. 2004. Ectopic expression of *URA3* can influence the virulence phenotypes and proteome of *Candida albicans* but can be overcome by targeted reintegration of *URA3* at the *RPS10* locus. *Eukaryotic Cell,* 3**,** 900-909.

BRANNAN, C. I. & BARTOLOMEI, M. S. 1999. Mechanisms of genomic imprinting. *Current Opinion in Genetics Development,* 9**,** 164-170.

BRAUN, B. R., HEAD, W. S., WANG, M. X. & JOHNSON, A. D. 2000. Identification and characterization of *TUP1*-regulated genes in *Candida albicans*. *Genetics,* 156**,** 31-44.

BRAUN, B. R., VAN HET HOOG, M., D'ENFERT, C., MARTCHENKO, M., DUNGAN, J., KUO, A., INGLIS, D. O., UHL, M. A., HOGUES, H., BERRIMAN, M., LORENZ, M., LEVITIN, A., OBERHOLZER, U., BACHEWICH, C., HARCUS, D., MARCIL, A., DIGNARD, D., LOUK, T., ZITO, R., FRANGEUL, L., TEKAIA, F., RUTHERFORD, K., WANG, E., MUNRO, C. A., BATES, S., GOW, N. A. R., HOYER, L. L., KOHLER, G., MORSCHHAUSER, J., NEWPORT, G., ZNAIDI, S., RAYMOND, M., TURCOTTE, B., SHERLOCK, G., COSTANZO, M. C., IHMELS, J., BERMAN, J., SANGLARD, D., AGABIAN, N., MITCHELL, A. P., JOHNSON, A., WHITEWAY, M. & NANTEL, A. 2005. A human-curated annotation of the *Candida albicans* genome. *PLoS genetics,* 1, 36-57.

BRETAGNE, S., COSTA, J. M., BESMOND, C., CARSIQUE, R. & CALDERONE, R. 1997. Microsatellite polymorphism in the promoter sequence of the elongation factor 3 gene of *Candida albicans* as the basis for a typing system. *Journal of Clinical Microbiology,* 35**,** 1777-1780.

BRUNO, V. M., WANG, Z., MARJANI, S. L., EUSKIRCHEN, G. M., MARTIN, J., SHERLOCK, G. & SNYDER, M. 2010. Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Research,* 20**,** 1451-1458.

BRUZUAL, I. & KUMAMOTO, C. 2011. An *MDR1* promoter allele with higher promoter activity is common in clinically isolated strains of *Candida albicans*. *Molecular Genetics and Genomics***,** 286, 347-357.

BUENO, A. & RUSSELL, P. 1992. Dual functions of *CDC6*: a yeast protein required for DNA replication also inhibits nuclear division. *The EMBO Journal,* 11**,** 2167.

BULLARD, J., PURDOM, E., HANSEN, K. & DUDOIT, S. 2010a. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics,* 11**,** 94.

BULLARD, J. H., MOSTOVOY, Y., DUDOIT, S. & BREM, R. B. 2010b. Polygenic and directional regulatory evolution across pathways in *Saccharomyces*. *Proceedings of the National Academy of Sciences,* 107**,** 5058-5063.

BUSTIN, S. A., BENES, V., GARSON, J. A., HELLEMANS, J., HUGGETT, J., KUBISTA, M., MUELLER, R., NOLAN, T., PFAFFL, M. W., SHIPLEY, G. L., VANDESOMPELE, J. & WITTWER, C. T. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry,* 55**,** 611-622.

BUTLER, G., RASMUSSEN, M. D., LIN, M. F., SANTOS, M. A. S., SAKTHIKUMAR, S., MUNRO, C. A., RHEINBAY, E., GRABHERR, M., FORCHE, A. & REEDY, J. L. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature,* 459**,** 657-662.

CALDERONE, R. 2002a. Introduction and Historical Perspectives. *In:* CALDERONE, R. (ed.) *Candida and candidiasis.* Washington D.C.: ASM Press.

CALDERONE, R. A. 2002b. *Candida and candidiasis*, Amer Society for Microbiology.

CASTAÑO, I., PAN, S., ZUPANCIC, M., HENNEQUIN, C., DUJON, B. & CORMACK, B. 2005. Telomere length control and transcriptional regulation of subtelomeric adhesins in *Candida glabrata*. *Molecular Microbiology,* 55**,** 1246-1258.

CHANDRA, J., KUHN, D. M., MUKHERJEE, P. K., HOYER, L. L., MCCORMICK, T. & GHANNOUM, M. A. 2001. Biofilm Formation by the Fungal Pathogen *Candida albicans*: Development, Architecture, and Drug Resistance. *Journal of Bacteriology,* 183**,** 5385-5394.

CHANDRA, J., MUKHERJEE, P. K. & GHANNOUM, M. A. 2012. *Candida* biofilms associated with CVC and medical devices. *Mycoses,* 55**,** 46-57.

CHEETHAM, J. 2008. *The regulation of the Hog1 stress activated protein kinase in Candida albicans* Ph.D., Newcastle University.

CHEN, W. H., TRACHANA, K., LERCHER, M. J. & BORK, P. 2012. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol Biol Evol,* 29**,** 1703-1706.

CHEN, X., WEAVER, J., BOVE, B. A., VANDERVEER, L. A., WEIL, S. C., MIRON, A., DALY, M. B. & GODWIN, A. K. 2008. Allelic imbalance in *BRCA1* and *BRCA2* gene expression is associated with an increased breast cancer risk. *Human Molecular Genetics,* 17**,** 1336-1348.

CHESS, A. 2012. Mechanisms and consequences of widespread random monoallelic expression. *Nat Rev Genet,* 13**,** 421-428.

CHESS, A., SIMON, I., CEDAR, H. & AXEL, R. 1994. Allelic inactivation regulates olfactory receptor gene expression. *Cell,* 78**,** 823-834.

CHEUNG, V. G., NAYAK, R. R., WANG, I. X., ELWYN, S., COUSINS, S. M., MORLEY, M. & SPIELMAN, R. S. 2010. Polymorphic *cis*- and *trans*-regulation of human gene expression. *PLoS Biol,* 8**,** e1000480.

COCKER, J. H., PIATTI, S., SANTOCANALE, C., NASMYTH, K. & DIFFLEY, J. F. X. 1996. An essential role for the Cdc6 protein in forming the pre-replicative complexes of budding yeast. *Nature,* 379**,** 180-182.

COGHLAN, A. & WOLFE, K. H. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast,* 16**,** 1131-1145.

COSTE, A., TURNER, V., ISCHER, F., MORSCHHAUSER, J., FORCHE, A., SELMECKI, A., BERMAN, J., BILLE, J. & SANGLARD, D. 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.

COTE, P., HOGUES, H. & WHITEWAY, M. 2009. Transcriptional analysis of the *Candida albicans* cell cycle. *Molecular biology of the cell,* 20**,** 3363-3373.

COTTER, G., DOYLE, S. & KAVANAGH, K. 2000. Development of an insect model for the *in vivo* pathogenicity testing of yeasts. *FEMS Immunology & Medical Microbiology,* 27**,** 163-169.

CULLEN, B. R., LOMEDICO, P. T. & JU, G. 1984. Transcriptional interference in avian retroviruses - implications for the promoter insertion model of leukemogenesis. *Nature,* 307**,** 241-245.

DA SILVA, C. R., DE ANDRADE NETO, J. B., CAMPOS, R. D. S., FIGUEIREDO, N. S., SAMPAIO, L. S., FERREIRA MAGALHÃES, H. I., CAVALCANTI, B. C., MACEDO, D. G., DE ANDRADE, G. M., PAMPOLHA LIMA, I. S., VIANA, G. S. D. B., DE MORAES, M. O., PINTO LOBO, M. D., GRANGEIRO, T. B. & NOBRE JÚNIOR, H. V. 2013. Synergistic effect of the flavonoids catechin, quercetin and epigallocatechin gallate with fluconazole induce apoptosis in *Candida tropicalis* resistant to fluconazole. *Antimicrobial agents and chemotherapy*, 58, 1468-1478.

DAGUE, E., BITAR, R., RANCHON, H., DURAND, F., YKEN, H. M. & FRANÇOIS, J. M. 2010. An atomic force microscopy analysis of yeast mutants defective in cell wall architecture. *Yeast,* 27**,** 673-684.

DEGNER, J. F., MARIONI, J. C., PAI, A. A., PICKRELL, J. K., NKADORI, E., GILAD, Y. & PRITCHARD, J. K. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA sequencing data. *Bioinformatics,* 25**,** 3207-3212.

DENG, Q., RAMSKÖLD, D., REINIUS, B. & SANDBERG, R. 2014. Single-Cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science,* 343**,** 193-196.

DENNISON, P. M. J., RAMSDALE, M., MANSON, C. L. & BROWN, A. J. P. 2005. Gene disruption in *Candida albicans* using a synthetic, codon-optimised Cre-*loxP* system. *Fungal Genetics and Biology,* 42**,** 737-748.

DETWEILER, C. S. & LI, J. J. 1997. Cdc6p establishes and maintains a state of replication competence during G1 phase. *Journal of Cell Science,* 110**,** 753-763.

DEVEALE, B., VAN DER KOOY, D. & BABAK, T. 2012. Critical evaluation of imprinted gene expression by RNA–seq: a new perspective. *PLoS Genet,* 8**,** e1002600.

DODGSON, A., DODGSON, K., PUJOL, C., PFALLER, M. & SOLL, D. 2004. Clade-specific flucytosine resistance is due to a single nucleotide change in the *FUR1* gene of *Candida albicans*. *Antimicrobial agents and chemotherapy,* 48**,** 2223.

DUDLEY, A. M., JANSE, D. M., TANAY, A., SHAMIR, R. & CHURCH, G. M. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol,* 1, 2005.0001.

DUMITRU, R., NAVARATHNA, D. H. M. L. P., SEMIGHINI, C. P., ELOWSKY, C. G., DUMITRU, R. V., DIGNARD, D., WHITEWAY, M., ATKIN, A. L. & NICKERSON, K. W. 2007. *In vivo* and *in vitro* anaerobic mating in *Candida albicans*. *Eukaryotic Cell,* 6**,** 465-472.

DVIR, S., VELTEN, L., SHARON, E., ZEEVI, D., CAREY, L. B., WEINBERGER, A. & SEGAL, E. 2013. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences,* 110**,** E2792-E2801.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32**,** 1792-1797.

EDMOND, M. B., WALLACE, S. E., MCCLISH, D. K., PFALLER, M. A., JONES, R. N. & WENZEL, R. P. 1999. Nosocomial bloodstream infections in United States hospitals: a three-year analysis. *Clinical Infectious Diseases,* 29**,** 239-244.

ENJALBERT, B., NANTEL, A. & WHITEWAY, M. 2003. Stress-induced gene expression in *Candida albicans*: absence of a general stress response. *Molecular biology of the cell,* 14**,** 1460-1467.

ENJALBERT, B., SMITH, D. A., CORNELL, M. J., ALAM, I., NICHOLLS, S., BROWN, A. J. P. & QUINN, J. 2006. Role of the *Hog1* stress-activated protein kinase in the global transcriptional response to stress in the fungal pathogen *Candida albicans*. *Molecular Biology of the Cell,* 17**,** 1018-1032.

ESTEVE-CODINA, A., KOFLER, R., PALMIERI, N., BUSSOTTI, G., NOTREDAME, C. & PEREZ-ENCISO, M. 2011. Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics,* 12**,** 552.

FEINBERG, A. P. 1993. Genomic imprinting and gene activation in cancer. *Nat Genet,* 4**,** 110-113.

FERNÁNDEZ-ARENAS, E., CABEZÓN, V., BERMEJO, C., ARROYO, J., NOMBELA, C., DIEZ-OREJAS, R. & GIL, C. 2007. Integrated proteomics and genomics strategies bring new insight into *Candida albicans*

response upon macrophage interaction. *Molecular & Cellular Proteomics,* 6**,** 460-478.

FONZI, W. A. & IRWIN, M. Y. 1993. Isogenic strain construction and gene mapping in *Candida albicans*. *Genetics,* 134**,** 717-728.

FORCHE, A., ABBEY, D., PISITHKUL, T., WEINZIERL, M. A., RINGSTROM, T., BRUCK, D., PETERSEN, K. & BERMAN, J. 2011. Stress alters rates and types of loss of heterozygosity in *Candida albicans*. *mBio,* 2.

FORCHE, A., ALBY, K., SCHAEFER, D., JOHNSON, A. D., BERMAN, J. & BENNETT, R. J. 2008. The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biol,* 6**,** e110.

FORCHE, A., MAY, G. & MAGEE, P. T. 2005. Demonstration of loss of heterozygosity by single-nucleotide polymorphism microarray analysis and alterations in strain morphology in *Candida albicans* strains during infection. *Eukaryotic Cell,* 4**,** 156-165.

FRADIN, C., DE GROOT, P., MACCALLUM, D., SCHALLER, M., KLIS, F., ODDS, F. C. & HUBE, B. 2005. Granulocytes govern the transcriptional response, morphology and proliferation of *Candida albicans* in human blood. *Molecular Microbiology,* 56**,** 397-415.

FU, Y., IBRAHIM, A. S., FONZI, W., ZHOU, X., RAMOS, C. F. & GHANNOUM, M. A. 1997. Cloning and characterization of a gene (*LIP1*) which encodes a lipase from the pathogenic yeast *Candida albicans*. *Microbiology,* 143**,** 331.

GAGNEUR, J., SINHA, H., PEROCCHI, F., BOURGON, R., HUBER, W. & STEINMETZ, L. M. 2009. Genome-wide allele- and strand-specific expression profiling. *Mol Syst Biol,* 5, 274.

GARAIZAR, J., BRENA, S., BIKANDI, J., REMENTERIA, A. & PONTÓN, J. 2006. Use of DNA microarray technology and gene expression profiles to investigate the pathogenesis, cell biology, antifungal susceptibility and diagnosis of *Candida albicans*. *FEMS Yeast Research,* 6**,** 987-998.

GARTLER, S. M. & GOLDMAN, M. A. 2005. X chromosome inactivation. *Encyclopedia of Life Sciences*.

GAUR, U., LI, K., MEI, S. & LIU, G. 2013. Research progress in allele-specific expression and its regulatory mechanisms. *Journal of Applied Genetics***,** 1-13.

GE, B., GURD, S., GAUDIN, T., DORE, C., LEPAGE, P., HARMSEN, E., HUDSON, T. J. & PASTINEN, T. 2005. Survey of allelic expression using EST mining. *Genome Research,* 15**,** 1584-1591.

GEISS, G. K., BUMGARNER, R. E., BIRDITT, B., DAHL, T., DOWIDAR, N., DUNAWAY, D. L., FELL, H. P., FERREE, S., GEORGE, R. D., GROGAN, T., JAMES, J. J., MAYSURIA, M., MITTON, J. D., OLIVERI, P., OSBORN, J. L., PENG, T., RATCLIFFE, A. L., WEBSTER, P. J., DAVIDSON, E. H., HOOD, L. & DIMITROV, K. 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotech,* 26**,** 317-325.

GERADS, M. & ERNST, J. F. 1998. Overlapping coding regions and trancriptional units of two essential chromosomal genes (*CCT8, TRP1*) in the fungal pathogen *Candida albicans*. *Nucleic Acids Research,* 26**,** 5061-5066.

GHANNOUM, M. A. & RICE, L. B. 1999. Antifungal agents: mode of action, mechanisms of resistance, and correlation of these mechanisms with bacterial tesistance. *Clinical Microbiology Reviews,* 12**,** 501-517.

GILLUM, A., TSAY, E. & KIRSCH, D. 1984. Isolation of the *Candida albicans* gene for orotidine-5 -phosphate decarboxylase by complementation of *S. cerevisiae ura3* and *E. coli pyrF* mutations. *Molecular and General Genetics MGG,* 198**,** 179-182.

GIMELBRANT, A., HUTCHINSON, J. N., THOMPSON, B. R. & CHESS, A. 2007. Widespread monoallelic expression on human autosomes. *Science,* 318**,** 1136-1140.

GIRARDOT, M., FEIL, R. & LLÈRES, D. 2013. Epigenetic deregulation of genomic imprinting in humans: causal mechanisms and clinical implications. *Epigenomics,* 5**,** 715-728.

GÓMEZ-RAJA, J., ANDALUZ, E., MAGEE, B., CALDERONE, R. & LARRIBA, G. 2008. A single SNP, G929T (Gly310Val), determines the presence of a functional and a non-functional allele of *HIS4* in *Candida albicans* SC5314: detection of the non-functional allele in laboratory strains. *Fungal Genetics and Biology,* 45**,** 527-541.

GONCALVES, I., DURET, L. & MOUCHIROUD, D. 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Research,* 10**,** 672-678.

GOTTSCHLING, D. E., APARICIO, O. M., BILLINGTON, B. L. & ZAKIAN, V. A. 1990. Position effect at *S. cerevisiae* telomeres: Reversible repression of Pol II transcription. *Cell,* 63**,** 751-762.

GREGG, C., ZHANG, J., WEISSBOURD, B., LUO, S., SCHROTH, G. P., HAIG, D. & DULAC, C. 2010. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science,* 329**,** 643-648.

GU, Z., STEINMETZ, L. M., GU, X., SCHARFE, C., DAVIS, R. W. & LI, W.-H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature,* 421**,** 63-66.

GUDLAUGSSON, O., GILLESPIE, S., LEE, K., BERG, J. V., HU, J. F., MESSER, S., HERWALDT, L., PFALLER, M. & DIEKEMA, D. 2003. Attributable mortality of nosocomial candidemia, revisited. *Clinical Infectious Diseases,* 37**,** 1172-1177.

GUIDA, A., LINDSTADT, C., MAGUIRE, S., DING, C., HIGGINS, D., CORTON, N., BERRIMAN, M. & BUTLER, G. 2011. Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast *Candida parapsilosis*. *BMC Genomics,* 12**,** 628.

GUO, M., RUPE, M. A., ZINSELMEIER, C., HABBEN, J., BOWEN, B. A. & SMITH, O. S. 2004. Allelic variation of gene expression in maize hybrids. *The Plant Cell Online,* 16**,** 1707-1716.

HALL, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series,* 41**,** 95-98.

HAMPSEY, M. 1997. A Review of Phenotypes in *Saccharomyces cerevisiae*. *Yeast,* 13**,** 1099-1133.

HARRIES, L. W., ELLARD, S., STRIDE, A., CONSORTIUM, T. E. M., MORGAN, N. G. & HATTERSLEY, A. T. 2006. Isomers of the *TCF1* gene encoding hepatocyte nuclear factor-1 alpha show differential expression in the pancreas and define the relationship between mutation position and clinical phenotype in monogenic diabetes. *Human Molecular Genetics,* 15**,** 2216-2224.

HARTWELL, L. H., MORTIMER, R. K., CULOTTI, J. & CULOTTI, M. 1973. Genetic control of the cell division cycle in yeast: V. Genetic analysis of *cdc* mutants. *Genetics,* 74**,** 267.

HAWSER, S. P. & DOUGLAS, L. J. 1994. Biofilm formation by *Candida* species on the surface of catheter materials *in vitro*. *Infection and Immunity,* 62**,** 915-921.

HAWSER, S. P. & DOUGLAS, L. J. 1995. Resistance of *Candida albicans* biofilms to antifungal agents *in vitro*. *Antimicrobial agents and chemotherapy,* 39**,** 2128-31.

HEAP, G. A., YANG, J. H. M., DOWNES, K., HEALY, B. C., HUNT, K. A., BOCKETT, N., FRANKE, L., DUBOIS, P. C., MEIN, C. A., DOBSON, R. J., ALBERT, T. J., RODESCH, M. J., CLAYTON, D. G., TODD, J. A., VAN HEEL, D. A. & PLAGNOL, V. 2010. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human Molecular Genetics,* 19**,** 122-134.

HERNÁNDEZ, R., NOMBELA, C., DIEZ-OREJAS, R. & GIL, C. 2004. Two-dimensional reference map of *Candida albicans* hyphal forms. *Proteomics,* 4**,** 374-382.

HICKMAN, M. A., ZENG, G., FORCHE, A., HIRAKAWA, M. P., ABBEY, D., HARRISON, B. D., WANG, Y.-M., SU, C.-H., BENNETT, R. J., WANG, Y. & BERMAN, J. 2013. The 'obligate diploid' *Candida albicans* forms mating-competent haploids. *Nature*, 494, 55-59.

HOBSON, R. P. 2003. The global epidemiology of invasive *Candida* infections—is the tide turning? *Journal of Hospital Infection,* 55**,** 159-168.

HOEHAMER, C. F., CUMMINGS, E. D., HILLIARD, G. M., MORSCHHÄUSER, J. & ROGERS, P. D. 2009. Proteomic analysis of Mrr1p- and Tac1p-associated differential protein expression in azole-resistant clinical isolates of *Candida albicans*. *Proteomics – Clinical Applications,* 3**,** 968-978.

HOEHAMER, C. F., CUMMINGS, E. D., HILLIARD, G. M. & ROGERS, P. D. 2010. Changes in the proteome of *Candida albicans* in response to azole, polyene, and echinocandin antifungal agents. *Antimicrob Agents Chemother,* 54**,** 1655-1664.

HOEPFNER, D., VAN DEN BERG, M., PHILIPPSEN, P., TABAK, H. F. & HETTEMA, E. H. 2001. A role for Vps1p, actin, and the Myo2p motor in peroxisome abundance and inheritance in *Saccharomyces cerevisiae*. *The Journal of Cell Biology,* 155**,** 979.

HOFFMAN, C. S. & WINSTON, F. 1987. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli Gene,* 57**,** 267 - 272.

HOLMES, A. R., TSAO, S., ONG, S. W., LAMPING, E., NIIMI, K., MONK, B. C., NIIMI, M., KANEKO, A., HOLLAND, B. R. & SCHMID, J. 2006. Heterozygosity and functional allelic variation in the *Candida albicans* efflux pump genes *CDR1* and *CDR2*. *Molecular Microbiology,* 62**,** 170-186.

HOMANN, O. R., DEA, J., NOBLE, S. M. & JOHNSON, A. D. 2009. A phenotypic profile of the *Candida albicans* regulatory network. *PLoS Genet,* 5**,** e1000783.

HONG, Z., MANN, P., BROWN, N. H., TRAN, L. E., SHAW, K. J., HARE, R. S. & DIDOMENICO, B. 1994. Cloning and characterization of *KNR4*, a yeast gene involved in (1,3)-beta-glucan synthesis. *Molecular and cellular biology,* 14**,** 1017-1025.

HOPE, W. W., TABERNERO, L., DENNING, D. W. & ANDERSON, M. J. 2004. Molecular mechanisms of primary resistance to flucytosine in *Candida albicans*. *Antimicrobial agents and chemotherapy,* 48**,** 4377-4386.

HOU, C. & CORCES, V. 2012. Throwing transcription for a loop: expression of the genome in the 3D nucleus. *Chromosoma***,** 121, 107-116.

HOYER, L. L. 2001. The *ALS* gene family of *Candida albicans*. *Trends in Microbiology,* 9**,** 176-180.

HULL, C. M., RAISNER, R. M. & JOHNSON, A. D. 2000. Evidence for mating of the "asexual" yeast *Candida albicans* in a mammalian host. *Science,* 289**,** 307-310.

IBM 2012. IBM SPSS Statistics. 21 ed. Chicago, Illinois.

INGLIS, D. O., ARNAUD, M. B., BINKLEY, J., SHAH, P., SKRZYPEK, M. S., WYMORE, F., BINKLEY, G., MIYASATO, S. R., SIMISON, M. & SHERLOCK, G. 2012. The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Research,* 40**,** D667-74.

INGOLIA, N. T., GHAEMMAGHAMI, S., NEWMAN, J. R. S. & WEISSMAN, J. S. 2009. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science,* 324**,** 218-223.

IRNIGER, S., EGLI, C., KUENZLER, M. & BRAUS, G. H. 1992. The yeast actin intron contains a cryptic promoter that can be switched on by preventing transcriptional interference. *Nucleic Acids Research,* 20**,** 4733-4739.

IWABE, N. & MIYATA, T. 2001. Overlapping genes in parasitic protist *Giardia lamblia*. *Gene,* 280**,** 163-167.

JACKSON, B. E., MITCHELL, B. M. & WILHELMUS, K. R. 2007. Corneal virulence of *Candida albicans* strains deficient in *Tup1*-regulated genes. *Investigative Ophthalmology & Visual Science,* 48**,** 2535-2539.

JANSEN, R. & GERSTEIN, M. 2000. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Research,* 28**,** 1481-1488.

JARVIS, W. R. 1995. Epidemiology of nosocomial fungal-infections, with emphasis on *Candida* species. *Clinical Infectious Diseases,* 20**,** 1526-1530.

JINEK, M. & DOUDNA, J. A. 2009. A three-dimensional view of the molecular machinery of RNA interference. *Nature,* 457**,** 405-412.

JOHNSON, A. 2003. The biology of mating in *Candida albicans*. *Nature Reviews Microbiology,* 1**,** 106-116.

JOHNSON, Z. I. & CHISHOLM, S. W. 2004. Properties of overlapping genes are conserved across microbial genomes. *Genome Research,* 14**,** 2268-2272.

JONES, T., FEDERSPIEL, N. A., CHIBANA, H., DUNGAN, J., KALMAN, S., MAGEE, B. B., NEWPORT, G., THORSTENSON, Y. R., AGABIAN, N., MAGEE, P. T., DAVIS, R. W. & SCHERER, S. 2004. The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 7329-7334.

KADOSH, D. & JOHNSON, A. D. 2001. *Rfg1*, a protein related to the *Saccharomyces cerevisiae* hypoxicr regulator *Rox1*, controls filamentous growth and virulence in *Candida albicans*. *Molecular and cellular biology,* 21**,** 2496-2505.

KHAN, Z., BLOOM, J. S., AMINI, S., SINGH, M., PERLMAN, D. H., CAUDY, A. A. & KRUGLYAK, L. 2012. Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Mol Syst Biol,* 8, 602.

KIM, J. & SUDBERY, P. 2011. *Candida albicans,* a major human fungal pathogen. *The Journal of Microbiology,* 49**,** 171-177.

KIRSCH, D. R. & WHITNEY, R. R. 1991. Pathogenicity of *Candida albicans* auxotrophic mutants in experimental infections. *Infection and Immunity,* 59**,** 3297-3300.

KRUEGER, F. 2012. ASAP - Allele-specific alignment pipeline. Babraham Institute, Cambridge, UK.

KRYAZHIMSKIY, S. & PLOTKIN, J. B. 2008. The population genetics of dN/dS. *PLoS Genet,* 4**,** e1000304.

KULLBERG, B. J. & FILLER, S. G. 2002. Candidemia. *In:* CALDERONE, R. (ed.) *Candida and candidiasis.* Washington D.C.: ASM Press.

KURAVI, K., NAGOTU, S., KRIKKEN, A. M., SJOLLEMA, K., DECKERS, M., ERDMANN, R., VEENHUIS, M. & VAN DER KLEI, I. J. 2006. Dynamin-related proteins Vps1p and Dnm1p control peroxisome abundance in *Saccharomyces cerevisiae*. *Journal of Cell Science,* 119**,** 3994-4001.

KUSCH, H., ENGELMANN, S., ALBRECHT, D., MORSCHHÄUSER, J. & HECKER, M. 2007. Proteomic analysis of the oxidative stress response in *Candida albicans*. *PROTEOMICS,* 7**,** 686-697.

KYOTO UNIVERSITY BIOINFORMATICS CENTER. 2010. *Multiple Sequence Alignment by CLUSTALW* [Online]. Available: http://align.genome.jp/ [Accessed November - December 2010].

LACHKE, S. A., LOCKHART, S. R., DANIELS, K. J. & SOLL, D. R. 2003. Skin facilitates *Candida albicans* mating. *Infection and Immunity,* 71**,** 4970-4976.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10**,** R25.

LEE, K.-H., JUN, S., HUR, H.-S., RYU, J.-J. & KIM, J. 2005. *Candida albicans* protein analysis during hyphal differentiation using an integrative HA-tagging method. *Biochemical and Biophysical Research Communications,* 337**,** 784-790.

LEE, R. D.-W., SONG, M.-Y. & LEE, J.-K. 2013. Large-scale profiling and identification of potential regulatory mechanisms for allelic gene expression in colorectal cancer cells. *Gene,* 512**,** 16-22.

LEFEBVRE, J. F., VELLO, E., GE, B., MONTGOMERY, S. B., DERMITZAKIS, E. T., PASTINEN, T. & LABUDA, D. 2012. Genotype-based test in mapping *cis*-regulatory variants from allele-specific expression data. *PLoS ONE,* 7**,** e38667.

LEGRAND, M., LEPHART, P., FORCHE, A., MUELLER, F. M. C., WALSH, T., MAGEE, P. T. & MAGEE, B. B. 2004. Homozygosity at the *MTL* locus in clinical strains of *Candida albicans*: karyotypic rearrangements and tetraploid formation. *Molecular Microbiology,* 52**,** 1451-1462.

LEVESQUE, M. J., GINART, P., WEI, Y. & RAJ, A. 2013. Visualizing SNVs to quantify allele-specific expression in single cells. *Nat Meth***,** 10, 865-867.

LEVIN, J. Z., YASSOUR, M., ADICONIS, X., NUSBAUM, C., THOMPSON, D. A., FRIEDMAN, N., GNIRKE, A. & REGEV, A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Meth,* 7**,** 709-715.

LEVY, S., SUTTON, G., NG, P. C., FEUK, L., HALPERN, A. L., WALENZ, B. P., AXELROD, N., HUANG, J., KIRKNESS, E. F., DENISOV, G., LIN, Y., MACDONALD, J. R., PANG, A. W. C., SHAGO, M., STOCKWELL, T. B., TSIAMOURI, A., BAFNA, V., BANSAL, V., KRAVITZ, S. A., BUSAM, D. A., BEESON, K. Y., MCINTOSH, T. C., REMINGTON, K. A., ABRIL, J. F., GILL, J., BORMAN, J., ROGERS, Y.-H., FRAZIER, M. E., SCHERER, S. W., STRAUSBERG, R. L. & VENTER, J. C. 2007. The diploid genome sequence of an individual human. *PLoS Biol,* 5**,** e254.

LI, G., BAHN, J. H., LEE, J.-H., PENG, G., CHEN, Z., NELSON, S. F. & XIAO, X. 2012. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, 40, e104.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transformation. *Bioinformatics,* 25**,** 1754-1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

LIANG, R. M., YONG, X. L., JIANG, Y. P., TAN, Y. H., DAI, B. D., WANG, S. H., HU, T. T., CHEN, X., LI, N., DONG, Z. H., HUANG, X. C., CHEN, J., CAO, Y. B. & JIANG, Y. Y. 2011. 2-Amino-nonyl-6-methoxyl-tetralin muriate activity against *Candida albicans* augments endogenous reactive

oxygen species production – a microarray analysis study. *FEBS Journal,* 278**,** 1075-1085.

LISTER, R., O'MALLEY, R. C., TONTI-FILIPPINI, J., GREGORY, B. D., BERRY, C. C., MILLAR, A. H. & ECKER, J. R. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell,* 133**,** 523-536.

LIU, T. T., ZNAIDI, S., BARKER, K. S., XU, L., HOMAYOUNI, R., SAIDANE, S., MORSCHHÄUSER, J., NANTEL, A., RAYMOND, M. & ROGERS, P. D. 2007. Genome-wide expression and location analyses of the *Candida albicans* Tac1p regulon. *Eukaryotic Cell,* 6**,** 2122-2138.

LO, H. S., WANG, Z., HU, Y., YANG, H. H., GERE, S., BUETOW, K. H. & LEE, M. P. 2003. Allelic variation in gene expression is common in the human genome. *Genome Research,* 13**,** 1855-1862.

LOCKHART, S. R., DANIELS, K. J., ZHAO, R., WESSELS, D. & SOLL, D. R. 2003. Cell biology of mating in *Candida albicans*. *Eukaryotic Cell,* 2**,** 49-61.

LONFAT, N., MONTAVON, T., JEBB, D., TSCHOPP, P., NGUYEN HUYNH, T. H., ZAKANY, J. & DUBOULE, D. 2013. Transgene- and locus-dependent imprinting reveals allele-specific chromosome conformations. *Proceedings of the National Academy of Sciences*, 16, 11946-11951.

LORENZ, M. C., BENDER, J. A. & FINK, G. R. 2004. Transcriptional response of *Candida albicans* upon internalization by macrophages. *Eukaryotic Cell,* 3**,** 1076-1087.

LYON, M. F. 1961. Gene action in the X-chromosome of the mouse (*Mus musculus L.*). *Nature,* 190**,** 372-373.

MACCALLUM, D. M., CASTILLO, L., NATHER, K., MUNRO, C. A., BROWN, A. J. P., GOW, N. A. R. & ODDS, F. C. 2009. Property differences among the four major *Candida albicans* strain clades. *Eukaryotic Cell***,** 8 373-387.

MACISAAC, J. L., BOGUTZ, A. B., MORRISSY, A. S. & LEFEBVRE, L. 2011. Tissue-specific alternative polyadenylation at the imprinted gene *Mest* regulates allelic usage at *Copg2*. *Nucleic Acids Research*, 40, 1523-1535.

MAGEE, B. B. & MAGEE, P. T. 2000. Induction of mating in *Candida albicans* by construction of *MTL*a and *MTL*α strains. *Science,* 289**,** 310-313.

MAIN, B., BICKEL, R., MCINTYRE, L., GRAZE, R., CALABRESE, P. & NUZHDIN, S. 2009. Allele-specific expression assays using Solexa. *BMC Genomics,* 10**,** 422.

MARÍN, A., GALLARDO, M., KATO, Y., SHIRAHIGE, K., GUTIÉRREZ, G., OHTA, K. & AGUILERA, A. 2003. Relationship between G+C content, ORF-length and mRNA concentration in *Saccharomyces cerevisiae*. *Yeast,* 20**,** 703-711.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. & GILAD, Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research,* 18**,** 1509-1517.

MARKOVICH, S., YEKUTIEL, A., SHALIT, I., SHADKCHAN, Y. & OSHEROV, N. 2004. Genomic approach to identification of mutations affecting caspofungin susceptibility in *Saccharomyces cerevisiae*. *Antimicrobial agents and chemotherapy,* 48**,** 3871-3876.

MARTÍNEZ, A. I., CASTILLO, L., GARCERÁ, A., ELORZA, M. V., VALENTÍN, E. & SENTANDREU, R. 2004. Role of *PIR1* in the construction of the *Candida albicans* cell wall. *Microbiology,* 150**,** 3151-3161.

MARTTILA, E., BOWYER, P., SANGLARD, D., UITTAMO, J., KAIHOVAARA, P., SALASPURO, M., RICHARDSON, M. & RAUTEMAA, R. 2013. Fermentative 2-carbon metabolism produces carcinogenic levels of acetaldehyde in *Candida albicans*. *Molecular Oral Microbiology,* 28**,** 281-291.

MASSEY, S. E., MOURA, G., BELTRAO, P., ALMEIDA, R., GAREY, J. R., TUITE, M. F. & SANTOS, M. A. 2003. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in *Candida* spp. *Genome Res,* 13**,** 544-557.

MAVRICH, T. N., IOSHIKHES, I. P., VENTERS, B. J., JIANG, C., TOMSHO, L. P., QI, J., SCHUSTER, S. C., ALBERT, I. & PUGH, B. F. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research,* 18**,** 1073-1083.

MBOUP, M., FISCHER, I., LAINER, H. & STEPHAN, W. 2012. *Trans*-species polymorphism and allele-specific expression in the CBF gene family of wild tomatoes. *Molecular Biology and Evolution,* 29**,** 3641-3652.

MCCULLOUGH, M. J., CLEMONS, K. V. & STEVENS, D. A. 1999. Molecular and phenotypic characterization of genotypic *Candida albicans* subgroups and comparison with *Candida dubliniensis* and *Candida stellatoidea*. *Journal of Clinical Microbiology,* 37**,** 417-421.

MCINTYRE, L., LOPIANO, K., MORSE, A., AMIN, V., OBERG, A., YOUNG, L. & NUZHDIN, S. 2011. RNA-seq: technical variability and sampling. *BMC Genomics,* 12**,** 293.

MCMANUS, C. J., MAY, G. E., SPEALMAN, P. & SHTEYMAN, A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Research,* 24**,** 422-430.

MEACHAM, F., BOFFELLI, D., DHAHBI, J., MARTIN, D., SINGER, M. & PACHTER, L. 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics,* 12**,** 451.

MERKL, R. 2003. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *Journal of Molecular Evolution,* 57**,** 453-466.

MILLER, L. G., HAJJEH, R. A. & EDWARDS, J. E. 2001. Estimating the cost of nosocomial candidemia in the United States. *Clinical Infectious Diseases,* 32**,** 1110.

MILNE, S. W., CHEETHAM, J., LLOYD, D., AVES, S. & BATES, S. 2011. Cassettes for PCR-mediated gene tagging in *Candida albicans* utilizing nourseothricin resistance. *Yeast,* 28**,** 833-841.

MINOCHE, A., DOHM, J. & HIMMELBAUER, H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology,* 12**,** R112.

MIRAMÓN, P., DUNKER, C., WINDECKER, H., BOHOVYCH, I. M., BROWN, A. J. P., KURZAI, O. & HUBE, B. 2012. Cellular responses of *Candida albicans* to phagocytosis and the extracellular activities of neutrophils are critical to counteract carbohydrate starvation, cxidative and nitrosative stress. *PLoS ONE,* 7**,** e52850.

MISHRA, P. K., BAUM, M. & CARBON, J. 2011. DNA methylation regulates phenotype-dependent transcriptional activity in *Candida albicans*. *Proceedings of the National Academy of Sciences,* 108**,** 11965-11970.

MIYAJIMA, N., HORIUCHI, R., SHIBUYA, Y., FUKUSHIGE, S.-I., MATSUBARA, K.-I., TOYOSHIMA, K. & YAMAMOTO, T. 1989. Two

erbA homologs encoding proteins with different T3 binding capacities are transcribed from opposite DNA strands of the same genetic locus. *Cell,* 57**,** 31-39.

MOAZENI, M., KHORAMIZADEH, M. R., KORDBACHEH, P., SEPEHRIZADEH, Z., ZERAATI, H., NOORBAKHSH, F., TEIMOORI-TOOLABIM, L. & REZAIE, S. 2012. RNA-mediated gene silencing in *Candida albicans*: Inhibition of hyphae formation by use of RNAi technology. *Mycopathologia,* 174**,** 177-185.

MONOD, M., TOGNI, G., HUBE, B. & SANGLARD, D. 1994. Multiplicity of genes encoding secreted aspartic proteinases in *Candida* species. *Molecular Microbiology,* 13**,** 357-368.

MORISON, I. M., RAMSAY, J. P. & SPENCER, H. G. 2005. A census of mammalian imprinting. *Trends in Genetics,* 21**,** 457-465.

MORSCHHÄUSER, J., MICHEL, S. & STAIB, P. 1999. Sequential gene disruption in *Candida albicans* by FLP-mediated site-specific recombination. *Molecular Microbiology,* 32**,** 547-556.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth,* 5**,** 621-628.

MORTON, B. R. 1993. Chloroplast DNA codon use: Evidence for selection at the *psb A* locus based on tRNA availability. *Journal of molecular evolution,* 37**,** 273-280.

MUKHERJEE, P. K., CHANDRA, J., KUHN, D. M. & GHANNOUM, M. A. 2003. Mechanism of fluconazole resistance in *Candida albicans* biofilms: Phase-specific role of efflux pumps and membrane sterols. *Infection and Immunity,* 71**,** 4333-4340.

MULLER, H. 1930. Types of visible variations induced by X-rays in *Drosophila*. *Journal of Genetics,* 22**,** 299-334.

MURAD, A. M. A., LEE, P. R., BROADBENT, I. D., BARELLE, C. J. & BROWN, A. J. P. 2000. CIp10, an efficient and convenient integrating vector for *Candida albicans*. *Yeast,* 16**,** 325-327.

MUZZEY, D., SCHWARTZ, K., WEISSMAN, J. & SHERLOCK, G. 2013. Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biology,* 14**,** R97.

MUZZEY, D., SHERLOCK, G. & WEISSMAN, J. S. 2014. Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Research*, 24, 963-973.

NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M. & SNYDER, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science,* 320**,** 1344-1349.

NAKAMURA, K., OSHIMA, T., MORIMOTO, T., IKEDA, S., YOSHIKAWA, H., SHIWA, Y., ISHIKAWA, S., LINAK, M. C., HIRAI, A., TAKAHASHI, H., ALTAF-UL-AMIN, M., OGASAWARA, N. & KANAYA, S. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research,* 39**,** e90.

NANNAPANENI, S., WANG, D., JAIN, S., SCHROEDER, B., HIGHFILL, C., REUSTLE, L., PITTSLEY, D., MAYSENT, A., MOULDER, S., MCDOWELL, R. & KIM, K. 2010. The yeast dynamin-like protein Vps1:vps1 mutations perturb the internalization and the motility of endocytic vesicles and endosomes via disorganization of the actin cytoskeleton. *European Journal of Cell Biology,* 89**,** 499-508.

NANTEL, A., DIGNARD, D., BACHEWICH, C., HARCUS, D., MARCIL, A., BOUIN, A.-P., SENSEN, C. W., HOGUES, H., VAN HET HOOG, M., GORDON, P., RIGBY, T., BENOIT, F., TESSIER, D. C., THOMAS, D. Y. & WHITEWAY, M. 2002. Transcription profiling of *Candida* albicans cells undergoing the yeast-to-hyphal transition. *Molecular biology of the cell,* 13**,** 3452-3465.

NETT, J. E., SANCHEZ, H., CAIN, M. T., ROSS, K. M. & ANDES, D. R. 2011. Interface of *Candida albicans* biofilm matrix-associated drug resistance and cell wall integrity regulation. *Eukaryotic Cell,* 10**,** 1660-1669.

NEWELL-PRICE, J., CLARK, A. J. L. & KING, P. 2000. DNA methylation and silencing of gene expression. *Trends in Endocrinology & Metabolism,* 11**,** 142-148.

NGUYEN, M., PEACOCK, J. E. & JR TANNER, D. C. 1995. Therapeutic approaches in patients with candidemia: Evaluation in a multicenter, prospective, observational study. *Archives of Internal Medicine,* 155**,** 2429-2435.

NIELSEN, R. & YANG, Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Molecular Biology and Evolution,* 20**,** 1231-1239.

NOBILE, CLARISSA J., FOX, EMILY P., NETT, JENIEL E., SORRELLS, TREVOR R., MITROVICH, QUINN M., HERNDAY, AARON D., TUCH, BRIAN B., ANDES, DAVID R. & JOHNSON, ALEXANDER D. 2012. A recently evolved transcriptional network controls biofilm development in *Candida albicans*. *Cell,* 148**,** 126-138.

NOBILE, C. J. & MITCHELL, A. P. 2005. Regulation of cell-surface genes and biofilm formation by the *C. albicans* transcription factor Bcr1p. *Current Biology,* 15**,** 1150-1155.

NOBLE, S. M., FRENCH, S., KOHN, L. A., CHEN, V. & JOHNSON, A. D. 2010. Systematic screens of a *Candida albicans* homozygous deletion library decouple morphogenetic switching and pathogenicity. *Nat Genet,* 42**,** 590-598.

NOBLE, S. M. & JOHNSON, A. D. 2005. Strains and strategies for large-scale gene deletion studies of the diploid human fungal pathogen *Candida albicans*. *Eukaryotic Cell,* 4**,** 298-309.

NOTHWEHR, S., BRYANT, N. & STEVENS, T. 1996. The newly identified yeast *GRD* genes are required for retention of late- Golgi membrane proteins. *Mol. Cell. Biol.,* 16**,** 2700-2707.

ODDS, F. C. 1993. Resistance of yeasts to azole-derivative antifungals. *Journal of Antimicrobial Chemotherapy,* 31**,** 463-471.

ODDS, F. C., BOUGNOUX, M.-E., SHAW, D. J., BAIN, J. M., DAVIDSON, A. D., DIOGO, D., JACOBSEN, M. D., LECOMTE, M., LI, S.-Y., TAVANTI, A., MAIDEN, M. C. J., GOW, N. A. R. & D'ENFERT, C. 2007. Molecular phylogenetics of *Candida albicans*. *Eukaryotic Cell,* 6**,** 1041-1052.

ODDS, F. C., BROWN, A. J. P. & GOW, N. A. R. 2004. *Candida albicans* genome sequence: a platform for genomics in the absence of genetics. *Genome Biology,* 5**,** 3.

ODDS, F. C. & JACOBSEN, M. D. 2008. Multilocus sequence typing of pathogenic *Candida* Species. *Eukaryotic Cell,* 7**,** 1075-1084.

ODDS, F. C. & WEBSTER, C. E. 1988. Effects of azole antifungals in vitro on host/parasite interactions relevant to C*andida* infections. *Journal of Antimicrobial Chemotherapy,* 22**,** 473-481.

OHAMA, T., SUZUKI, T., MORI, M., OSAWA, S., UEDA, T., WATANABE, K. & NAKASE, T. 1993. Non-universal decoding of the leucine codon CUG in several *Candida* species. *Nucleic Acids Res,* 21**,** 4039-4045.

OHLSSON, R., TYCKO, B. & SAPIENZA, C. 1998. Monoallelic expression: `there can only be one'. *Trends in Genetics,* 14**,** 435-438.

OSBORNE, J. W. 2005. Notes on the use of data transformations. *Notes,* 9**,** 42.

PALIGE, K., LINDE, J., MARTIN, R., BÖTTCHER, B., CITIULO, F., SULLIVAN, D. J., WEBER, J., STAIB, C., RUPP, S., HUBE, B., MORSCHHÄUSER, J. & STAIB, P. 2013. Global transcriptome sequencing identifies chlamydospore specific markers in *Candida albicans* and *Candida dubliniensis*. *PLoS ONE,* 8**,** e61940.

PANDEY, R. V., FRANSSEN, S. U., FUTSCHIK, A. & SCHLÖTTERER, C. 2013. Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular Ecology Resources***,** 13, 740-745.

PASTINEN, T., GE, B. & HUDSON, T. J. 2006. Influence of human genome polymorphism on gene expression. *Human Molecular Genetics,* 15**,** R9-R16.

PASTINEN, T. & HUDSON, T. J. 2004. *Cis*-acting regulatory variation in the human genome. *Science,* 306**,** 647-650.

PEREIRA, J. P., GIRARD, R., CHABY, R., CUMANO, A. & VIEIRA, P. 2003. Monoallelic expression of the murine gene encoding Toll-like receptor 4. *Nat Immunol,* 4**,** 464-470.

PESTOV, D. G., STOCKELMAN, M. G., STREZOSKA, Z. & LAU, L. F. 2001. *ERB1*, the yeast homolog of mammalian *Bop1*, is an essential gene required for maturation of the 25S and 5.8S ribosomal RNAs. *Nucleic Acids Research,* 29**,** 3621-3630.

PETERS, C., BAARS, T. L., BÜHLER, S. & MAYER, A. 2004. Mutual control of membrane fission and fusion proteins. *Cell,* 119**,** 667-678.

PFALLER, M. A. & DIEKEMA, D. J. 2007. Epidemiology of Invasive *Candidiasis*: a Persistent Public Health Problem. *Clinical Microbiology Reviews,* 20**,** 133-163.

PIDSLEY, R., DEMPSTER, E., TROAKES, C., AL-SARRAJ, S. & MILL, J. 2012. Epigenetic and genetic variation at the *IGF2/H19* imprinting control

region on 11p15.5 is associated with cerebellum weight. *Epigenetics,* 7**,** 155-163.

PORMAN, A. M., ALBY, K., HIRAKAWA, M. P. & BENNETT, R. J. 2011. Discovery of a phenotypic switch regulating sexual mating in the opportunistic fungal pathogen *Candida tropicalis*. *Proceedings of the National Academy of Sciences,* 108**,** 21158-21163.

PRATT, J. M., SIMPSON, D. M., DOHERTY, M. K., RIVERS, J., GASKELL, S. J. & BEYNON, R. J. 2006. Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protocols,* 1**,** 1029-1043.

PUJOL, C., JOLY, S., LOCKHART, S. R., NOEL, S., TIBAYRENC, M. & SOLL, D. R. 1997. Parity among the randomly amplified polymorphic DNA method, multilocus enzyme electrophoresis, and Southern blot hybridization with the moderately repetitive DNA probe Ca3 for fingerprinting *Candida albicans*. *Journal of Clinical Microbiology,* 35**,** 2348-2358.

PUJOL, C., PFALLER, M. & SOLL, D. R. 2002. Ca3 fingerprinting of *Candida albicans* bloodstream isolates from the United States, Canada, South America, and Europe reveals a European clade. *Journal of Clinical Microbiology,* 40**,** 2729-2740.

PUJOL, C., PFALLER, M. & SOLL, D. R. 2004. Flucytosine resistance is restricted to a single genetic clade of *Candida albicans*. *Antimicrobial agents and chemotherapy,* 48**,** 262-266.

QUINN, A., JUNEJA, P. & JIGGINS, F. M. 2014. Estimates of allele-specific expression in *Drosophila* with a single genome sequence and RNA-seq data. *Bioinformatics*.

RAJ, A. & VAN OUDENAARDEN, A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell,* 135**,** 216-226.

RAMAGE, G., TOMSETT, K., WICKES, B. L., LÓPEZ-RIBOT, J. L. & REDDING, S. W. 2004. Denture stomatitis: a role for *Candida* biofilms. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology,* 98**,** 53-59.

RAMAGE, G., VANDEWALLE, K., LOPEZ-RIBOT, J. L. & WICKES, B. L. 2002. The filamentation pathway controlled by the Efg1 regulator protein is

required for normal biofilm formation and development in *Candida albicans*. *FEMS Microbiology Letters,* 214**,** 95-100.

RAMAGE, G., VANDEWALLE, K., WICKES, B. L. & LOPEZ-RIBOT, J. L. 2001. Characteristics of biofilm formation by *Candida albicans*. *Rev Iberoam Micol,* 18**,** 163 - 170.

REUß, O., VIK, Å., KOLTER, R. & MORSCHHÄUSER, J. 2004. The *SAT1* flipper, an optimized tool for gene disruption in *Candida albicans*. *Gene,* 341**,** 119-127.

REX, J. H., RINALDI, M. G. & PFALLER, M. 1995. Resistance of *Candida* species to fluconazole. *Antimicrobial agents and chemotherapy,* 39**,** 1-8.

RICHMOND, T. J. & DAVEY, C. A. 2003. The structure of DNA in the nucleosome core. *Nature,* 423**,** 145-150.

ROBBINS, N., UPPULURI, P., NETT, J., RAJENDRAN, R., RAMAGE, G., LOPEZ-RIBOT, J. L., ANDES, D. & COWEN, L. E. 2011. *Hsp90* governs dispersion and drug resistance of fungal biofilms. *PLoS Pathog,* 7**,** e1002257.

ROBINSON, M., MCCARTHY, D. & SMYTH, G. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-140.

ROBINSON, M. & OSHLACK, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology,* 11**,** R25.

ROOIJ, I. I. S.-D., ALLWOOD, E. G., AGHAMOHAMMADZADEH, S., HETTEMA, E. H., GOLDBERG, M. W. & AYSCOUGH, K. R. 2010. A role for the dynamin-like protein Vps1 during endocytosis in yeast. *Journal of Cell Science,* 123**,** 3496-3506.

ROSE, J. & EISENMENGER, F. 1991. A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. *Journal of molecular evolution,* 32**,** 340-354.

ROZOWSKY, J., ABYZOV, A., WANG, J., ALVES, P., RAHA, D., HARMANCI, A., LENG, J., BJORNSON, R., KONG, Y., KITABAYASHI, N., BHARDWAJ, N., RUBIN, M., SNYDER, M. & GERSTEIN, M. 2011. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol,* 7**,** 522.

RUHNKE, M. 2002. Skin and Mucous Membrane Infections. *In:* CALDERONE, R. (ed.) *Candida and Candidiasis.* Washington D.C.: ASM Press.

SABATINOS, S. & FORSBURG, S. 2009. Measuring DNA content by flow cytometry in fission yeast. *In:* VENGROVA, S. & DALGAARD, J. Z. (eds.) *DNA Replication.* Humana Press.

SALICHOS, L. & ROKAS, A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature,* 497**,** 327-331.

SANGLARD, D., KUCHLER, K., ISCHER, F., PAGANI, J. L., MONOD, M. & BILLE, J. 1995. Mechanisms of resistance to azole antifungal agents in *Candida albicans* isolates from AIDS patients involve specific multidrug transporters. *Antimicrobial agents and chemotherapy,* 39**,** 2378-2386.

SANTIAGO, T. C., PURVIS, I. J., BETTANY, A. J. E. & BROWN, A. J. P. 1986. The relationship between mRNA stability and length in *Saccharomyces cerevisiae*. *Nucleic Acids Research,* 14**,** 8347-8360.

SANZ, M., VALLE, R. & RONCERO, C. 2007. Promoter heterozygosity at the *Candida albicans CHS7* gene is translated into differential expression between alleles. *FEMS Yeast Research,* 7**,** 993-1003.

SAVOVA, V., VIGNEAU, S. & GIMELBRANT, A. A. 2013. Autosomal monoallelic expression: genetics of epigenetic diversity? *Current Opinion in Genetics & Development*, 23, 642-648.

SCANNELL, D. R., BUTLER, G. & WOLFE, K. H. 2007. Yeast genome evolution—the origin of the species. *Yeast,* 24**,** 929-942.

SCHERBAKOV, D. & GARBER, M. 2000. Overlapping genes in bacterial and phage genomes. *Molecular Biology,* 34**,** 485-495.

SCHERER, S. & STEVENS, D. A. 1987. Application of DNA typing methods to epidemiology and taxonomy of *Candida* species. *Journal of Clinical Microbiology,* 25**,** 675-679.

SCHERF, A., HERNANDEZ-RIVAS, R., BUFFET, P., BOTTIUS, E., BENATAR, C., POUVELLE, B., GYSIN, J. & LANZER, M. 1998. Antigenic variation in malaria: *in situ* switching, relaxed and mutually exclusive transcription of *var* genes during intra-erythrocytic development in *Plasmodium falciparum*. *The EMBO Journal,* 17**,** 5419-5426.

SEGAL, E., FONDUFE-MITTENDORF, Y., CHEN, L., THASTROM, A., FIELD, Y., MOORE, I. K., WANG, J.-P. Z. & WIDOM, J. 2006. A genomic code for nucleosome positioning. *Nature,* 442**,** 772-778.

SELLAM, A., HOGUES, H., ASKEW, C., TEBBJI, F., VAN HET HOOG, M., LAVOIE, H., KUMAMOTO, C., WHITEWAY, M. & NANTEL, A. 2010. Experimental annotation of the human pathogen *Candida albicans* coding and noncoding transcribed regions using high-resolution tiling arrays. *Genome Biology,* 11**,** R71.

SERRE, D., GURD, S., GE, B., SLADEK, R., SINNETT, D., HARMSEN, E., BIBIKOVA, M., CHUDIN, E., BARKER, D. L. & DICKINSON, T. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. *PLoS genetics,* 4**,** e1000006.

SETIADI, E. R., DOEDT, T., COTTIER, F., NOFFZ, C. & ERNST, J. F. 2006. Transcriptional response of *Candida albicans* to hypoxia: Linkage of oxygen sensing and Efg1p-regulatory networks. *Journal of Molecular Biology,* 361**,** 399-411.

SHARP, P. M. & LI, W. H. 1987. The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research,* 15**,** 1281.

SHEN, J. Q., GUO, W. H. & KOHLER, J. R. 2005. *CaNAT1*, a heterologous dominant selectable marker for transformation of candida albicans and other pathogenic *Candida* species. *Infection and Immunity,* 73**,** 1239-1242.

SHINDE, R., CHAUHAN, N., RAUT, J. & KARUPPAYIL, S. 2012. Sensitization of *Candida albicans* biofilms to various antifungal drugs by cyclosporine A. *Annals of Clinical Microbiology and Antimicrobials,* 11**,** 27.

SHINTANI, S., O'HUIGIN, C., TOYOSAWA, S., MICHALOVA, V. & KLEIN, J. 1999. Origin of gene overlap: the case of *TCP1* and *ACAT2*. *Genetics,* 152**,** 743-754.

SILVA, L. V., SANGUINETTI, M., VANDEPUTTE, P., TORELLI, R., ROCHAT, B. & SANGLARD, D. 2013. Milbemycins: more than efflux inhibitors for fungal pathogens. *Antimicrobial agents and chemotherapy,* 57**,** 873-886.

SIMON, M. D., PINTER, S. F., FANG, R., SARMA, K., RUTENBERG-SCHOENBERG, M., BOWMAN, S. K., KESNER, B. A., MAIER, V. K., KINGSTON, R. E. & LEE, J. T. 2013. High-resolution *Xist* binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, 504, 465-469.

SINGH, P., WU, X., LEE, D.-H., LI, A. X., RAUCH, T. A., PFEIFER, G. P., MANN, J. R. & SZABO, P. E. 2011. Chromosome-wide analysis of parental allele-specific chromatin and DNA methylation. *Mol. Cell. Biol.***,** MCB.00961-10.

SKELLY, D. A., JOHANSSON, M., MADEOY, J., WAKEFIELD, J. & AKEY, J. M. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research,* 21**,** 1728-1737.

SLUTSKY, B., STAEBELL, M., ANDERSON, J., RISEN, L., PFALLER, M. & SOLL, D. R. 1987. "White-opaque transition": a second high-frequency switching system in *Candida albicans*. *J. Bacteriol.,* 169**,** 189-197.

SOLL, D. R. 2002. *Candida* commensalism and virulence: the evolution of phenotypic plasticity. *Acta Tropica,* 81**,** 101-110.

SONG, M.-Y., KIM, H.-E., KIM, S., CHOI, I.-H. & LEE, J.-K. 2011. SNP-based large-scale identification of allele-specific gene expression in human B cells. *Gene*, 493, 211-218.

SOUTHERN, E. M. 1975. Detection of specific sequences among DNA fragments seperated by gel electrophoresis. *Journal of Molecular Biology,* 98**,** 503-517.

STAIB, P., KRETSCHMAR, M., NICHTERLEIN, T., HOF, H. & MORSCHHÄUSER, J. 2002. Host versus *in vitro* signals and intrastrain allelic differences in the expression of a *Candida albicans* virulence gene. *Molecular Microbiology,* 44**,** 1351-1366.

STAIB, P. & MORSCHHÄUSER, J. 2007. Chlamydospore formation in *Candida albicans* and *Candida dubliniensis*– an enigmatic developmental programme. *Mycoses,* 50**,** 1-12.

STEVENSON, K., COOLON, J. & WITTKOPP, P. 2013. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics,* 14**,** 536.

SUN, C., SOUTHARD, C., WITONSKY, D. B., KITTLER, R. & DI RIENZO, A. 2010. Allele-specific down-regulation of *RPTOR* expression induced by retinoids contributes to climate adaptations. *PLoS Genet,* 6**,** e1001178.

SUNG, H.-M., WANG, T.-Y., WANG, D., HUANG, Y.-S., WU, J.-P., TSAI, H.-K., TZENG, J., HUANG, C.-J., LEE, Y.-C., YANG, P., HSU, J., CHANG, T., CHO, C.-Y., WENG, L.-C., LEE, T.-C., CHANG, T.-H., LI, W.-H. & SHIH,

M.-C. 2009. Roles of *trans* and *cis* variation in yeast intraspecies evolution of gene expression. *Molecular Biology and Evolution,* 26**,** 2533-2538.

SUPEK, F. & VLAHOVI EK, K. 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics,* 20**,** 2329.

TANIGUCHI, Y., CHOI, P. J., LI, G.-W., CHEN, H., BABU, M., HEARN, J., EMILI, A. & XIE, X. S. 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science,* 329**,** 533-538.

TEARE, M. D., PINYAKORN, S., HEIGHWAY, J. & SANTIBANEZ KOREF, M. F. 2011. Comparing methods for mapping *cis* acting polymorphisms using allelic expression ratios. *PLoS ONE,* 6**,** e28636.

TENG, X., LIU, J. Y., LI, D., FANG, Y., WANG, X. Y., MA, Y. X., CHEN, S. J., ZHAO, Y. X., XU, W. Z. & GU, H. X. 2011. Application of allele-specific RNAi in hepatitis B virus lamivudine resistance. *Journal of Viral Hepatitis,* 18**,** e491-e498.

THE R FOUNDATION FOR STATISTICAL COMPUTING 2010. R: A language and environment for statistical computing. *In:* R DEVELOPMENT CORE TEAM (ed.) 2.12.0 ed. Vienna, Austria.

TIERNEY, L., LINDE, J., MULLER, S., BRUNKE, S., MOLINA, J. C., HUBE, B., SCHOCK, U., GUTHKE, R. & KUCHLER, K. 2012. An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Frontiers in Microbiology,* 3, 85.

TIROSH, I., REIKHAV, S., LEVY, A. A. & BARKAI, N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science,* 324**,** 659-662.

TSANG, P. W. K., CAO, B., SIU, P. Y. L. & WANG, J. 1999. Loss of heterozygosity, by mitotic gene conversion and crossing over, causes strain-specific adenine mutants in constitutive diploid *Candida albicans*. *Microbiology-UK,* 145**,** 1623-1629.

TUCH, B. B., LABORDE, R. R., XU, X., GU, J., CHUNG, C. B., MONIGHETTI, C. K., STANLEY, S. J., OLSEN, K. D., KASPERBAUER, J. L., MOORE, E. J., BROOMER, A. J., TAN, R., BRZOSKA, P. M., MULLER, M. W., SIDDIQUI, A. S., ASMANN, Y. W., SUN, Y., KUERSTEN, S., BARKER,

M. A., DE LA VEGA, F. M. & SMITH, D. I. 2010a. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS ONE,* 5**,** e9317.

TUCH, B. B., MITROVICH, Q. M., HOMANN, O. R., HERNDAY, A. D., MONIGHETTI, C. K., DE LA VEGA, F. M. & JOHNSON, A. D. 2010b. The transcriptomes of two heritable cell types illuminate the circuit governing their differentiation. *PLoS Genet,* 6**,** e1001070.

TUNG, J., AKINYI, M. Y., MUTURA, S., ALTMANN, J., WRAY, G. A. & ALBERTS, S. C. 2011. Allele-specific gene expression in a wild nonhuman primate population. *Molecular Ecology,* 20**,** 725-739.

TURRO, E., SU, S.-Y., GONCALVES, A., COIN, L., RICHARDSON, S. & LEWIN, A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology,* 12**,** R13.

UPPULURI, P., NETT, J., HEITMAN, J. & ANDES, D. 2008. Synergistic effect of calcineurin inhibitors and fluconazole against *Candida albicans* biofilms. *Antimicrobial Agents and Chemotherapy,* 52**,** 1127-1132.

URRUTIA, A. O. & HURST, L. D. 2003. The signature of selection mediated by expression on human genes. *Genome Research,* 13**,** 2260-2264.

VAN HET HOOG, M., RAST, T. J., MARTCHENKO, M., GRINDLE, S., DIGNARD, D., HOGUES, H., CUOMO, C., BERRIMAN, M., SCHERER, S. & MAGEE, B. 2007. Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biology,* 8**,** R52.

VAN LEEUWEN, F., WIJSMAN, E. R., KIEFT, R., VAN DER MAREL, G. A., VAN BOOM, J. H. & BORST, P. 1997. Localization of the modified base J in telomeric *VSG* gene expression sites of*Trypanosoma brucei*. *Genes & Development,* 11**,** 3232-3241.

VAN VOORST, F., HOUGHTON-LARSEN, J., JØNSON, L., KIELLAND-BRANDT, M. C. & BRANDT, A. 2006. Genome-wide identification of genes required for growth of *Saccharomyces cerevisiae* under ethanol stress. *Yeast,* 23**,** 351-359.

VASUDEVAN, S. & PELTZ, S. W. 2001. Regulated ARE-mediated mRNA decay in Saccharomyces cerevisiae. *Molecular Cell,* 7**,** 1191-1200.

VEERAMACHANENI, V., MAKALOWSKI, W., GALDZICKI, M., SOOD, R. & MAKALOWSKA, I. 2004. Mammalian overlapping genes: the comparative perspective. *Genome Research,* 14**,** 280-286.

VERMES, A., GUCHELAAR, H.-J. & DANKERT, J. 2000. Flucytosine: a review of its pharmacology, clinical indications, pharmacokinetics, toxicity and drug interactions. *Journal of Antimicrobial Chemotherapy,* 46**,** 171-179.

VERSTEEG, ROGIER, VAN SCHAIK, B. D. C., VAN BATENBURG, M. F., ROOS, M., MONAJEMI, R., CARON, H., BUSSEMAKER, H. J. & VAN KAMPEN, A. H. C. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Research,* 13**,** 1998-2004.

VIAENE, J., TIELS, P., LOGGHE, M., DEWAELE, S., MARTINET, W. & CONTRERAS, R. 2000. *MET15* as a visual selection marker for *Candida albicans*. *Yeast,* 16**,** 1205-1215.

VIDAL, D. O., DE SOUZA, J. E. S., PIRES, L. C., MASOTTI, C., SALIM, A. C. M., COSTA, M. C. F., GALANTE, P. A. F., DE SOUZA, S. J. & CAMARGO, A. A. 2011. Analysis of allelic differential expression in the human genome using allele-specific serial analysis of gene expression tags. *Genome,* 54**,** 120-127.

VIZEACOUMAR, F. J., VREDEN, W. N., FAGARASANU, M., EITZEN, G. A., AITCHISON, J. D. & RACHUBINSKI, R. A. 2006. The dynamin-like protein Vps1p of the yeast *Saccharomyces cerevisiae* associates with peroxisomes in a Pex19p-dependent manner. *Journal of Biological Chemistry,* 281**,** 12817-12823.

WANG, J., WANG, W., LI, R., LI, Y., TIAN, G., GOODMAN, L., FAN, W., ZHANG, J., LI, J., ZHANG, J., GUO, Y., FENG, B., LI, H., LU, Y., FANG, X., LIANG, H., DU, Z., LI, D., ZHAO, Y., HU, Y., YANG, Z., ZHENG, H., HELLMANN, I., INOUYE, M., POOL, J., YI, X., ZHAO, J., DUAN, J., MA, L., LI, G., YANG, Z., ZHANG, G., YANG, B., YU, C., LIANG, F., LI, W., SHAOCHAUN, L., LI, D., NI, P., RUAN, J., LI, Q., ZHU, H., LIU, D., LU, Z., LI, N., GUO, G., ZHANG, J., YE, J., FANG, L., HAO, Q., CHEN, Q., LIANG, Y., SU, Y., SAN, A., PING, C., YANG, S., CHEN, F., LI, L., ZHOU, K., ZHENG, H., REN, Y., YANG, L., GAO, Y., YANG, G., LI, Z., FENG, X., KRISTIANSEN, K., WONG, G. K.-S., NIELSEN, R., DURBIN,

R., BOLUND, L., ZHANG, X., LI, S., YANG, H. & WANG, J. 2008. The diploid genome sequence of an Asian individual. *Nature,* 456**,** 60-65.

WANG, Y., LIU, C. L., STOREY, J. D., TIBSHIRANI, R. J., HERSCHLAG, D. & BROWN, P. O. 2002. Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences,* 99**,** 5860-5865.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WHELAN, W. L. & MAGEE, P. T. 1981. Natural heterozygosity in *Candida albicans*. *Journal of Bacteriology,* 145**,** 896-903.

WHELAN, W. L., PARTRIDGE, R. M. & MAGEE, P. T. 1980. Heterozygosity and segregation in *Candida albicans*. *Molecular and General Genetics MGG,* 180**,** 107-113.

WHITE, T. C. 1997. The presence of an *R467K* amino acid substitution and loss of allelic variation correlate with an azole-resistant lanosterol 14alpha demethylase in *Candida albicans*. *Antimicrobial agents and chemotherapy,* 41**,** 1488.

WILHELM, B. T., MARGUERAT, S., WATT, S., SCHUBERT, F., WOOD, V., GOODHEAD, I., PENKETT, C. J., ROGERS, J. & BAHLER, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature,* 453**,** 1239-1243.

WILSBACH, K. & PAYNE, G. 1993. Vps1p, a member of the dynamin GTPase family, is necessary for Golgi membrane protein retention in *Saccharomyces cerevisiae*. *The EMBO Journal,* 12**,** 3049.

WITTKOPP, P. J., HAERUM, B. K. & CLARK, A. G. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature,* 430**,** 85-88.

WRIGHT, C. J., BURNS, L. H., JACK, A. A., BACK, C. R., DUTTON, L. C., NOBBS, A. H., LAMONT, R. J. & JENKINSON, H. F. 2013. Microbial interactions in building of communities. *Molecular Oral Microbiology,* 28**,** 83-101.

WYRICK, J. J., HOLSTEGE, F. C. P., JENNINGS, E. G., CAUSTON, H. C., SHORE, D., GRUNSTEIN, M., LANDER, E. S. & YOUNG, R. A. 1999. Chromosomal landscape of nuclesome-dependent gene expression and silencing in yeast. *Nature,* 402**,** 418-421.

XU, D., JIANG, B., KETELA, T., LEMIEUX, S., VEILLETTE, K., MARTEL, N., DAVISON, J., SILLAOTS, S., TROSOK, S., BACHEWICH, C., BUSSEY, H., YOUNGMAN, P. & ROEMER, T. 2007. Genome-wide fitness test and mechanism-of-action studies of inhibitory compounds in *Candida albicans*. *PLoS Pathog,* 3**,** e92.

XU, J. P., MITCHELL, T. G. & VILGALYS, R. 1999. PCR-restriction fragment length polymorphism (RFLP) analyses reveal both extensive clonality and local genetic differences in *Candida albicans*. *Molecular Ecology,* 8**,** 59-73.

YAN, H., DOBBIE, Z., GRUBER, S. B., MARKOWITZ, S., ROMANS, K., GIARDIELLO, F. M., KINZLER, K. & VOGELSTEIN, B. 2002a. Small changes in expression affect predisposition to tumorigenesis. *Nature Genetics,* 30**,** 25 - 26.

YAN, H., YUAN, W., VELCULESCU, V. E., VOGELSTEIN, B. & KINZLER, K. W. 2002b. Allelic variation in human gene expression. *Science,* 297**,** 1143.

YANG, Y., GRAZE, R. M., WALTS, B. M., LOPEZ, C. M., BAKER, H. V., WAYNE, M. L., NUZHDIN, S. V. & MCINTYRE, L. M. 2011. Partitioning transcript variation in *Drosophila*: abundance, isoforms, and alleles. *G3: Genes, Genomes, Genetics,* 1**,** 427-436.

YOSHIKAWA, K., TANAKA, T., FURUSAWA, C., NAGAHISA, K., HIRASAWA, T. & SHIMIZU, H. 2009. Comprehensive phenotypic analysis for identification of genes affecting growth under ethanol stress in *Saccharomyces cerevisiae*. *FEMS Yeast Research,* 9**,** 32-44.

YU, L., CASTILLO, L. P. A., MNAIMNEH, S., HUGHES, T. R. & BROWN, G. W. 2006. A survey of essential gene function in the yeast cell division cycle. *Molecular biology of the cell,* 17**,** 4736-4747.

YU, L., ZHAO, J., FENG, J., FANG, J., FENG, C., JIANG, Y., CAO, Y. & JIANG, L. 2010. *Candida albicans CaPTC6* is a functional homologue for *Saccharomyces cerevisiae ScPTC6* and encodes a type 2C protein phosphatase. *Yeast,* 27**,** 197-206.

YU, X. & CAI, M. 2004. The yeast dynamin-related GTPase Vps1p functions in the organization of the actin cytoskeleton via interaction with Sla1p. *Journal of Cell Science,* 117**,** 3839-3853.

ZHAI, R., FENG, Y., ZHAN, X., SHEN, X., WU, W., YU, P., ZHANG, Y., CHEN, D., WANG, H., LIN, Z., CAO, L. & CHENG, S. 2013. Identification of transcriptome SNPs for assessing allele-specific gene expression in a super-hybrid rice Xieyou9308. *PLoS ONE,* 8**,** e60668.

ZHANG, M., ZHAO, H., XIE, S., CHEN, J., XU, Y., WANG, K., ZHAO, H., GUAN, H., HU, X., JIAO, Y., SONG, W. & LAI, J. 2011. Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. *Proceedings of the National Academy of Sciences*, 108, 20042-20047.

ZHANG, T., LI, W., LI, D., WANG, Y. & SANG, J. 2008. Role of *CaECM25* in cell morphogenesis, cell growth and virulence in *Candida albicans*. *Science in China Series C: Life Sciences,* 51**,** 362-372.

ZHANG, Z., HUANG, S., WANG, J., ZHANG, X., PARDO MANUEL DE VILLENA, F., MCMILLAN, L. & WANG, W. 2013. GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference owing to RNA-seq reads misalignment. *Bioinformatics,* 29**,** i291-i299.

ZHANG, Z. H., JHAVERI, D. J., MARSHALL, V. M., BAUER, D. C., EDSON, J., NARAYANAN, R. K., ROBINSON, G. J., LUNDBERG, A. E., BARTLETT, P. F., WRAY, N. R. & ZHAO, Q. 2014. A comparative study of techniques for differential expression analysis on RNA-Seq data. *bioRxiv*.

ZHAO, X., OH, S.-H., YEATER, K. M. & HOYER, L. L. 2005. Analysis of the *Candida albicans* Als2p and Als4p adhesins suggests the potential for compensatory function within the Als family. *Microbiology,* 151**,** 1619-1630.

ZHAO, X., OH, S. H. & HOYER, L. L. 2007. Unequal contribution of *ALS9* alleles to adhesion between *Candida albicans* and human vascular endothelial cells. *Microbiology,* 153**,** 2342.

ZHAO, X. M., PUJOL, C., SOLL, D. R. & HOYER, L. L. 2003. Allelic variation in the contiguous loci encoding *Candida albicans ALS5, ALS1* and *ALS9*. *Microbiology-Sgm,* 149**,** 2947-2960.

ZHENG, W., CHUNG, L. & ZHAO, H. 2011. Bias detection and correction in RNA sequencing data. *BMC Bioinformatics,* 12**,** 290.

ZORDAN, R. E., MILLER, M. G., GALGOCZY, D. J., TUCH, B. B. & JOHNSON, A. D. 2007. Interlocking transcriptional feedback loops control white-opaque switching in *Candida albicans*. *PLoS Biol,* 5**,** e256.

ZÖRGÖ, E., GJUVSLAND, A., CUBILLOS, F. A., LOUIS, E. J., LITI, G., BLOMBERG, A., OMHOLT, S. W. & WARRINGER, J. 2012. Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast. *Molecular Biology and Evolution,* 29**,** 1781-1789.

ZUO, T., WANG, L., MORRISON, C., CHANG, X., ZHANG, H., LI, W., LIU, Y., WANG, Y., LIU, X., CHAN, M. W. Y., LIU, J.-Q., LOVE, R., LIU, C.-G., GODFREY, V., SHEN, R., HUANG, T. H. M., YANG, T., PARK, B. K., WANG, C.-Y., ZHENG, P. & LIU, Y. 2007. *FOXP3* Is an X-Linked breast cancer suppressor gene and an important repressor of the *HER-2/ErbB2* oncogene. *Cell,* 129**,** 1275-1286.