



**Evolution of DNA binding preferences in a family of  
eukaryotic transcription regulators**

**Evolutionäre Entwicklung der Bindeaffinität an bestimmte  
DNA Sequenzen in einer Familie von eukaryotischen  
Transkriptionsfaktoren**

Doctoral thesis for a doctoral degree  
at the Graduate School of Life Sciences,  
Julius-Maximilians-Universität Würzburg,  
Section Infection and Immunity

submitted by

**Valentina del Olmo Toledo**

from

**Barcelona (Spain)**

Würzburg **2019**

Submitted on:

Members of the Doctoral Thesis Committee

Chairperson:

Primary Supervisor: Dr. J. Christian Pérez

Supervisor (Second): Prof. T. Nicolai Siegel, PhD

Supervisor (Third): Prof. Dr. Joachim Morschhäuser

Supervisor (Fourth): Dr. Polly M. Fordyce

Date of Public Defence:

Date of Receipt of Certificates:

*“There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved”*

Charles Darwin (*The origin of species*, 1859)

## **Affidavit**

I hereby confirm that my thesis entitled “Evolution of DNA binding preferences in a family of eukaryotic transcription regulators” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

\_\_\_\_\_  
Place, Date

\_\_\_\_\_  
Signature

## **Eidesstattliche Erklärung**

Hiermit erkläre ich an Eides statt, die Dissertation „Evolutionäre Entwicklung der Bindeaffinität an bestimmte DNA Sequenzen in einer Familie von eukaryotischen Transkriptionsfaktoren“ eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

\_\_\_\_\_  
Ort, Datum

\_\_\_\_\_  
Unterschrift

## Acknowledgments

I would like to start by thanking my primary supervisor Christian Pérez. Thank you for offering me this project that I have ended up loving and for *always* finding the time to discuss and answer all my questions without fail. Thank you for taking into consideration my career interests and for your endless support, I truly appreciate it.

I thank Joachim Morschhäuser for his great input and discussions in everything concerning *C. albicans*; it has been a great pleasure to work closely with such an expert in the field.

I would also like to thank Nicolai Siegel for his expertise in chromatin and DNA, for providing great input and always finding the time to join all the meetings.

A special thank you goes to Polly Fordyce, for hosting me in her lab and giving me the great opportunity to learn and work with MITOMI. Being able to spend time in Stanford and USCF was an incredible experience for me. I am really grateful to all members of her lab that made my stay even more enjoyable. Finally, thank you for your help in writing and reviewing the manuscript and being actively involved until the very end of the publishing process.

I would also like to express my gratitude to the GSLS for their support, training and for letting me take part in their study program as a fellow.

I am infinitely grateful for the colleagues I have had the pleasure of working with in the Pérez lab: Lena, Sanda, Juliane, Marie, Tobias, Philipp and Sergio. Thank you for building the best imaginable lab environment, it has been amazing to work next to you. The regular visitors from the first floor: Fabiön and Gianluca, always ready for tea at three. I would also like to thank Austin and Bernardo for their great company in the office. I feel immensely lucky for knowing you all. I will miss our Friday “scientific discussions” and the lunch breaks where I laughed so hard that my muscles hurt. I have enjoyed every second of your company and will always have you in my heart wherever I go.

I would like to give a huge thanks to Danny for his infinite patience and support, for putting up with me every time I came home stressed out and for proofreading my abstracts and listening to my practice presentations without too many complaints.

Por ultimo, doy las gracias a mi familia, que pese a la distancia, no han dejado de apoyarme ni un segundo durante los últimos cuatro años.

## Summary

Regulation of gene expression by the control of transcription is essential for any cell to adapt to the environment and survive. Transcription regulators, i.e. sequence-specific DNA binding proteins that regulate gene expression, are central elements within the gene networks of most organisms. Transcription regulators are grouped into distinct families based on structural features that determine, to a large extent, the DNA sequence(s) that they can recognise and bind. Less is known, however, about how the DNA binding preferences can diversify within transcription regulator families during evolutionary timescales, and how such diversification can affect the biology of the organism.

In this dissertation I study the SREBP (**s**terol **r**egulatory **e**lement **b**inding **p**rotein) family of transcriptional regulators in yeasts, and in *Candida albicans* in particular, as an experimental system to address these questions. The SREBPs are conserved from fungi to humans and represent a subgroup of basic helix-loop-helix DNA binding proteins. Early chromatin immunoprecipitation experiments with SREBPs from humans and yeasts showed that these proteins bound *in vivo* to the canonical DNA sequence, termed *E-box*, most basic helix-loop-helix proteins bind to. By contrast, most recent analysis carried out with less-studied fungal SREBPs revealed a non-canonical DNA motif to be the most overrepresented sequence in the bound regions.

This study aims to establish the intrinsic DNA binding preferences of key branches of this family and to determine how the divergence in DNA binding affinities originated. To this end, I combined phylogenetic and ancestral reconstruction with extensive biochemical characterisation of key SREBP proteins. The results indicated that while the most-studied SREBPs (in mammals) indeed show preference for the *E-box*, a second branch of the family preferentially binds the *non-E-box*, and a third one is able to bind both sequences with similar affinity. The preference for one or the other DNA sequence is an intrinsic property of each protein because their purified DNA binding domain was sufficient to recapitulate their *in vivo* binding preference. The ancestor that gave rise to these two different types of SREBPs (the branch that binds *E-box* and the one that binds *non-E-box* DNA) appears to be a protein with a broader DNA binding capability that had a slight preference for the non-canonical motif. Thus, the results imply these two branches originated by either enhancing the original ancestral preference for *non-E-box* or tilting it towards the *E-box* DNA and flipping the preference for this sequence.

The main function associated with members of the SREBP family in most eukaryotes is the control of lipid biosynthesis. I have further studied the function of these proteins in the lineage that encompasses the human associated yeast *C. albicans*. Strikingly, the three SREBPs present in the fungus' genome contribute to the colonisation of the mammalian gut by regulating cellular processes unrelated to lipid metabolism. Here I describe that two of the three *C. albicans* SREBPs form a regulatory cascade that regulates morphology and cell wall modifications under anaerobic conditions, whereas the third SREBP has been shown to be involved in the regulation of glycolysis genes.

Therefore, I posit that the described diversification in DNA binding specificity in these proteins and the concomitant expansion of targets of regulation were key in enabling this fungal lineage to associate with animals.

## Zusammenfassung

Für jede Zelle ist es essenziell die Transkription über die Genexpression zu regulieren, um sich an unterschiedliche Lebensbedingungen anzupassen. Regulatoren der Transkription, zum Beispiel sequenzspezifische DNA-bindende Proteine, sind ein zentrales Element des Genregulationsnetzwerks in den meisten Organismen. Auf Grund ihres Aufbaus sowie der daraus resultierenden spezifischen Eigenschaften DNA zu binden, werden diese Regulatoren in unterschiedliche Familien unterteilt. Bisher ist wenig darüber bekannt, wie unterschiedlich die DNA Sequenzen sein können, welche von einer Familie von Transkriptionsregulatoren gebunden werden, wie sich diese Diversität der Bindung in der Evolution über die Zeit verändert hat und ob diese unterschiedlichen Bindeaffinitäten die Biologie eines Organismus beeinflussen.

In dieser Dissertation befaße ich mich mit der Transkriptionsregulator Familie der SREBPs (**sterol regulatory element binding protein**) in Hefen, als Modelorganismus diente dabei *Candida albicans*. Die Familie der SREBPs ist vom Pilz zu den Menschen genetisch weitestgehend konserviert und repräsentiert eine Unterfamilie der Helix-loop-helix DNA-bindende Proteine. Erste Chromatin-Immunpräzipitation Experimente der SREBPs in Menschen und Hefen zeigen *in vivo* eine Bindung an eine kanonische DNA Sequenz genannt E-box, welche von den meisten der Helix-loop-helix Proteine gebunden wird. Im Gegensatz zeigen neuere Analysen, welche mit weniger bekannten SREBPs aus Pilzen durchgeführt wurden, dass hauptsächlich nicht-kanonische DNA Sequenzen gebunden werden.

Diese Arbeit versucht die Präferenzen, mit welchen einige der wichtigsten Mitglieder der Familie der SREBPs an bestimmte DNA Sequenzen binden aufzudecken und heraus zu finden wie es innerhalb dieser Gruppe zu unterschiedlichen Bindungsaffinitäten kam. Dafür wurden phylogenetische Rekonstruktionsanalysen und aufwändige biochemische Charakterisierungen einiger der Proteine der SREBP Familie durchgeführt. Die Ergebnisse zeigen, dass die meisten der bisher charakterisierten SREBPs (in Säugetieren) es vorziehen an die E-box Sequenz zu binden, ein anderer Zweig des SREBP Familienstammbaums bevorzugt hingegen die non-E-box Sequenz, ein dritter Zweig des Stammbaums ist in der Lage beide Sequenzen mit gleicher Affinität zu binden. Das Bevorzugen einer der beiden DNA Sequenzen ist eine natürliche Eigenschaft des jeweiligen Proteins, da in Experimenten die isolierte DNA-bindende Domäne der Proteine



ausreichend war, um die *in vivo* Bindepräferenzen zu replizieren. Der Ursprung dieser beiden Gruppen (der E-box bindenden Gruppe und der Gruppe die non-E-box Sequenzen bindet) liegt wahrscheinlich in einem Protein, welches beide Sequenzen binden konnte, mit einem Vorzug für die nicht-kanonische Sequenz. Dies impliziert, dass die Gruppen entstanden sind indem sich entweder eine Präferenz des Vorgängerproteins für die nicht-kanonische Sequenz durchgesetzt hat oder, dass sich eine Präferenz für die E-box bindende Sequenz durchgesetzt hat und somit die Affinität dahingehend verschoben wurde.

Die Hauptfunktion der meisten Proteine der SREBP Familie in Eukaryoten ist die Kontrolle der Lipid Biosynthese. In meiner Arbeit habe ich mich auf die Erforschung der SREBPs in einer Gruppe von Organismen zugewandt, die auch den mit dem Menschen assoziierten Hefepilz *Candida albicans* umfasst. Erstaunlicherweise beeinflussen die drei SREBPs die im *Candida albicans* Genom zu finden sind, die Kolonisierung des Säugetierdarms, jedoch nicht durch die Kontrolle der Lipid Biosynthese. Im Folgenden werde ich beschreiben wie zwei der drei SREBPs aus *Candida albicans* eine regulatorische Kaskade bilden, welche Einfluss auf die Regulierung der Morphologie und der Zellwandzusammensetzung des Pilzes unter anaeroben Bedingungen hat, wohingegen das dritte Protein der SREBP Familie für die Regulierung der Glykolyse von Bedeutung ist. Ich habe festgestellt, dass die beschriebene Vielfalt mit der diese Proteine an bestimmte DNA Sequenzen binden und die damit einhergehende Expansion der regulierbaren Ziele ein wesentlicher Grund dafür ist, dass Organismen dieses Stammbaums erfolgreich Säugetiere kolonisieren können.

## Abbreviation index

% (v/v)	% (volume/volume)
°C	degree Celsius
A (amino acid)	alanine
A (nucleobase)	adenine
<i>A. fumigatus</i>	<i>Aspergillus fumigatus</i>
aa	amino acid
Amp	ampicillin
Anc	ancestor
APR	ancestral protein reconstruction
APS	adenosine 5'-phosphosulfate
BAM	binary alignment map
bHLH	basic helix-loop-helix
bp	base pair(s)
BSA	bovine serum albumin
C	(amino acid) cysteine
C	(nucleobase) cytosine
<i>C. albicans</i>	<i>Candida albicans</i>
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>
<i>C. neoformans</i>	<i>Cryptococcus neoformans</i>
<i>C. parapsilosis</i>	<i>Candida parapsilosis</i>
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
chr	chromosome
D	aspartic acid
<i>D. melanogaster</i>	<i>Drosophila melanogaster</i>
Da	Dalton
DBD	DNA binding domain
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphate
dsDNA	double-stranded DNA
DTT	dithiothreitol
E	glutamic acid
<i>E. coli</i>	<i>Escherichia coli</i>
EDTA	ethylene diamine tetraacetic acid
EMSA	electrophoretic mobility shift assay
ER	endoplasmic reticulum
F	phenylalanine
G (amino acid)	glycine
G (nucleobase)	guanine
gDNA	genomic DNA

GEO	Gene Expression Omnibus
GI	gastrointestinal
H	histidine
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HGT	horizontal gene transfer
HRP	horseradish peroxidase
I	isoleucine
IP	immunoprecipitation
K	lysine
Kan	kanamycin
kb	kilo base pairs
kDa	kilo Dalton
L	leucine
LB	lysogeny broth
LDL	low density lipoprotein
M (amino acid)	methionine
M	molar
min	minute
MITOMI	mechanically-induced trapping of molecular interactions
mRNA	messenger RNA
N	asparagine
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
Ni-NTA	nickel-nitrilotriacetic acid
nt	nucleotide
OD <sub>600</sub>	optical density at 600 nm
oligo	oligonucleotide
ON	overnight
ORF	open reading frame
P	proline
PAA	polyacrylamide
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCA	principal component analysis
PCR	polymerase chain reaction
PMSF	phenylmethylsulfonyl fluoride
PNK	polynucleotide kinase
Pol	polymerase
Q	glutamine
qPCR	quantitative PCR
R	arginine
RNA	ribonucleic acid
RNA-seq	RNA sequencing

rpm	revolutions per minute
rRNA	ribosomal RNA
RT	room temperature
S	serine
<i>S. cerevisiae</i>	<i>Saccharomyces cerevisiae</i>
<i>S. pombe</i>	<i>Schizosaccharomyces pombe</i>
SAM	sequence alignment map
SD	standard deviation
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
sec	second
SREBP	sterol regulatory element binding protein
T (amino acid)	threonine
T (nucleobase)	thymine
TEMED	N,N,N,N,-Tetramethylethylenediamin
TH	Todd-Hewitt
TM	transmembrane domain
TR	transcription regulator
U (amino acid)	uridine
U	unit
V (amino acid)	valine
V	volt
vol	volume
W	tryptophan
WB	western blot
WIG	wiggle
WT	wild-type
Y	tyrosine
<i>Y. lipolytica</i>	<i>Yarrowia lipolytica</i>
YPD	yeast peptone dextrose

## List of figures

<b>Figure 1</b> Canonical SREBPs recognise two DNA sequences.....	27
<b>Figure 2</b> Ancestral protein reconstruction and synthesis. ....	42
<b>Figure 3</b> Phylogenetic reconstruction of the SREBP fungal family.....	46
<b>Figure 4</b> <i>In vitro</i> DNA binding preferences of the <i>C. albicans</i> SREBPs Hms1p, Cph2p and Tye7p in MITOMI.....	49
<b>Figure 5</b> <i>CaCph2</i> binds to a single half-site in cooperation with other co-factors.....	55
<b>Figure 6</b> Several fungal SREBPs exhibit an intrinsic ability to bind a non-palindromic DNA motif. ....	58
<b>Figure 7</b> Gel shift assay with fluorescently labelled DNA sequences competing for the same pool of <i>CaHms1</i> protein. ....	61
<b>Figure 8</b> The DNA binding specificity in of <i>CaHms1p</i> is conferred by residues in the first helix and loop region of its DNA binding domain.....	63
<b>Figure 9</b> <i>CaTye7</i> shows a preference for the canonical E-box motif over the non- palindromic DNA. ....	65
<b>Figure 10</b> Ancestral protein reconstruction at two selected nodes of the fungal SREBP phylogeny.....	67
<b>Figure 11</b> DNA binding preferences of the reconstructed ancestral proteins.....	68
<b>Figure 12</b> Divergence of DNA binding preferences in fungal SREBPs.....	69
<b>Figure 13</b> Mammalian SREBP pathway. ....	74
<b>Figure 14</b> SREBPs regulate a morphological switch in the Saccharomycotina clade.....	77
<b>Figure 15</b> The <i>C. albicans</i> SREBPs <i>CPH2</i> and <i>HMS1</i> form a regulatory cascade. ....	86
<b>Figure 16</b> The <i>C. albicans</i> SREBPs <i>CPH2</i> and <i>HMS1</i> affect morphology.....	91
<b>Figure 17</b> Anaerobic conditions induce massive changes in the transcriptome of <i>C. albicans</i> .....	93
<b>Figure 18</b> The inactive form of Cph2p localises in the membrane of the ER whereas the active form is restricted to the nucleus.....	98
<b>Figure 19</b> Changing levels of nutrients in the medium trigger the proteolytic cleavage of the <i>CaCph2p</i> .....	101
<b>Figure 20</b> Overview of the function of SREBP proteins in <i>C. albicans</i> . ....	105
<b>Appendix Figure 21</b> MITOMI replicate agreement. ....	123
<b>Appendix Figure 22</b> Extended phylogenetic tree of fungal SREBPs.....	129

## List of tables

<b>Table 1</b> List of DNA regions occupied by <i>CaCph2p</i> based on ChIP-seq experiments.....	52
<b>Appendix Table 2</b> Oligonucleotides used in MITOMI experiments.....	112
<b>Appendix Table 3</b> Primers used in this study.....	124
<b>Appendix Table 4</b> Protein expression plasmids generated in this study .....	126
<b>Appendix Table 5</b> Sequences of the DBD of fungal SREBPs included in the phylogenetic tree.....	127
<b>Appendix Table 6</b> Comprehensive lists of DNA motifs derived from MITOMI. ....	130
<b>Appendix Table 7</b> <i>C. albicans</i> strains used in this dissertation.....	137
<b>Appendix Table 8</b> List including the most differentially expressed transcripts under anaerobic conditions.....	138

# Table of contents

<b>AFFIDAVIT</b> .....	<b>I</b>
<b>ACKNOWLEDGMENTS</b> .....	<b>II</b>
<b>SUMMARY</b> .....	<b>III</b>
<b>ZUSAMMENFASSUNG</b> .....	<b>V</b>
<b>ABBREVIATION INDEX</b> .....	<b>VII</b>
<b>LIST OF FIGURES</b> .....	<b>X</b>
<b>LIST OF TABLES</b> .....	<b>XI</b>
<b>TABLE OF CONTENTS</b> .....	<b>XII</b>
<b>1. INTRODUCTION</b> .....	<b>15</b>
1.1. Origins of phenotypic novelty.....	<b>15</b>
1.2. Transcription regulators, i.e. sequence-specific DNA binding proteins.....	<b>17</b>
1.3. Evolution of transcriptional regulatory networks as a source of phenotypic diversity.....	<b>18</b>
1.3.1. Changes in <i>cis</i> -regulatory elements.....	<b>18</b>
1.3.2. Changes in <i>trans</i> -regulatory elements .....	<b>19</b>
1.4. Ancestral protein reconstruction .....	<b>20</b>
1.5. The ascomycete yeasts .....	<b>21</b>
1.6. <i>Candida albicans</i> is a gut commensal in humans .....	<b>22</b>
1.7. Yeasts as a model organism to study how new traits arise .....	<b>22</b>
<b>2. RESULTS</b> .....	<b>24</b>
<b>2.1. Diversification of DNA binding preferences in a conserved family of transcription regulators</b> .....	<b>24</b>
2.1.1. Summary.....	<b>24</b>
2.1.2. Introduction.....	<b>25</b>
2.1.2.1. The SREBP family of transcription regulators.....	<b>25</b>
2.1.2.2. DNA binding preferences of SREBPs.....	<b>25</b>
2.1.2.3. Different methodologies employed to determine transcription regulators' DNA binding specificity.....	<b>28</b>

2.1.3. Materials and methods .....	31
2.1.3.1. SREBP nomenclature .....	31
2.1.3.2. Phylogenetic reconstruction.....	31
2.1.3.3. MITOMI 2.0.....	32
2.1.3.4. ChIP-seq analysis.....	32
2.1.3.5. Reverse transcription and real-time PCR .....	36
2.1.3.6. Plasmid construction .....	37
2.1.3.7. Protein purification .....	38
2.1.3.8. Electrophoretic mobility shift assays and competition assays.....	39
2.1.3.9. Ancestral protein reconstruction .....	41
2.1.4. Results.....	43
2.1.4.1. The SREBP family of transcription regulators in fungi.....	43
2.1.4.2. Different branches of the SREBP phylogenetic tree show distinct DNA binding preferences.....	47
2.1.4.3. Several fungal SREBPs show higher affinity for a non-canonical motif.	57
2.1.4.4. Tracing the structure of the DNA binding domain that confers specificity to <i>CaHms1p</i> .....	62
2.1.4.5. Ancestral protein reconstruction reveals pattern of divergence of SREBP's DNA binding preferences.....	64
2.1.5. Discussion.....	70
<b>2.2. Diversification of the biological functions controlled by the SREBP family of transcription regulators .....</b>	<b>72</b>
2.2.1. Summary.....	72
2.2.2. Introduction.....	73
2.2.2.1. SREBPs regulate the expression of sterol biosynthesis genes in most eukaryotes .....	73
2.2.2.2. SREBPs have adopted different functions in yeasts .....	75
2.2.2.3. <i>Candida albicans</i> morphologies: yeast and filamentous forms .....	78
2.2.2.4. Anaerobiosis in <i>C. albicans</i> .....	79
2.2.3. Materials and methods .....	80
2.2.3.1. Transcriptome analyses .....	80
2.2.3.2. Cell morphology determination.....	81
2.2.3.3. Protein immunostaining.....	82



2.2.3.4. Western blot analysis .....	83
2.2.4. Results.....	85
2.2.4.1. The <i>C. albicans</i> SREBPs Hms1p and Cph2p form a regulatory cascade.	85
2.2.4.2. SREBPs regulate morphology in <i>C. albicans</i> .....	89
2.2.4.3. Anaerobic conditions induce changes in the <i>C. albicans</i> cell wall .....	92
2.2.4.4. Intracellular localisation of the Cph2 protein in <i>C. albicans</i> .....	95
2.2.4.5. The signal that triggers the cleavage and release of Cph2p from the membrane differs from most SREBPs .....	99
2.2.5. Discussion.....	102
<b>3. DISCUSSION.....</b>	<b>106</b>
<b>APPENDIX.....</b>	<b>111</b>
<b>REFERENCES.....</b>	<b>140</b>
<b>CURRICULUM VITAE.....</b>	<b>CLIV</b>
<b>LIST OF PUBLICATIONS.....</b>	<b>CLVII</b>
<b>CONFERENCES AND COURSES ATTENDED.....</b>	<b>CLVIII</b>

# 1. Introduction

## 1.1. Origins of phenotypic novelty

Evolution encompasses all living things and has shaped every single organism since life started. Organisms that belong to a population are generally conditioned by variation that results in their evolution and adaptation to changes in the environment. New traits in any given organism can emerge from a diverse variety of sources, for instance by mutations in their DNA sequences, acquisition of new genetic material via duplication or transfer from another organism or by mutations in non-coding elements.

Mutations in the DNA sequences of existing genes can result beneficial by contributing to efficiency or increasing the fitness of the organism as well as expanding the variety of conditions the organism can survive in [1,2].

Acquisition of new genes by gene duplication also represents one of the most important processes by which organisms generate new genes [3–5]. It can occur by duplication of the entire genome [6], a chromosome (or part of a chromosome) or a single gene (or group of genes) [7]. The situation that immediately arises after gene duplication is the existence of two identical copies of the duplicated gene. This situation of redundancy is evolutionarily predicted to be short-lived (studies suggest that between 50% and 92% of genes that arise after duplication get lost [4]). The newly arisen gene copy is no longer subjected to selective pressure and will most likely accumulate deleterious mutations and become inactivated leading to gene loss [4]. However, from time to time accumulated mutations might lead to the acquisition of a novel function in the new duplicated gene and/or a change of its expression pattern. An example of how gene duplication can originate evolutionary innovation can be found in the vertebrate globin family [7,8] where a series of genome and gene duplication events directed the evolution and physiological division of function in the ancestor of jawed vertebrates that resulted in the functional specialisation of haemoglobin (specialised in carrying oxygen) and myoglobin (specialised in its storage) [7].

Another process by which organisms can acquire new traits is horizontal (or lateral) gene transfer (HGT). It is a common process in bacteria and archaea, and is defined as the sharing of genetic material between lineages that are not in a parent-offspring relationship (this distinguishes it from the standard vertical gene transfer from parent to offspring) [9–11]. It was initially thought that for a transferred gene to survive in the recipient lineage, that gene must contribute to the organism with a beneficial or selective advantage. It is now clear, however, that many cases of HGT result in neutral or nearly neutral effects in the recipient organism but can later be the source of new combinations of genetic material for selection to act on [10].

Changes in non-coding elements (*i.e.* regulatory DNA sequences located upstream of coding regions and that are recognised by transcriptional regulatory proteins to promote or repress expression) and/or in regulatory proteins also comprise an important source of phenotypic diversity. In the early seventies two influential studies [12,13] already highlighted the importance of regulatory mutations in the generation of new phenotypes. Many examples have been studied to date where mutations and changes in regulatory elements generate new traits. For instance, in the fruit fly *Drosophila melanogaster*, naturally occurring transposable elements have been reported to disrupt the promoter function of the *hsp70* heat shock protein, which results in increased thermal tolerance [14,15]. Another example can be found in humans; a relatively recent adaptation in our diet is the ability to digest lactose once we have reached adulthood (which is termed lactase persistence). *LCT* (lactase-phlorizin hydrolase) is the gene encoding the enzyme that catalyses the hydrolysis of lactose into glucose and galactose. Its levels reach a maximum at birth and then decrease through aging. However, in a minority of adults, high levels persist during adulthood allowing lactose persistence. Researchers found that mutations in regulatory regions were genetically associated with lactose persistence because they elevate *LCT* transcription. These mutations are highly common in individuals from northern Europe causing these populations to be more lactose persistent than in other regions [16,17]. Thus, mutations in the regulatory region of a key enzyme contribute to changes in an ecologically relevant trait.

Defining the effects of mutations in coding DNA typically relies on studies based on scanning coding DNA and protein sequences. The comparison of sequences allows identification of mutations that alter the protein structure such as non-synonymous

substitutions, frame shifts, early stop codons, etc. However, most regulatory mutations in non-coding DNA sequences can only be identified through functional or biochemical tests often making them under-represented in evolutionary studies [18].

## **1.2. Transcription regulators, i.e. sequence-specific DNA binding proteins**

Gene regulation is crucial for the maintenance and survival of most cells. Cells usually adapt to fluctuating conditions by expressing adequate sets of genes in a specific spatial and temporal manner to respond appropriately. Transcription regulators (TRs) are central players in the regulation of gene transcription. They are called *trans*-regulatory elements because they are regulatory agents that are not part of the regulated genes. They function by recognising and binding to specific short DNA sequences (*cis*-regulatory elements) that are typically located within neighbouring regions upstream of the gene being regulated [19]. Generally, a TR can bind to many locations in the genome that harbour its target sequence therefore regulating the expression of tens or hundreds of genes. Upon binding, TRs can influence transcription in a positive or negative manner. Many regulators belonging to the same protein family often work cooperatively to control the expression of the same gene. Additionally, several of the genes being regulated can be in turn TRs themselves, expanding the downstream effects and creating large and complex transcriptional regulatory networks [20]. Much of the diversity and complexity in higher eukaryotes can be attributed to the evolution of elaborate networks rather than to a larger number of genes [21]. TRs influence many critical aspects of the biology of the cell such as development, cell type and morphology. Thus, a cell deleted of a TR gene might exhibit strong abnormalities in organisation and development.

Transcription regulators are classified into different families according to the sequence and structure of the DNA-binding domain that they use to directly interact with DNA [22]. Some of the most conserved and functionally relevant examples are: the zinc finger family, generally implicated in chromatin remodelling and DNA structures [23]; the leucine zipper family in which hydrophobic leucines allow dimerization of the proteins and have roles in cell division and protein interactions [24,25]; and the basic helix-loop-

helix (bHLH) which are involved in development of muscle, nerve, blood and pancreatic cells in higher eukaryotes and regulate metabolic pathways for nutrient uptake and lipid production in yeasts [26,27]. In this thesis I study a group of transcriptional regulators that belong to the bHLH family, which is extensively described in chapter 2.1.2.1.

### **1.3. Evolution of transcriptional regulatory networks as a source of phenotypic diversity**

Transcription regulation is the primary step by which most cells control gene expression and is mediated by TRs (*trans*-regulatory elements) that bind to specific short DNA motifs (*cis*-regulatory elements). Transcription regulators are often conserved and retain the same DNA-binding specificity across large phylogenetic distances. However, in the past decades there have been a number of studies in multiple model organisms ranging from vertebrates to unicellular fungi that have revealed that evolutionary changes either in the DNA regulatory elements or in the TRs themselves underlie the origin of many traits such as morphological innovations or the ability to colonise new environments [20,28]. Additionally, analyses of transcription regulatory circuits in yeast have shown that network rewiring (breaking of old connections between TR and DNA target sequence, and formation of new ones) happens at high rates over relatively short evolutionary timescales [29–31].

#### **1.3.1. Changes in *cis*-regulatory elements**

As mentioned above, *cis*-regulatory elements are short DNA sequences recognised and bound by TRs. Gains and losses of *cis*-regulatory sequences have been found to be an important driver of evolution and underlie many cases of transcriptional rewiring [28,32,33]. One well-studied example of *cis*-regulatory change giving rise to a new phenotype is the origin of wing pigmentation patterns in *Drosophila melanogaster* by a gene called *Yellow*. Species of the fruit fly *Drosophila* display distinct patterns of

pigmentation on their bodies and wings. A common feature in male flies is the presence of a dark spot close to the tip of the wing. The production of these dark spots requires the activity of enzymes able to synthesise melanin. The *Yellow* gene (called yellow because its mutation turns dark areas of pigmentation into yellow or brown colour) plays a key role in the divergence of melanin pigmentation patterns. On the one hand, in species with dark spots *Yellow* is highly expressed in wing cells causing the production of dark spots. On the other hand, species that lack wing pigmentation express low levels of *Yellow* at the wing tips. Theoretically, differences in expression of *Yellow* between species could be due to differences in the spatial-temporal presence of TRs that control the expression of the gene (changes in *trans*-) and/or differences in the *cis*-regulatory sequences in the vicinity of *Yellow* that also contribute to the control of its expression. To get insights on the mechanisms that regulate this gene, the activity of *cis*-regulatory sequences that regulate *Yellow* was analysed by placing them upstream of a reporter gene and introducing them into *D. melanogaster*. The results revealed that while *cis*-regulatory elements at the wing tip of species that lack dark spots express low levels of the reporter gene at the wing end, the corresponding element in pigmented species (like *D. biarmipes* or *D. elegans*) expressed high levels of the reporter gene. These observations showed that changes in the sequence and function of the *cis*-regulatory element in different species were responsible for changes in the regulation of *Yellow* and contributed to the generation of dark pigmentation in the wing. In conclusion, mutations of ancestral *cis*- elements generated new *cis*- binding sites [34]. By mutating only a *cis*- sequence, effects of mutations in coding sequences of genes that act in multiple processes or in TRs that control the expression of hundreds of genes can be avoided. Other examples where changes in *cis*- DNA sequences underlie interesting alterations in phenotypes are pesticide resistance in fruit flies, paternal care behaviour in rodents and susceptibility to schizophrenia or lactose intolerance in humans [18]

### **1.3.2. Changes in *trans*-regulatory elements**

For many years changes in *cis*-acting elements were considered to be the main drivers of network rewiring because they were hypothesised to have fewer pleiotropic

effects than mutations in TRs that can affect the expression of hundreds of genes [35]. However, although TRs tend to be largely conserved across large phylogenetic distances, several studies point to evolutionary changes in TRs as a source of phenotypic novelty [20]. One clear example can be found in the *LYS* transcription regulators in fungi. In the environmental yeast *S. cerevisiae*, Lys14 regulates lysine biosynthesis. However, in *C. albicans*, this very conserved TR underwent successive duplication events that led to the generation of four paralogs (Lys14, Lys142, Lys143 and Lys144) that diverged in their DNA binding profiles. This diversification resulted in each paralog regulating a different set of genes and in the generation of novel regulatory connections that appear to be crucial for *C. albicans*' ability to colonise its mammalian host [36].

In this thesis I study a family of TRs that also appears to have diversified the DNA binding preferences of some of its members. I investigate the mechanisms by which the diversification took place and how these changes may have contributed to the biology of the fungus *C. albicans*.

#### **1.4. Ancestral protein reconstruction**

Being able to study how the structure and properties of biological molecules change over time is essential to understand their evolution [37]. However, for many years, researchers did not have at their disposal tools and techniques that allowed them to empirically address such questions. Conventional evolutionary and biochemical approaches to study how proteins evolve focus mainly on horizontal comparisons; that is, comparisons of protein homologs from different branches of a phylogenetic tree. Ancestral protein reconstruction is a relatively new technique that allows researchers to study the evolutionary processes of how protein structures diversify and to reveal how new architectures and structures impact their evolution. It allows the “resurrection” of proteins within the nodes of a phylogenetic tree in a vertical manner [37–39]. By employing ancestral protein reconstruction, important questions about reversibility of evolution, complexity or functionality of ancestors can be tackled. The resurrection of ancestral proteins can thus illuminate relevant aspects in molecular evolution because it

can reveal the mechanisms by which historical mutations led to the emergence of new traits and give insights about how the physical properties of a protein shaped the evolutionary processes.

Ancestor protein reconstruction has been employed, for instance, to study the evolution of hormone specificity in the vertebrate glucocorticoid receptor (GR). The family of glucocorticoid receptors is divided in two groups; one group contains the elasmobranch GRs, mineralocorticoid receptors (MRs) and agnathan corticosteroid receptors (CRs), which are sensitive to both cortisol and mineralocorticoid hormones, and the other group harbours the teleost and tetrapod GRs which are sensitive to cortisol only. The study shows that the cortisol-specificity of the “modern” glucocorticoid receptor ligand-binding domain (LBD) evolved from a more promiscuous ancient receptor that was activated by the mineralocorticoids aldosterone and deoxycorticosterone (DOC) and, although in a weaker manner, also by cortisol [40,41].

## **1.5. The ascomycete yeasts**

Fungi are one of the most extensive eukaryotic kingdoms comprising an estimated of 1.5 – 5 million species [42,43]. The Ascomycota is the largest of the fungal phyla with around 64.000 known species. It is a monophyletic group meaning that in it are contained all the descendants of one common ancestor. They are divided in three subphyla: the Pezizomycotina which is the largest group and includes macroscopic fungi like truffles; the Taphrinomycotina in which fission yeasts (*Schizosaccharomyces*) are included and the Saccharomycotina which comprises most of the ascomycete yeasts such as the *Saccharomyces* and *Candida* clades. The Saccharomycotina subphylum (the ascomycete yeasts) contains a single class (Saccharomycetes), which in turn contains a single order, the Saccharomycetales. In this monophyletic lineage are comprised about 1.000 known species [44]. The ascomycete yeasts comprise some of the most economically relevant fungal species used in industry to brew, bake, and ferment food products. They can be found in a wide range of environments like soil, salt or fresh water and in the atmosphere as well as in association with plants or animals. This thesis will be primarily focused on



the ascomycete yeast *Candida albicans*, an opportunistic commensal pathogen that colonises several mucosal surfaces in the human body (see chapter 1.6).

### **1.6. *Candida albicans* is a gut commensal in humans**

In contrast to environmental yeasts such as *Saccharomyces cerevisiae*, *Candida albicans* has no known natural reservoir outside the host and resides in association with warm-blooded animals being a member of the normal human microbiota [45]. *C. albicans* resides in several body sites such as skin, oral cavity, genitourinary tract and gastrointestinal (GI) tract in healthy individuals [46]. In most cases, humans are life-long colonised by harmless *C. albicans*. Overgrowth of the fungus is believed to be controlled and limited by microbial competitors [47]. However, certain conditions such as prolonged antibiotic treatment, chemotherapy or immunosuppression may interfere with the equilibrium of the microbiome and the balance between *C. albicans* and microbial community can be disrupted. Such dysbiosis may lead to a decrease of the populations of bacterial competitors that help to keep *C. albicans* “in line.” This in turn could result in an uncontrolled overgrowth of the fungus which can cause serious and life threatening infections [48]. Fungi-host interactions have typically been studied in the context of infections (*i.e.* when they cause disease); by contrast, the commensal interplay between humans and fungi and the genetic determinants that enable the fungus to inhabit the mammalian intestine remain to be elucidated [45].

### **1.7. Yeasts as a model organism to study how new traits arise**

Ascomycete yeasts are a powerful model to study how eukaryotic transcriptional regulatory networks are modified by evolutionary processes to generate phenotypic diversity. First of all, genomes have been sequenced and are available for a high proportion of ascomycete yeasts. Second, there is a large amount of evolutionary history

within the ascomycetes. Phylogeny based on single gene analyses has shown that ascomycetes are a highly divergent group despite morphological similarities between most species [44]. Third, in comparison with higher eukaryotes, ascomycete yeasts have relatively small genomes (the genome of *S. cerevisiae* is approximately 12Mb and *C. albicans* is 16Mb compared to the human's 3,000Mb) and are genetically tractable. Techniques based on next generation sequencing (NGS) can be used to experimentally map regulatory circuits. For instance full genome chromatin immunoprecipitation followed by sequencing (ChIP-seq) can be used to determine all the genomic locations where transcription regulators bind; or a transcription regulator can be easily knocked out and the effects can be assessed by sequencing the full transcriptome (RNA-seq) [20]. Finally, due to their unicellular nature, transcriptional regulatory networks have a tendency to be simpler than those of multi cellular organisms and therefore easier to dissect [20].

## 2. RESULTS

### 2.1. Diversification of DNA binding preferences in a conserved family of transcription regulators

#### 2.1.1. Summary

The Sterol Regulatory Element Binding Proteins (SREBPs) are transcription regulators that belong to the basic helix-loop-helix (bHLH) family. These proteins are highly conserved among eukaryotes and as their name implies, they regulate lipid biosynthesis genes in most organisms. However, in the ascomycete yeasts, an unrelated transcription regulator (belonging to the zinc finger family) has taken over this function. Despite this fact, SREBPs have expanded in some ascomycetes such as *Candida spp.*, which raises questions about their function and evolution in these organisms. In general, most bHLH proteins are known to recognise and bind to a canonical E-box DNA sequence. In this chapter, I examine the DNA binding preferences of fungal SREBPs and report that several branches of the fungal SREBP phylogeny have higher affinity for a non-canonical DNA sequence over the palindromic E-box sequence. Additionally, I employ ancestral protein reconstruction and characterisation to establish that the divergence in DNA binding preferences in this particular family originated from a likely ancestor that was promiscuous and bound both DNA sequences. Thus, the repertoire of DNA binding affinities of the current SREBP members originated as a partition and further specialisation of an ancestral function.

## **2.1.2. Introduction**

### **2.1.2.1. The SREBP family of transcription regulators**

The SREBPs are a family of proteins highly conserved through eukaryotes ranging from unicellular fungi to higher and more complex organisms such as the fruit fly *Drosophila melanogaster* or humans. They were first described in the early 90s when the human SREBP1 protein was reported to function by binding a short DNA sequence (i.e. a regulatory element) upstream of the promoter of a sterol receptor [49], which lead to the name SREBP (**S**terol **R**egulatory **E**lement **B**inding **P**rotein). SREBPs are transcription regulatory proteins and therefore their main function is to recognise and bind to specific DNA sequences to promote or repress transcription. The DNA binding domain of the SREBPs is composed of a basic region followed by two amphipathic  $\alpha$ -helices connected by a loop. Thus, the SREBPs belong to the basic helix-loop-helix (bHLH) family of transcription regulators, which has been well studied and is one of the most functionally relevant family of regulators in all eukaryotes [50]. The architecture of the DNA binding domain of a transcription regulator determines, in most cases, the DNA sequence (or sequences) that the regulator recognises [51]. SREBP proteins harbour a bHLH DNA binding domain and thus the premise is that they should have DNA binding properties similar to those of bHLH proteins.

### **2.1.2.2. DNA binding preferences of SREBPs**

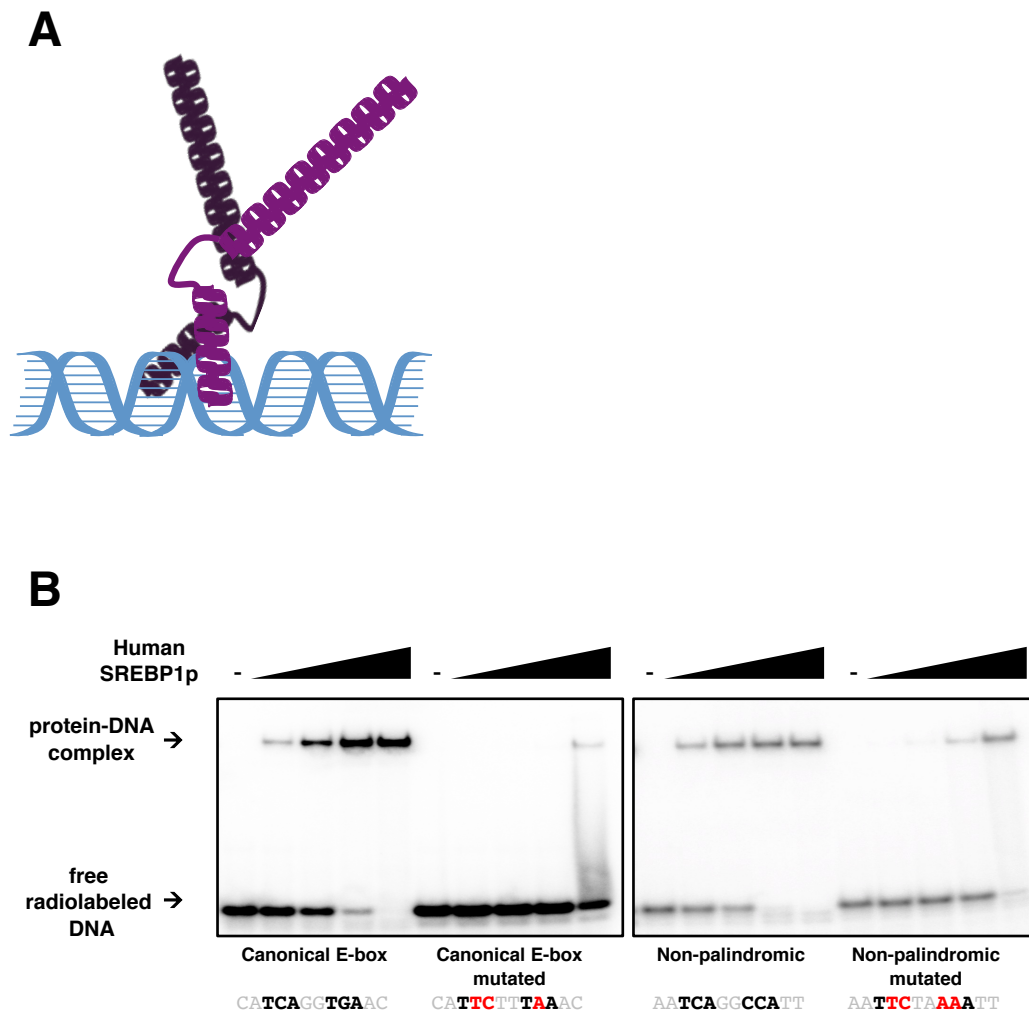
As previously mentioned, SREBPs belong to the bHLH family of regulators and as such they possess a bHLH DNA binding domain. Typically, bHLH proteins contain a characteristic 60-to-100-residue DNA binding domain composed of two segments that form amphipathic  $\alpha$ -helices separated by a loop region that varies in sequence and length. Basic HLH proteins have been well studied and characterised [50,52,53]. Most of them

form homo- or heterodimers where the first helix of each monomer establishes specific contacts with the DNA (Figure 1A).

They typically bind to the so-called E-box DNA motifs which are consensus 5'-CANNTG-3' sequences [27,54]. This palindromic sequence has been shown to be bound by many bHLH proteins in a variety of organisms [50,55–57].

The SREBPs also possess a DNA binding domain composed of a basic region and two helices separated by a loop. In the case of SREBPs the loop is also highly variable in length (for instance Hms1p of the fungus *Candida albicans* contains 124 amino acids in its loop whereas the loop of the human SREBP1p contains only 8 amino acids). However, a characteristic feature that makes the SREBPs a unique and distinct subgroup within the bHLH family is the presence of a tyrosine residue at a very conserved position of the first helix of their DNA binding domain where all other bHLH proteins have an arginine instead [58,59]. Crystal structure studies have shown that this tyrosine residue plays a critical role in DNA recognition and binding. In particular, in the human SREBP the tyrosine is responsible for allowing SREBP1 to bind to an additional DNA sequence besides the canonical E-box (Figure 1B and [59]). In contrast to the palindromic 5'-CANNTG-3, this second DNA sequence that SREBPs are able to recognise is non-palindromic (5'-TCACCCCA-3') and significantly different from the canonical E-box. *In vitro* electrophoretic mobility shift assays confirm the “dual” binding ability of SREBPs which recognise and bind to both the palindromic and non-palindromic DNA sequences with similar affinities [58,59].

However, despite the clear *in vitro* dual binding ability, when *in vivo* chromatin immunoprecipitation (ChIP) experiments were performed for the human SREBP, the protein showed binding only to the canonical, palindromic E-box sequence that all other bHLH proteins bind [56]. Thus, the significance of the “dual” DNA binding preferences of the SREBPs remains unclear.



**Figure 1 Canonical SREBPs recognise two DNA sequences.**

(A) Schematic representation of the structure of the DNA binding domain of a typical SREBP homodimer bound to DNA. Shown in the figure is an SREBP homodimer (one monomer in purple and a second one in black) binding a DNA strand (in blue). Each monomer contacts the DNA strand through the first helix. (B) The SREBP family of regulators is known to bind *in vitro* to two related DNA sequences. Shown is an example of a gel-shift assay carried out with the purified DNA binding domain of the human SREBP1p. The protein shows specific binding to a palindromic E-box variant as well as to a non-palindromic DNA sequence. Protein concentrations added in each lane are 0.56, 6.25, 25 and 100 nM.

### 2.1.2.3. Different methodologies employed to determine transcription regulators' DNA binding specificity

A number of methods including experimental and computational, *in vivo* and *in vitro*, specific and comprehensive, have been developed to unveil and measure protein-DNA interactions. A first classification of these methods is *in vivo*- versus *in vitro*-based experiments. On the one hand, *in vitro* approaches are generally used to identify transcription regulator (TR) consensus binding sites or define the biophysics of the TR-DNA interaction and therefore can produce qualitative and quantitative data. Additionally, they allow capturing direct binding of the TR to the DNA and avoiding the need for signals that trigger the activation of the TRs. On the other hand, *in vivo*-based methods aim to provide information about the consensus TR binding site as well as the biological context of the specific interaction [60].

A traditional biochemical method used to characterise transcription regulator binding sites *in vitro* is electrophoretic mobility shift assays (EMSA). In this approach (which will be used throughout this thesis), a purified protein (or its purified DNA binding domain) is incubated with a radioactively labelled DNA probe harbouring the putative DNA sequence bound by the protein of interest. The mixture is resolved by gel electrophoresis and the binding can be measured by analysing the “shift” in size that is produced when the protein is bound to the DNA compared to the smaller size of free-radiolabelled DNA [61]. A major downside of this approach is that it does not allow high-throughput and systematic characterisation of binding sites. Moreover, it requires previous knowledge and information about the potential TR binding site.

One of the first *in vitro* approaches to overcome the need for previous information about the binding sites is the so-called systematic evolution of ligands by exponential enrichment (SELEX) [62]. In this method a purified transcription regulator is incubated with a pool of random DNA oligomers. Oligos bound by the TR are selected and amplified by PCR. These amplified oligos are then re-incubated with the TR and repeated rounds of selection allow the identification of high-affinity binders or the consensus DNA binding site that the protein binds to [62,63]. The main disadvantage of this approach is that only a small percentage of high-affinity binding sites are selected and then amplified making it hard to capture the relationship between sequence and binding affinity of the DNA

binding sites [64]. This limitation however can be overcome by carrying out massive parallel sequencing after only one round of *in vitro* binding selection in SELEX-seq approaches although this method does not provide information about affinity [65].

Another widely used *in vitro* approach consists on performing the TR-DNA binding reactions on immobilised double-stranded DNA microarray chips. TRs are allowed to bind to protein-binding microarrays (PBMs) and after a series of washes the binding events are quantified by measuring intensities of emitted by fluorophore-coupled antibodies specific for the protein of interest [66,67]. The use of PBMs increases the resolution of detection of TR motifs and additionally enables the use of a single microarray design to examine a wide range of TR binding preferences [66]. PBMs allow capturing strong as well as weak TR-DNA interactions, which is not possible with classical biochemical approaches like EMSA. Weak affinity sites are now known to be evolutionarily conserved and relevant during transcription [68]. However, PBMs cannot measure reactions at equilibrium preventing affinity measurements and have a reduced sensitivity [69].

The limitation of traditional PBMs can be overcome with the generation of high-throughput microfluidic platforms that use novel detection methods based on mechanically-induced trapping of molecular interactions (MITOMI) [70,71]. In short, MITOMI microfluidic devices fabricated in polydimethylsiloxane (PDMS) using multilayer soft lithography are aligned to a glass slide where thousands of micro arrayed-DNA spots (each containing a different Cy5-labelled DNA sequence) are printed. Micromechanical valves control each chamber in the device. Fluorescently labelled TRs are generated by *in vitro* transcription and translation systems and loaded onto the device. The addition of TRs solubilises the spotted DNA in each chamber and allows it to interact with the protein. TR-DNA interactions are incubated and trapped at equilibrium. After a series of washes, where unbound protein and DNA are discarded, fluorescence intensities are measured in every chamber of the device. These intensity ratios reflect interaction affinities between the TR and a specific oligonucleotide. This technique is capable of measuring thousands of parallel binding interactions at equilibrium in one device and therefore allows the construction of detailed maps of binding energy landscapes. As well as PBMs, MITOMI can detect transient and low-affinity interactions [71].

A general major drawback from *in vitro* approaches is their inability to provide information about the biological context, *e.g.* interaction with co-factors or post-



translational modifications, that affect the binding abilities of the TR, accessibility of binding sites depending on chromatin packing or occupancy of binding sites by other factors. These questions can generally be answered by *in vivo* approaches.

The traditional, and one of the first developed, approaches to detect TR-DNA binding interactions *in vivo* involves promoter deletion analysis coupled to a reporter assay. In this type of experiments the promoter region (or fragments of it) upstream of a gene is deleted and fused to a reporter gene such as luciferase. Quantitative or qualitative measurements of the levels of expression of the reporter gene allow identification of regulatory regions of interest from the promoter [72]. Since this traditional approach involves cloning and modification of every specific site of interest it is not amenable to high throughput analyses. In the past decades, methods that produce high yield results have been developed. The most widely used whole genome binding approaches are chromatin-immunoprecipitation followed by either hybridisation to a microarray (ChIP-chip) or next generation sequencing (ChIP-seq), which allow *in vivo* mapping of the genome-wide distribution of TR binding sites at base-pair resolution. ChIP methods are based on an initial cross-link between the DNA binding protein and its cognate DNA followed by shredding of the DNA and further immunoprecipitation using an antibody directly targeted against the protein (which is still bound to the specific DNA sequences). Next, the DNA is purified and amplified so that DNA locations where the protein was bound are enriched. Further analysis of the bound regions uncovers overrepresented sequences which are likely motifs bound by the transcription regulator of interest [73]. Most high throughput *in vitro* and *in vivo* methods rely on substantial computational analyses to dissect the data and identify overrepresented sequences among all bound locations. However, not all the binding sites predicted *in silico* are actually being occupied *in vivo*. Thus, it is essential to be able to discriminate between biologically functional and non-functional binding sites by empirically validating *in silico* predicted targets.

### 2.1.3. Materials and methods

#### 2.1.3.1. SREBP nomenclature

The standard and systematic names of the main SREBP genes included in this dissertation are as follows: *TYE7* (*ORF19.4941* or *C1\_13140C\_A* in *C. albicans*; *YOR344C* in *S. cerevisiae*); *CPH2* (*ORF19.1187* or *C6\_00280W\_A* in *C. albicans*; *YOR032C* in *S. cerevisiae*); *HMS1* (*ORF19.921* or *C5\_00670C\_A* in *C. albicans*). Notice that although the standard name of the *S. cerevisiae* gene *YOR032C* in the *Saccharomyces* genome database is *HMS1*, the phylogenetic reconstruction presented in this study places it closer to the *C. albicans* *CPH2* branch (Figure 3).

#### 2.1.3.2. Phylogenetic reconstruction

Fungal protein sequences were retrieved from UniProt [74]. The maximum likelihood tree was constructed by aligning the basic region, first helix and second helix (55 amino acids in total) of the DNA binding domain of 198 fungal SREBPs in MEGA7 [75]. Because the loop of SREBP proteins is highly variable in sequence and length, only the last five residues of the loop were taken into account and the rest of the loop region was omitted from the alignment. Only proteins carrying the characteristic tyrosine residue in the first helix (a defining feature of the SREBP family) were included in the analysis. ProtTest [76] was used to find the best-fit model to infer the phylogenetic tree (LG+G). *OCTOPUS* software [77] was used to scan the full length sequence of the fungal SREBP proteins for the presence of transmembrane domains.

### 2.1.3.3. MITOMI 2.0

Linear templates encoding the putative DNA binding domains of *C. albicans* Hms1p (amino acids 463-685), Cph2p (amino acids 197-302) and Tye7p (amino acids 159-269) for protein expression were generated by a two-step PCR: First, a consensus sequence and a 6×His tag at 5' and 3' respectively were added using template specific primers. Second, using universal primers a T7 promoter sequence was added to the 5' end and a poly-A tail as well as a T7 terminator were added to the 3' end. An *in vitro* transcription-translation system (Promega) was used to generate GFP tagged versions of the constructs, which were then added to a microfluidic device (See section 2.1.2.3 of introduction and [71]) containing the Cy5-labelled DNA library. All experiments used a 740-oligonucleotide pseudorandom double-stranded 70-nt long DNA library containing all possible 8-nucleotide sequences (Appendix Table 2). Protein-DNA interactions were trapped at equilibrium. After a series of washings where unbound DNA and proteins were washed out, the GFP/Cy5 intensity ratio was measured in every chamber of the device. Experiments were performed in duplicates (see replicate agreement in Appendix Figure 21). Cytoscape (v3.4) [78] was used to visualise the data. MatrixREDUCE [79] was used to derive MITOMI motifs. Topologies X6, X7, X3N2X3 and X4N2X3 (scanning the forward, reverse or both strands) were employed to query the dataset. MatrixREDUCE calculates  $r^2$  and  $P$ -value for each motif, which were used to rank the resulting motifs (only motifs with  $P < 1 \times 10^{-10}$  were included in the final dataset). See full dataset of derived motifs in Appendix Table 6.

### 2.1.3.4. ChIP-seq analysis

Myc-tagged and untagged *C. albicans* strains (the latter served as a negative control) were grown in YPD broth at 30°C until mid-log phase. ChIP was carried out as described [80]. Briefly, 200 ml cultures grown under the conditions mentioned above were cross-linked with 37% formaldehyde for 5 minutes. The cross-linking was quenched by adding 2.5 M glycine to the cells. The samples were then centrifuged down, washed with TBS and frozen at -80°C. Cells were resuspended in lysis buffer and lysed mechanically with Zirconia

beads for approximately 4 hours. The lysate was recovered and subjected to three 15-minute rounds of sonication (30 seconds ON / 30 seconds OFF). Part of the lysate was stored to later be used as input. Samples for immunoprecipitation were incubated with mouse anti-Myc antibody overnight and the following addition of Sepharose beads was used to immunoprecipitate the DNA. IP samples were washed (using wash buffer) and eluted. All samples were incubated at 37°C for two hours and 66°C for 12 hours in the presence of proteinase K for cross-link reversal before being cleaned up (QIAGEN PCR clean-up kit). Input and immunoprecipitated DNAs were directly used to generate libraries for sequencing with the NEBNext® ChIP-Seq Library Prep Master Mix Set for Illumina (New England Biosciences). DNA sequencing was carried out by GATC Biotech (Konstanz, Germany) using standard procedures. The reads were aligned to the *C. albicans* genome using Bowtie2 [81]. The full alignment was performed according to the following steps:

Prior to the alignment, the *C. albicans* Built 21 genome was indexed:

```
$ bowtie2-build <reference genome.fasta> <output prefix>
```

Then sequencing unpaired reads in the format of FASTQ files were aligned to the genome:

```
$ bowtie2 -t -local -p 3 -x <indexed genome prefix> -U <sequencing  
reads.fastq> -S <output.sam>
```

The output SAM files were then filtered to remove non-uniquely aligned reads as well as reads with mismatches. For this and the following steps *samtools* was used:

```
$ samtools view -Sh <input.sam> | grep -e "^@" -e "XM:i:[012][^0-9]"  
| grep -v "XS:i:" > <output.filtered.sam>
```

Next, filtered SAM files were converted to BAM files, which were then sorted:

```
$ samtools view -bh <input.filtered.sam> > <output.bam>  
  
$ samtools sort -o <output.sorted.bam> <input.filtered.bam>
```



Diluent (pH 7.5)

0.143 M	NaCl
1.43 M	EDTA
71.43 mM	HEPES-KOH

Glycine solution

3 M	Glycine
20 mM	Tris-HCl pH 7.5

TBS

20 mM	Tris-HCl pH7.5
150 mM	NaCl

Lysis Buffer

50 mM	HEPES-KOH
140 mM	NaCl
1mM	EDTA
1 %	Triton X-100
0.1 %	Na-deoxycholate

Lysis Buffer w/500 mM NaCl

50 mM	HEPES-KOH
500 mM	NaCl
1 mM	EDTA
1 %	Triton X-100
0.1 %	Na-deoxycholate

Wash Buffer

10 mM	Tris-HCl pH8.0
250 m	MLiCl
0.5 %	NP40
0.5 %	Na-deoxycholate

1 mM        EDTA

Elution Buffer

50 mM        Tris-HCl pH8.0

10 mM        EDTA

1%            SDS

TE Buffer

10 mM        Tris-HCl pH8.0

1 mM        EDTA

Proteinase K solution

2  $\mu$ L        Proteinase K (20 mg/mL)

48  $\mu$ L        TE buffer pH 8.0

The CHIP-seq data generated in this study have been deposited in NCBI's Gene Expression Omnibus (GEO) under accession number GSE118416.

**2.1.3.5. Reverse transcription and real-time PCR**

Reference strain, *cph2* and *efg1* deletion mutants were grown under anaerobic conditions at 37°C in Todd-Hewitt broth to an OD<sub>600</sub> of 0.3. Total RNA was extracted using RiboPure™ RNA purification kit for yeast (Ambion, Life Technologies). Next, cDNA was synthesised using SuperScript II reverse Transcriptase kit™ (Life Technologies) following the manufacturer's instructions. Real time PCR was used to quantify specific transcripts (oligos listed in Appendix Table 3). The experimentally validated *TAF10* transcript [87] served as a reference control for the qPCR. The student *t*-test for unpaired samples was used to assess statistical differences between transcript levels.

### 2.1.3.6. Plasmid construction

All plasmids used for recombinant protein expression are listed in Appendix Table 4. The putative DNA binding domains of *CaCph2p* (amino acids 197-302), *CaHms1p* (amino acids 463-686), *CaTye7p* (amino acids 121-269), *CpHms1p* (amino acids 486-659) and *AfSrbAp* (amino acids 145-266) were amplified from genomic DNA of each species and introduced into plasmids pLIC-H3 [88] and pbRZ75 [89] (both derivatives of pET28b). These plasmids were designed to produce recombinant N-terminal 6×His or 6×His-MBP (maltose binding protein) tagged proteins.

Chimeric proteins were constructed by (1) replacing residues 211-232 from *CaCph2p* by residues 489-510 from *CaHms1p* to generate chimeric helix 1 protein; (2) replacing residues 255-281 from *CaCph2p* by residues 625-651 from *CaHms1p* to generate chimeric helix 2 protein; (3) replacing residues 211-255 from *CaCph2p* by residues 489-629 from *CaHms1p* to generate chimeric helix1-loop protein; and (4) replacing the loop of *CaCph2p* (236-255) by the loop of *CaHms1p* (506-625) to generate the chimeric loop protein.

The DNA fragments encoding the reconstructed ancestral proteins as well as the DNA binding domain of the human SREBP1p (aa 317-400) were generated by gene synthesis (Invitrogen GeneArt Gene Synthesis). These fragments included restriction sites for cloning into pLIC-H3 [88].

Restriction enzyme pairs *SmaI/XhoI* and *NheI/XhoI* (New England Biolabs) were used to digest the DNA fragments and insert them into pLIC-H3 and pbRZ75, respectively. Digestion reactions were performed according to manufacturer's indications. Correct digestion was verified by checking DNA sizes of digested and undigested fragments on gel electrophoresis. Ligation reactions were performed using T4 DNA ligase (New England Biolabs) as well as digested plasmid and insert in a 1:3 ratio, the reactions were incubated for 2 hours at RT followed by 10 minutes at 65°C to achieve inactivation of the T4 ligase. Ligation reactions were then transformed into competent *E. coli* DH5α cells. Briefly, competent cells were thawed on ice, 50µl transferred in a 1.5 ml tube and 5 µl of ligation reaction was added. Cells were incubated 30 min on ice followed by heat shock 45 sec at 42°C. Cells were incubated on ice for 1 min. Then 1 ml of LB medium was added, and cells



were incubated for 1 h at 30°C shaking at 200 rpm to recover. After recovering cells were centrifuged for 2 min at 5000 rpm, resuspended in 50 µl of LB medium and plated in LB agar plates supplemented with 100µg/ml kanamycin (both pLIC-H3 and pbRZ75 harbour a kanamycin resistance cassette). Incorporation of insert-containing plasmids in the growing colonies was checked by colony PCR and the correct insertion, as well as the insert sequence, was verified by Sanger sequencing (Eurofins Genomics). Verified colonies were then used to set up liquid overnight cultures in LB medium for further plasmid isolation (using NucleoSpin plasmid purification kit, Macherey-Nagel) and transformation into *E. coli* BL21 cells for protein purification.

#### **2.1.3.7. Protein purification**

*E. coli* BL21 was used as the host of the expression plasmids. For recombinant protein overexpression, 250 ml of bacterial cell cultures were grown at 37°C on a shaker to an OD<sub>600</sub> of approximately 0.8 and protein production was induced with 0.5 mM IPTG. Cultures were grown further for 3 hours at 30°C after induction and then pelleted down at 5000 rpm for 5 min. All following steps were performed at 4°C. Cells were resuspended in 15 ml of cold lysis buffer (containing freshly added protease inhibitor cocktail without EDTA (2 tablets, Roche), lysozyme (0.5 mg/ml), β-mercaptoethanol (5 mM) and PMSF (1 mM)) and sonicated for 15-30 minutes (rounds of 1-2 minutes with 5 minutes of incubation on ice in between to prevent the sample from warming up and damaging the proteins). Cells were centrifuged down for 20 minutes at 5000 rpm to discard cell debris. His-tagged proteins were affinity purified from the lysate using 1 ml of previously equilibrated Ni-NTA agarose beads (Qiagen) during 1 hour on a shaker. After a series of washes (4 times during 10 minutes each with cold wash buffer containing 1M PMSF, centrifuging 1 minute at 1000 rpm in between washes) the proteins were eluted from the beads using 2 ml of cold elution buffer (containing freshly added protease inhibitor cocktail without EDTA (1 tablet) and PMSF (1 mM)) and incubating for 1 hour on a shaker. The solution containing the eluted protein was then transferred to Amicon Ultra-15 centrifugal filters (Merck) (10 or 30K membranes depending on protein size) to exchange buffer and concentrate the proteins in storage buffer. Protein concentration was

estimated in Rothi® blue (Carl Roth, Germany) stained gels using known amounts of bovine serum albumin as standards.

Buffers and solutions for protein purification are listed below:

Lysis Buffer (pH 8.0)

50 mM      NaH<sub>2</sub>PO<sub>4</sub>  
600 mM     NaCl  
10 mM      Imidazole

Washing Buffer (pH 8.0)

50 mM      NaH<sub>2</sub>PO<sub>4</sub>  
600 mM     NaCl  
30 mM      Imidazole

Elution Buffer (pH 8.0)

50 mM      NaH<sub>2</sub>PO<sub>4</sub>  
400 mM     NaCl  
250 mM     Imidazole

Storage Buffer (pH 7-9 depending on each protein's pI)

40 mM      Tris base  
400 mM     NaCl  
5 mM        DTT

### **2.1.3.8. Electrophoretic mobility shift assays and competition assays**

Complementary pairs of 30-bp long oligonucleotides were mixed at equimolar concentrations and annealed to obtain 50µM double-stranded DNA probes. Labelling mix was prepared as follows:

Radiolabelling reaction

2,5µL	10× PNK buffer (New England Biolabs)
2µL	1µM annealed oligonucleotides
17,5µL	H <sub>2</sub> O
2µL	<sup>32</sup> P radioactive isotope
1µL	T4 polynucleotide kinase (New England Biolabs)

Tubes containing the radiolabelling reaction were incubated at 37°C for 20 minutes, then 65°C for 15 minutes and finally 95°C for 5 minutes. The radiolabelled mix was filtered through G25 columns (BioRad) by centrifugation 4 minutes at 1000 g before use. Binding conditions were 22.5 % 5× minimal buffer, 5 mM MgCl<sub>2</sub>, 1 mM DTT, 0.1 % NP40, 1 mM ZnSO<sub>4</sub> and 1 µg/µl BSA. Proteins were diluted in 1× minimal buffer to the desired concentration. In the case of competition assays, increasing concentrations of unlabelled DNA competitor DNA were added to the mix prior to the addition of the protein. Binding reactions (containing the protein and the radiolabelled DNA) were incubated for 15 minutes at room temperature and then run on gels composed of 6 % polyacrylamide, 0.5× Tris-Glycine-EDTA and 2.5% glycerol. Gels were run at 130 V for 90 minutes, dried and exposed to phosphor imaging screens (GE Healthcare) for at least 5 hours. Imaging was carried out with Typhoon imager (FLA 7000).

5× Minimal buffer

100mM	Tris pH 8.0
250mM	NaCl
25%	Glycerol

5× Tris-Glycine-EDTA buffer

125 mM	Tris pH 8.5
1M	Glycine
0.5 mM	EDTA

### **2.1.3.9. Ancestral protein reconstruction**

Phylobot [90] was used to infer ancestral protein sequences at the two specific nodes of the phylogenetic tree (Figure 2). The model used to infer the tree was PROTGAMMALG (tested for best-fit model with ProtTest) for all cases. The reconstructed protein sequence for Anc4 exhibited low levels of uncertainty. For Anc5, due to higher levels of sequence uncertainty, I carried out two reconstructions: (1) Using Pho4 as the only outsider sequence; and (2) including mouse and human SREBP sequences in addition to Pho4 as outsider sequences. Two alignment models, MUSCLE and msaprobs, were considered. Eight versions of Anc5, which differed from one another at residues in the first helix, were synthesised and their overall ability to bind to DNA was evaluated by EMSAs. Due to the high variability in the loop segment, this particular portion of the DNA binding domain could be neither properly aligned nor reconstructed. Given its short size (55 amino acid residues), the corresponding amino acid sequence of *CaCph2p* was used to fill the loop segment in all ancestral proteins.



## 2.1.4. Results

### 2.1.4.1. The SREBP family of transcription regulators in fungi

In addition to higher eukaryotes, the SREBP family of transcription regulators is widely represented in the fungal kingdom. A distinctive feature of this family—which distinguishes them from other bHLH proteins—is the presence of a tyrosine residue instead of an arginine in the first helix of the DNA binding domain (Figure 3A).

Using the presence of the tyrosine residue as hallmark to identify proteins that belong to the SREBP family, I retrieved the amino acid sequence of around 200 fungal SREBPs. An initial comparison of the protein sequences revealed a variety of sizes between SREBPs from different species. Several attempts to align the sequence of the full-length proteins indicated that sequence conservation between fungal members of the SREBP family was mainly restricted to the DNA binding domain. Therefore, in order to build a robust phylogeny, I built an alignment based only on the amino acid sequence of the DNA binding domain of the retrieved fungal SREBPs, which then served to generate a phylogenetic tree (see Appendix Table 5 and Appendix Figure 22 for extended phylogeny; models and computational procedures used for phylogenetic reconstruction are described under materials and methods section 2.1.3.2.). The resulting phylogenetic tree points to the existence of several sub-groups, which cluster separately, within the fungal SREBPs.

The two defining features of SREBPs are the tyrosine at position 12 in the first helix and the transmembrane domain that keeps the inactive form of the protein membrane-bound until regulated proteolysis takes place. The SREBPs from the model fission yeast *Schizosaccharomyces pombe* as well as the human SREBP among others possess transmembrane domains [56,91] that attach the proteins to the ER membrane. However, other SREBPs, such as those in the ascomycete model yeast *Saccharomyces cerevisiae* and other species of the *Candida* clade, lack such feature [92]. Therefore, in order to determine if the presence of such domain is widespread across the fungal SREBP phylogeny I scanned the full length of each protein for the presence of a putative transmembrane sequence(s), which are shown as blue dots in the phylogenetic tree of Figure 3C. I found

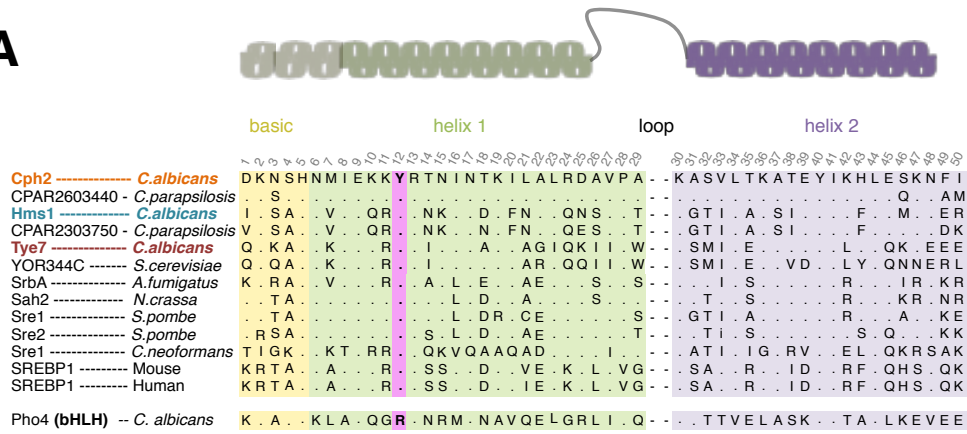
that the majority of the fungal sequences scanned (70.3%) harboured at least one putative transmembrane domain. The proteins with no transmembrane domains were found clustering together in branches of the phylogenetic tree.

One might expect that proteins from the same organism would cluster together if they originated by gene duplication after a recent speciation event. This is for instance what happens in the case of the fission yeast *S. pombe*: the organisms' SREBPs Sre1p and Sre2p cluster very close together in the same branch. Strikingly, in other cases they appear to be distributed in different branches of the phylogenetic tree far away from each other. This is the case of the ascomycete yeasts (i.e. the Saccharomycotina) where the SREBP members are distributed in different branches (which are labelled 1, 2 and 3 in the phylogenetic tree of Figure 3C). This separation in three distinct clusters of SREBPs from the same species is also supported by other independent, large-scale reconstructions of fungal gene families such as Fungal Orthogroups [93]. As in most other organisms (like humans, *D. melanogaster* or *C. elegans*), the majority of species in the Saccharomycotina encode no more than one or two SREBPs. A few species in the *Candida* clade, however, encode three SREBPs (namely Cph2p, Hms1p and Tye7p) (Figure 3B).

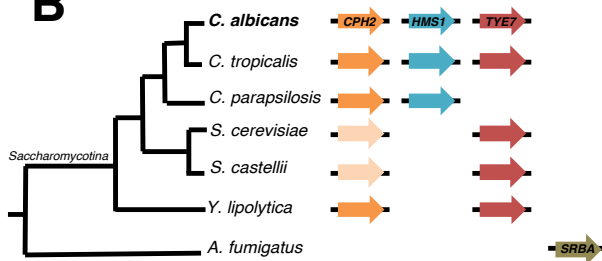
Remarkably, each one of these three proteins lies in a different branch of the phylogenetic tree (Figure 3C) indicating that the *Candida* proteins span a considerable distance in the phylogeny. Within the Saccharomycotina, only branch 1 (which includes the *Candida albicans* Cph2 protein) contains putative transmembrane domains whereas the other two groups (branches 2 and 3) do not. A sub-cluster of SREBPs in *Aspergillus* spp. is the only other group outside the Saccharomycotina that appears to lack transmembrane domains. In this thesis, I will focus on characterising SREBPs that are representative of branches 1, 2 and 3 of the fungal SREBP phylogenetic tree (Figure 3).

## 2.1 Diversification of DNA binding preferences in a conserved family of transcription regulators

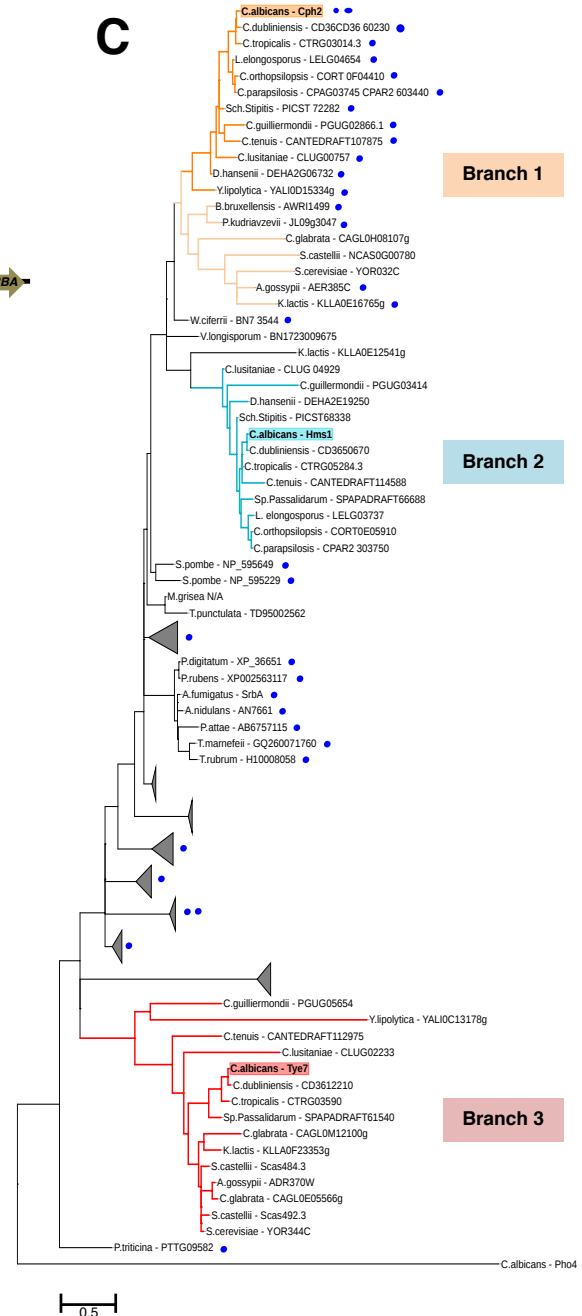
**A**



**B**



**C**





**Figure 3 Phylogenetic reconstruction of the SREBP fungal family.**

(A) Protein alignment of the SREBPs' DNA binding domain. The SREBPs are basic helix-loop-helix transcription regulators (linear structure drawn on top). Amino acids within the yellow shade belong to the basic region; in green the first helix; and in purple the second helix. The loop region is variable in sequence and length. The tyrosine residue that is the hallmark of the SREBP family is highlighted in pink. Most bHLH regulators (e.g. Pho4) have a conserved arginine instead of the tyrosine. Dots in the alignment represent the presence of the same amino acid as in the top sequence (B) Cladogram depicting the phylogenetic relationship among extant ascomycete yeasts (Saccharomycotina). The genes encoding SREBPs in each species are represented by coloured arrows. Same shade of colour portrays inferred orthology. The colours orange, blue and red represent SREBPs belonging to branches 1, 2 and 3 (respectively) of the phylogenetic reconstruction shown above (C) Phylogenetic tree of the SREBP fungal family. An alignment of the amino acid sequences of the DNA binding domain of 198 fungal SREBPs was employed to build the phylogeny. Redundant or uninformative sequences are omitted from the tree (see Appendix Figure 22 for an expanded version of the phylogenetic tree). Blue dots indicate the presence or absence and number of transmembrane domains in each SREBP. Highlighted in the tree are three branches I will focus on in this thesis. Each branch is represented by a different *C. albicans* SREBP: Cph2p (orange), Hms1p (cyan) and Tye7p (red)

#### 2.1.4.2. Different branches of the SREBP phylogenetic tree show distinct DNA binding preferences

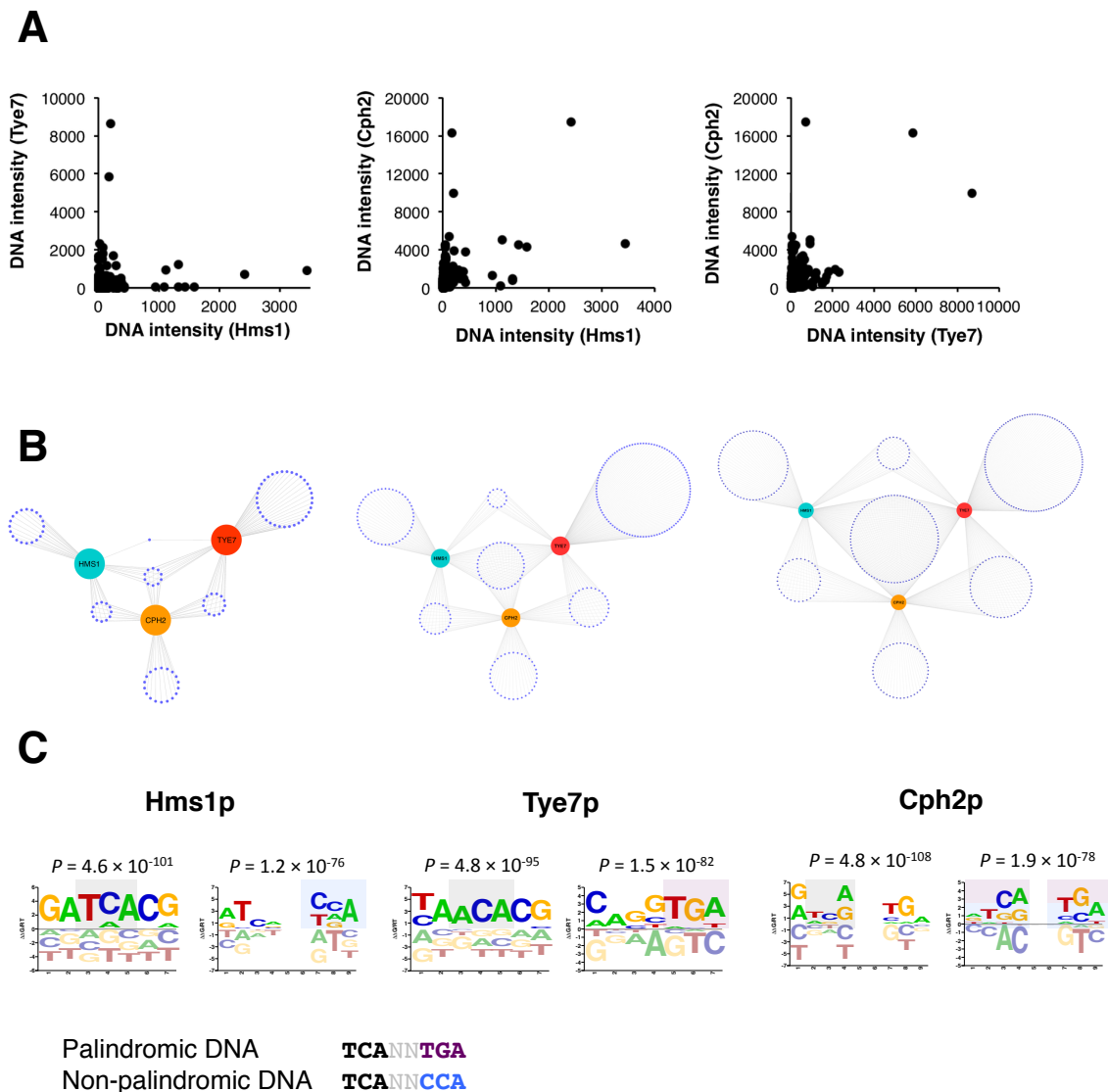
SREBPs are members of the bHLH family and like all these proteins, they are known to recognise and bind to variants of the palindromic E-box DNA sequence [54,56] as well as to a non-palindromic sequence [59]. Early classic *in vitro* binding assays confirmed the dual binding ability of the archetype and most studied SREBP, the human SREBP1 [58]. The protein, however, appears to preferentially bind *in vivo* to the same E-box DNA sequence all other bHLH bind to, as shown by more recent chromatin immunoprecipitation (ChIP) experiments [56]. Whether binding to this alternative DNA sequence happens only *in vitro* or also takes place *in vivo* remains unclear. In this chapter I sought to establish and evaluate the intrinsic DNA binding preferences of the three selected branches of fungal SREBPs, represented by *Candida albicans*' proteins Cph2p, Hms1p and Tye7p.

As a first systematic approach to establish the DNA binding preferences of the three representative proteins, I employed MITOMI [71], a large-scale microfluidic-based approach that enables the *in vitro* measurement of protein-DNA interactions at equilibrium between transcription regulators and a comprehensive library of oligonucleotides. In each experiment, I assessed binding to a set of 740 double-stranded 70-nt oligos designed so that all possible 8-mers were represented in the library. The binding was then quantified by measuring the ratio of fluorescence emitted from labelled DNA binding to surface-immobilised labelled transcription regulators [71]. Pair-wise comparisons of the oligos bound by the proteins indicate that the oligonucleotide binding patterns observed for Hms1p and Tye7p are largely orthogonal whereas the other two pairs (Cph2p - Hms1p and Cph2p - Tye7p) display some degree of overlapping binding preferences (Figure 4A).

Examining the sets of oligonucleotides bound by each protein with the highest scores shows that, to a significant extent, they bind different sets of DNA sequences (Figure 4B) suggesting they have their own unique affinity for DNA. Comparisons of the top oligonucleotides bound and shared by the proteins also reveals that *Ca*Hms1p and

*CaTye7p* are the most distant from each other whereas *CaCph2p* appears as an intermediate (i.e. it shares a similar number of bound oligomers with *CaHms1p* and with *CaTye7p*). A similar pattern can be observed if the top 30% or even the top 50% of oligomers are considered (Figure 4B).

I next used MatrixREDUCE [79] to analyse the binding intensities from all oligonucleotides and find DNA motifs overrepresented in the MITOMI data. I run several iterations of the software varying the following parameters: length of the motif to be found and including or not a 2-nucleotide spacer between the two half sites of the motif. I then compiled and ranked all the motifs generated by MatrixREDUCE ( $P < 1 \times 10^{-10}$ ) according to their  $r^2$  and  $P$ -values (a full list containing all the motifs can be found in Appendix Table 6). Representatives of the top scoring motifs for each protein are shown in Figure 4C. To a large extent these motifs represent either the palindromic E-box variant 5'-ATCANNTGA-3' or the non-palindromic sequence 5'-ATCANNCCA-3' (or their predicted half-sites). Importantly, and in agreement with the results above, the motifs derived for each of the proteins differ from one another suggesting each protein has a distinct preference for DNA. Consistent with the pattern of overlap in bound oligonucleotides, the *CaHms1p*- and *CaTye7p*-derived motifs were the least similar to each other. *CaCph2p*, on the other hand, appeared as an intermediate that could recognise both types of motifs.



**Figure 4** *In vitro* DNA binding preferences of the *C. albicans* SREBPs Hms1p, Cph2p and Tye7p in MITOMI.

(A) Each transcription regulator was evaluated for binding to a set of double stranded 740 70-nt oligos designed so that all possible 8-mers were represented. Binding was quantified by measuring the ratio of fluorescence emitted from labelled DNA bound to surface-immobilised labelled regulators. The intensity of binding (DNA intensity) to each oligonucleotide is plotted for each of the three proteins (one dot represents one oligonucleotide). Shown are pairwise comparisons among the regulators. Notice the orthogonal relationship among pairs, particularly between Tye7p and Hms1p. The MITOMI data was derived from two biological replicates. (B) Distribution of top scoring oligonucleotides bound by the *C. albicans* SREBPs Hms1p, Cph2p and Tye7p in MITOMI. The diagrams display the top 10% (left), 30% (middle) and 50% (right) scoring

oligonucleotides are shown. Each purple dot represents one oligonucleotide. The distances separating the three proteins (Hms1p in cyan, Cph2p in orange and Tye7p in red) are inversely proportional to the number of shared oligonucleotides. (C) MITOMI-derived motifs for the three *C. albicans* SREBPs. Two motifs resembling each of the half sites for either the palindromic (purple) or the non-palindromic.

As a complementary approach to determine the DNA binding preferences of the proteins, I analysed *in vivo* genome-wide chromatin immunoprecipitation (ChIP) data. While such datasets have been generated for all three proteins in *C. albicans*, a clear DNA motif could be derived only for Tye7p and Hms1p [86,94]. Thus, I performed a ChIP-Seq experiment of the third SREBP in *C. albicans*, Cph2p. The putative DNA binding domain of Cph2p is located at the N-terminal portion of the protein and is followed by two transmembrane domains that anchor Cph2p to an intracellular membrane. An unidentified signal is thought to trigger the cleavage and release of the N-terminal portion of Cph2p from the membrane and its posterior shuttle to the nucleus [95]. To circumvent the need for an “activating” signal, I generated a *C. albicans* strain encoding a truncated version of the protein, which ends immediately before the transmembrane domain and is Myc-tagged at this new C-terminus (Figure 5A) making the resulting construct constitutively active.

The ChIP-Seq experiment conducted with this strain identified 14 high-confidence binding regions located within intergenic sequences (Figure 5B) shown in Table 1.

**Table 1** List of DNA regions occupied by *CaCph2p* based on ChIP-seq experiments.

List of 14 bona fide chromosome locations where an enrichment peak could be observed indicating Cph2 binding.

ORF	Gene name	Peak value -log <sub>10</sub> P-value	Fold enrichment	Motif score
orf19.921	<i>HMS1</i>	61,27	5,17	4,1
orf19.3794	<i>CSR1</i>	42,25	4,49	3,9
orf19.3549	<i>CDC21</i>	34,07	3,45	3,2
orf19.3337		26,05	2,94	3,4
orf19.2333		50,89	4,16	4,8
orf19.4941	<i>TYE7</i>	20,14	2,97	3,1
orf19.6736		24,50	2,48	2,8
orf19.4167		35,59	3,88	3,3
orf19.2723	<i>HIT1</i>	32,56	3,24	3,2
orf19.4309	<i>GRP2</i>	25,16	3,02	3,8
orf19.7502		25,60	2,82	4,6
orf19.3261		17,48	2,29	4,2
orf19.7561	<i>DEF1</i>	32,49	3,29	-
orf19.610	<i>EFG1</i>	25,80	3,02	3,1

A clear DNA motif (shown in Figure 5C) could be derived from this *in vivo* Cph2p occupancy data set. The derived motif represents a bona fide binding sequence mainly because of two reasons: First, I tested the purified *Ca*Cph2p protein in gel shift assays and found that it was able to “shift” the DNA fragment containing an instance of the ChIP-derived motif; and second, the introduction of point mutations in the putative binding site strongly impaired binding of the protein (Figure 5D).

Since manual, detailed examination of the DNA regions occupied by the Cph2p protein *in vivo* produced no evidence of a composite motif (*i.e.* a second half-site), I considered the possibility that co-factors could contribute to this protein’s binding *in vivo*. Indeed, DNA motif searches in our *Ca*Cph2p ChIP dataset revealed the co-occurrence of a DNA sequence that closely resembles the DNA motif recognised by the *C. albicans* regulator Efg1p (Figure 5E).

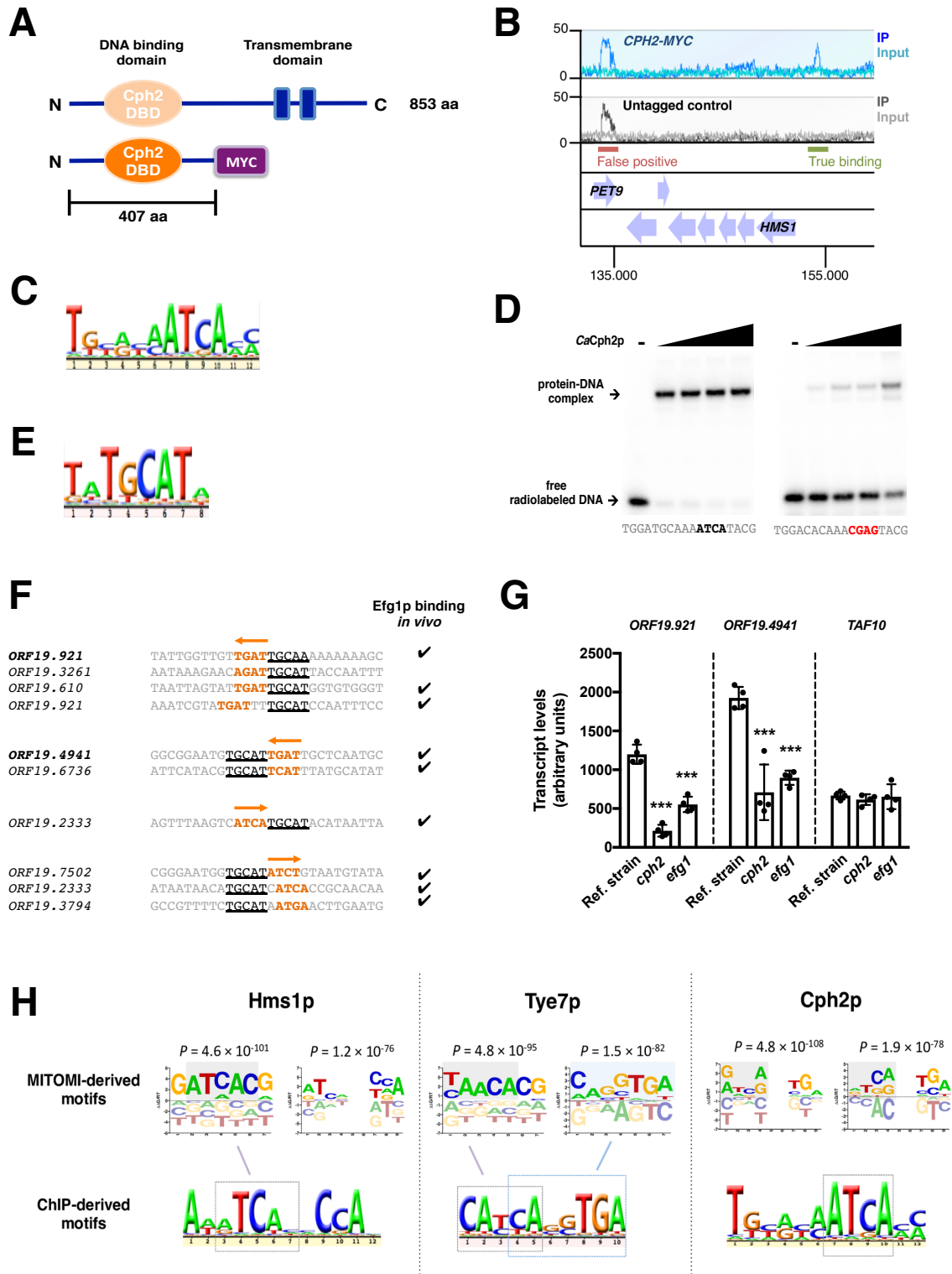
Consistent with this result, I found that a significant proportion of these sites are occupied by Efg1p *in vivo* ( $P = 2.6 \times 10^{-5}$ ) (Figure 5F; [85]). These observations suggested that the Cph2p and Efg1p proteins might interact *in vivo* (either by binding cooperatively or by competing with each other for binding) to regulate a subset of target promoters. Consistent with this hypothesis, we found that the expression of two direct targets of regulation (*ORF19.921* and *ORF19.4941*; each containing Cph2p- and Efg1p-binding sites in their putative promoter regions as indicated in Figure 5G) is dependent, at least in part, on *CPH2* and *EFG1*.

As illustrated in Figure 5H the MITOMI- and ChIP-derived DNA motifs were, to a significant extent, congruent with each other and revealed distinct DNA binding preferences for each protein. *Ca*Tye7p bound to a singular variant of the palindromic E-box motif that consisted of an extended left half-site (5'-CATCA-3') and a three-nucleotide right half-site (5'-TGA-3'). While the MITOMI analysis was unable to capture the full 10-nucleotide sequence in a single motif, the two separate *Ca*Tye7p MITOMI motifs could explain the full-length sequence when combined. *Ca*Hms1p bound to an alternative, non-palindromic sequence (5'-ATCANNCCA-3'). In this case, the Hms1p MITOMI motif that included a 2-nucleotide spacer was in very close agreement with the full *Ca*Hms1p ChIP motif. In contrast to Hms1p and *Ca*Tye7p, the *Ca*Cph2p MITOMI motifs suggested that, at least *in vitro*, this protein may recognise both (5'-ATCANNTGA-3') and (5'-ATCANNCCA-



3') sequences. The *CaCph2p* ChIP motif, on the other hand, indicated that, *in vivo*, this protein might simply bind to the left portion of either sequence (5'-A/CATCA-3').

Taken together, these results show that the three analysed proteins, which are representatives of three branches of the fungal SREBP phylogeny, have different DNA binding preferences. The data indicate that, on the one hand, *CaTye7p* binds to the canonical E-box motif whereas *CaHms1p* binds a non-palindromic DNA sequence; both of these motifs are composed of two half-sites separated by a 2-nucleotide spacer. On the other hand, *CaCph2p* appears *by itself* to recognise a shorter DNA sequence (composed of only one half-site) but may operate in concert with other co-factors such as *Efg1p*.



**Figure 5** *CaCph2* binds to a single half-site in cooperation with other co-factors.

(A) Schematic representation of the *CaCph2*-MYC construct used for ChIP. The MYC-tag was introduced immediately after the DNA binding domain in order to circumvent the need for an activating signal that would release Cph2 from the membrane and to have a

constitutively active and tagged version of the protein. **(B)** Identification of *C. albicans* genome regions bound *in vivo* by Cph2p. ChIP-Seq was carried out with a strain expressing a constitutively active MYC-tagged Cph2p (blue track) and an untagged control strain (grey track). Shown is a 25 kb region of chromosome 5 where a binding event (upstream of the *HMS1* gene) can be visualised and distinguished from a false positive (in the *PET9* gene). 14 binding regions were consistent across replicates and therefore used for the DNA motif analysis displayed in C and E **(C)** *CaCph2p* motif derived from sequences occupied *in vivo* which are listed in Table 1. Unlike the other two representative proteins, *CaHms1p* and *CaTye7p*, the motif is composed of a “single half-site” instead of two “half-sites” separated by a spacer. **(D)** Gel shift assay probing the binding of the purified *CaCph2* protein (0, 0.0012, 0.006, 0.02, 0.1 and 0.4 nM) to the indicated P<sup>32</sup>-labeled DNA fragment (taken from the upstream intergenic region of *ORF19.921*) which harbours an instance of the putative Cph2p motif (in black). DNA binding is strongly reduced when point mutations (in red) are introduced in the binding site. **(E)** Putative Efg1p motif. Search iterations revealed the presence of a second overrepresented motif in the set of DNA sequences occupied *in vivo* by Cph2p. This motif closely resembles the motif derived for Efg1p in [85]. **(F)** *CPH2* and *EFG1* co-regulate the expression of target genes. Distribution of Cph2p and Efg1p DNA binding sites in a subset of the sequences occupied by Cph2p. Notice the co-occurrence of the Cph2p and Efg1p binding sequences. Putative Cph2p sites are shown in orange whereas the predicted Efg1p sites are underlined. Check marks to the right indicate whether Efg1p has been found to bind *in vivo* to the respective target promoter [85]. **(G)** Cph2p and Efg1p function together to regulate some of their target genes. Total RNA was prepared from wild-type, *cph2* and *efg1* deletion mutant strains after a 24-hour incubation under anaerobic conditions (37°C). *ORF19.921*, *ORF19.4941* and *TAF10* (control) transcript levels were determined by quantitative real-time PCR. Plotted are the mean and SD of four biological replicates. Asterisks denote statistically significant differences compared to the reference strain. **(H)** DNA motifs preferred by the *C. albicans* SREBPs Hms1p, Cph2p and Tye7p. Shown are representative motifs derived from the MITOMI (top) and ChIP (bottom) datasets. Highlighted is the likely correspondence between MITOMI and ChIP motifs. Notice that the Hms1p MITOMI motif that includes a 2-nucleotide spacer is a very close match to the full Hms1p ChIP motif. The Cph2p MITOMI motifs imply that, *in vitro*, this protein might recognise both ATCANNTGA and ATCANNCCA sequences whereas the ChIP motif suggests binding to the left half-site only (ATCA). The ChIP-derived motif for Cph2 was derived from data included in this thesis. The ChIP motifs for Tye7p and Hms1p were derived from references [94] and [86], respectively.

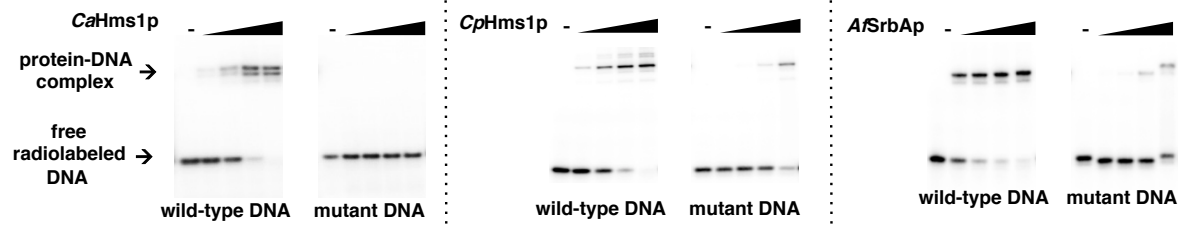
### 2.1.4.3. Several fungal SREBPs show higher affinity for a non-canonical motif

MITOMI and ChIP data clearly indicate that the Hms1p protein from *C. albicans* binds a non-palindromic DNA sequence, which is unusual because most bHLH proteins bind a palindromic DNA motif. We wondered whether this unusual binding preference was exclusive to this protein in this species or extended to other SREBPs. To address this question, in addition to *CaHms1p*, we purified the putative DNA binding domains of the *C. parapsilosis* Hms1p (CPAR2\_303750) and the *A. fumigatus* SrbAp (Afu2g01260) proteins and carried out electrophoretic mobility gel shift assays (EMSAs). As shown in Figure 6A, all three proteins bound to a DNA fragment harbouring an instance of the non-palindromic motif. This binding was specific to the analysed DNA sequence because point mutations introduced in the putative binding site abolished or severely impaired binding.

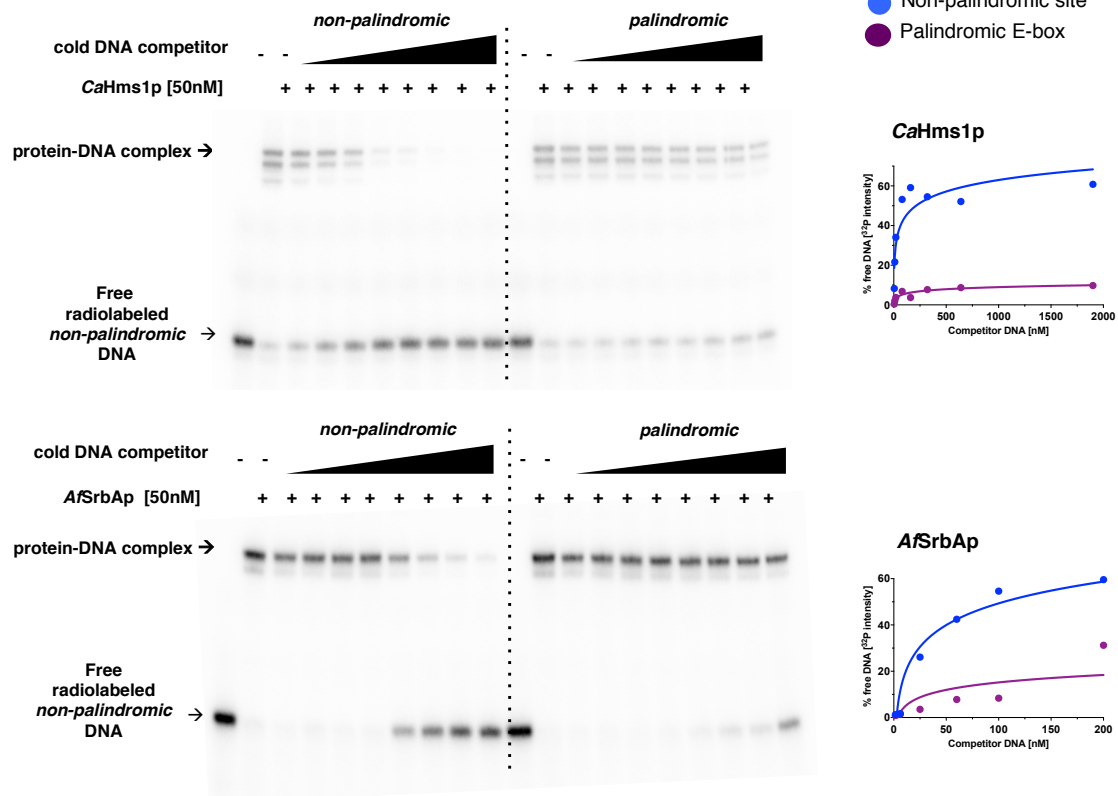
I next wanted to determine whether the proteins were able to discriminate between non-palindromic and canonical E-box sequences. For this, I carried out competition binding assays in which we incubated *CaHms1p* or *AfSrbAp* with a <sup>32</sup>P-labeled DNA fragment carrying the non-palindromic sequence. Upon binding, I competed the reactions with unlabelled DNA fragments harbouring either the non-palindromic site or the canonical E-box sequence (Figure 6B). In the case of both proteins, *CaHms1p* and *AfSrbAp*, the former DNA fragment was a much stronger competitor compared to the latter indicating that the proteins exhibit a marked preference for the non-palindromic sequence versus the canonical E-box motif.

**A**

wild-type `ACAAAAAAAAA`**TCA**GGCCA**TTT**GTAACT  
 mutant `ACAAAAAAAAA`**TTC**TAAA**TTT**GTAACT



**B**

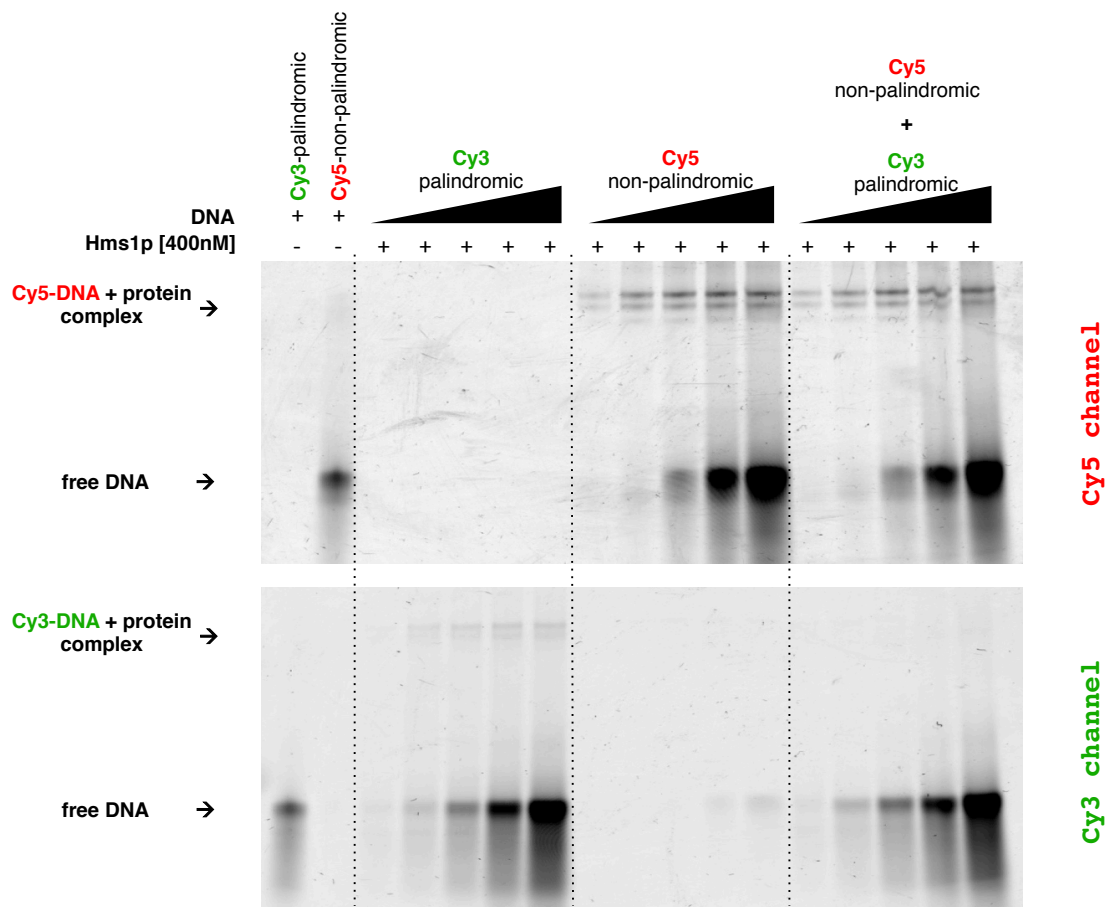


**Figure 6 Several fungal SREBPs exhibit an intrinsic ability to bind a non-palindromic DNA motif.**

(A) Gel shift assays showing binding of three fungal SREBPs to a non-palindromic DNA sequence. P<sup>32</sup>-labeled DNA fragments containing the wild-type or mutant non-palindromic binding site were incubated with increasing concentrations (0, 1.56, 6.25, 25 and 100 nM) of the purified DNA binding domain of *C. albicans* Hms1p (left), *C. parapsilosis* Hms1p (centre) or *A. fumigatus* SrbAp (right) for 90 min at room temperature in standard EMSA buffer and resolved in 6% polyacrylamide gels run with 0.5× TGE. Point

mutations introduced in the DNA binding site are shown in red (**B**) At least two branches of fungal SREBPs bind preferentially to a non-palindromic site over the canonical E-box. Competition experiments to determine the DNA binding preferences of the SREBPs *CaHms1p* (top) and *AfSrbAp* (bottom). The purified DNA binding domain of either protein was incubated with a P<sup>32</sup>-labeled DNA fragment containing the non-palindromic binding site. Upon binding, increasing concentrations of unlabelled competitor DNA fragments harbouring either palindromic E-box or non-palindromic binding sites were added to the reactions, and the mixtures were then resolved by polyacrylamide gel electrophoresis.

Even though competition gel shift assays are a useful method for quantification of protein-DNA binding preferences, they allow detection of only one of the evaluated sequences because only one of the two DNA fragments can be radioactively marked and therefore detected. In order to overcome this limitation and be able to detect both evaluated sequences I used fluorescently labelled DNA fragments (with Cy3 or Cy5) harbouring either the palindromic or the non-palindromic motifs. I incubated the *CaHms1p* together with equimolar concentrations of both fluorescently labelled DNA sequences. Hence, in this experimental setup, both sequences were competing for the same pool of protein. In agreement with previous observations, the non-palindromic DNA bound to a much higher extent to the *CaHms1* protein than the palindromic sequence (Figure 7). Taken together, these results demonstrate that several SREBPs bind preferentially to a non-palindromic DNA sequence over the canonical E-box motif. This property is shared by SREBPs both within and outside the ascomycete yeasts and it is additionally an intrinsic property of the proteins since the presence of extra co-factors or other proteins was required for recapitulating the *in vivo* preference.



**Figure 7 Gel shift assay with fluorescently labelled DNA sequences competing for the same pool of *CaHms1* protein.**

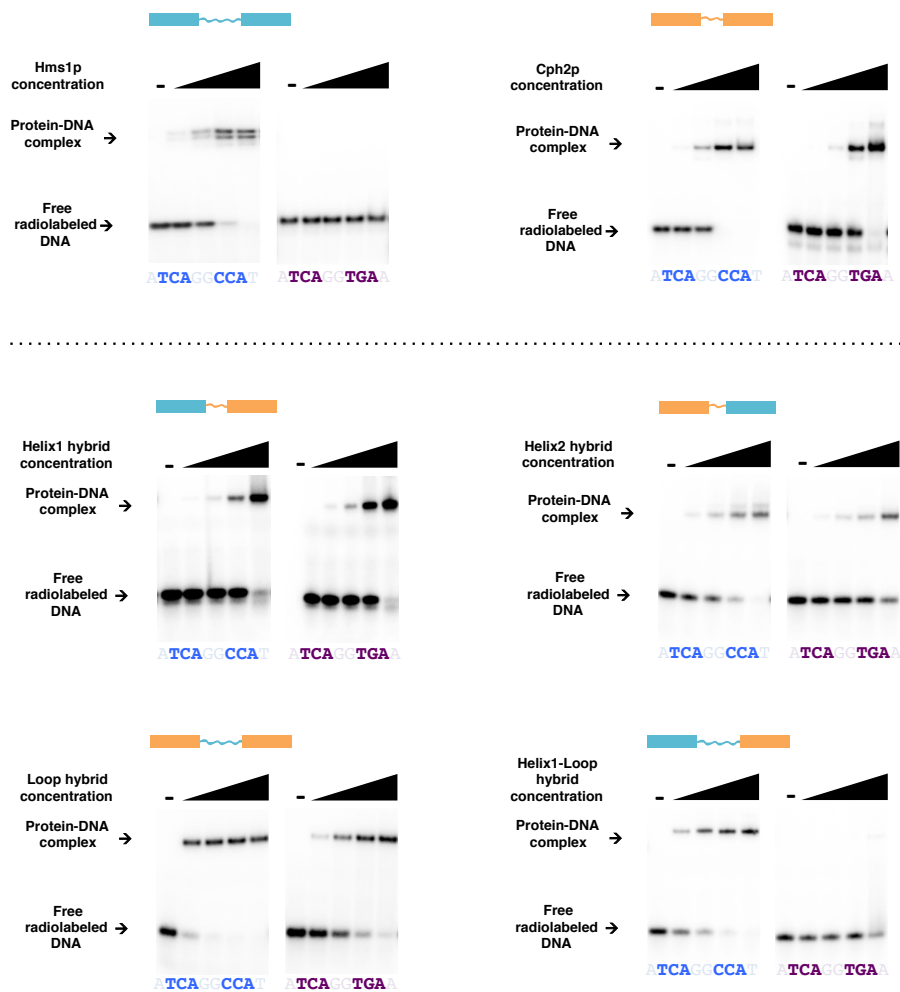
Increasing amounts (0.56, 1.67, 5, 15 and 45 ng) of Cy5-labeled non-palindromic and Cy3-labeled palindromic DNA fragments, alone or together, were incubated with purified Hms1 protein, and resolved in 6% polyacrylamide gels run with  $0.5 \times$  TGE. The images of the gels taken in the Cy5 and Cy3 channels are shown at the top and bottom, respectively. Notice the strong preference ( $>10$ -fold) of the protein for the Cy5-labeled non-palindromic site.



#### **2.1.4.4. Tracing the structure of the DNA binding domain that confers specificity to *CaHms1p***

As shown above, the purified DNA binding domain of the *CaHms1* protein binds to a DNA sequence different from the SREBP's canonical E-box motif. Such binding pattern is concordant with the fact that the purified DBD of the protein displays a strong preference for its cognate DNA binding sequence over the canonical E-box motif in competitive EMSAs (Figure 6B). Since the *CaHms1p*'s ability to discriminate between the DNA sequences evaluated here appears to be an intrinsic property of the protein I sought to determine what portion(s) of its DBD were responsible for conferring this ability.

I constructed several chimeric proteins by exchanging one of three portions (first helix, loop region or second helix) of the DNA binding domains of *CaHms1p* and *CaCph2p* (the latter protein displayed little, if any, ability to discriminate *in vitro* between the two DNA sequences evaluated here). I then employed EMSAs to probe each chimeric protein for their ability to bind DNA fragments harbouring either the cognate *HMS1* binding site or the canonical E-box sequence. We found that a chimeric protein consisting of the first helix and the loop from *CaHms1p* and the second helix from *CaCph2p* recapitulated almost completely the ability to discriminate between the two DNA sequences as the native *CaHms1p* (Figure 8). Chimeric proteins containing only the first helix or only the loop from *CaHms1p* showed no discrimination. Therefore, the conclusion withdrawn from these experiments is that that residues within the first helix combined with residues in the loop region of *CaHms1p* confer the ability to bind specifically to the non-palindromic sequence.

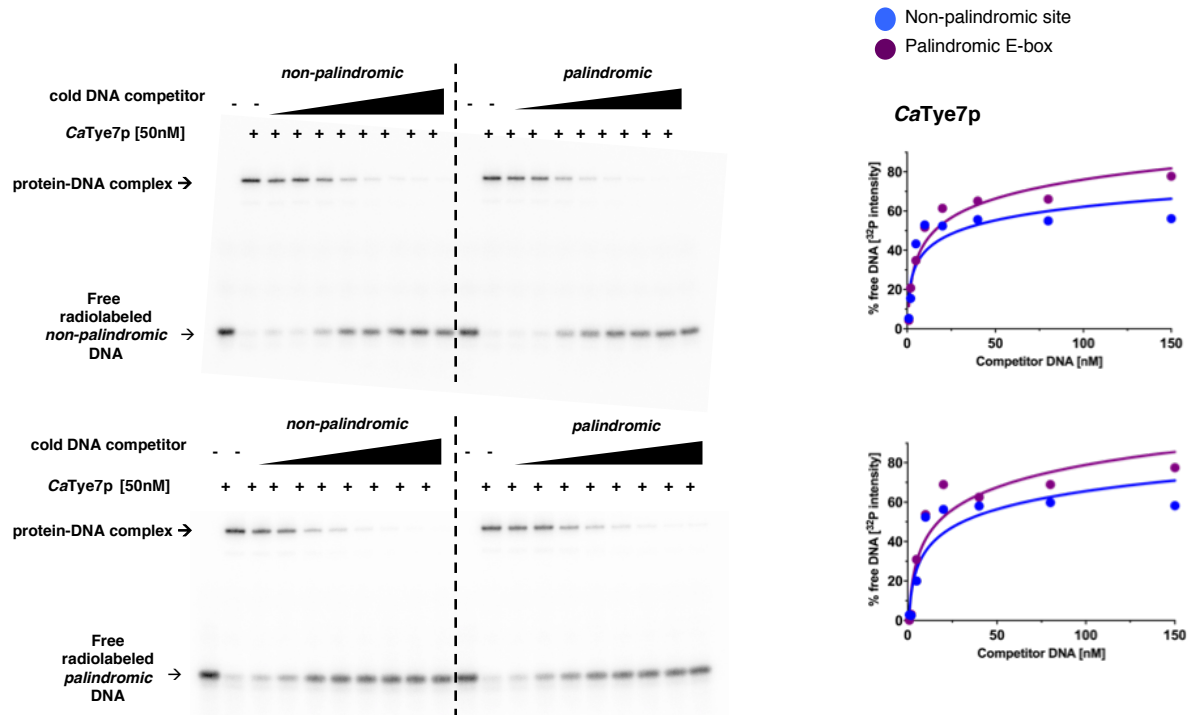


**Figure 8** The DNA binding specificity in of *CaHms1p* is conferred by residues in the first helix and loop region of its DNA binding domain.

Gel shift assays to determine the protein regions of the *C. albicans* SREBP Hms1p (cyan) that are necessary to discriminate between binding to its cognate DNA binding site (sequence in blue) and binding to the canonical E-box sequence (purple) typically recognised by other SREBPs. Chimeric proteins were generated by exchanging parts of the DNA-binding domain of *CaHms1p* (cyan) with that of *CaCph2p* (orange). The latter protein, as most SREBPs, cannot discriminate *in vitro* between both DNA binding sites. Chimeric proteins harbouring the *CaHms1p* first helix, second helix or loop alone displayed no discrimination between the two sequences. The chimeric protein containing the *CaHms1p* first helix and loop bound preferentially to the Hms1p's cognate DNA binding sequence. The protein concentrations evaluated in all gel shift assays included in this figure were 0, 1.56, 6.25, 25 and 100 nM.

#### **2.1.4.5. Ancestral protein reconstruction reveals pattern of divergence of SREBP's DNA binding preferences**

A major difference between branches 2 and 3 of the fungal SREBPs (Figure 3) (these branches are represented by *CaHms1p* and *CaTye7p*, respectively) is their intrinsic ability to discriminate between the canonical, palindromic E-box (core motif 5'-CANNTG-3') and the non-palindromic DNA sequence 5'-ATCANNCCA-3'. While Hms1p and related proteins exhibited a strong preference for the non-palindromic DNA (Figure 6), Tye7p showed a slight but consistent preference for the canonical E-box (Figure 9). The presence of a tyrosine residue in the DNA binding domain of the SREBPs (instead of a conserved arginine in other bHLH proteins) allows binding to either DNA sequence [59]. But the question of how the preference to bind one or the other sequence arose remains unanswered. There are at least two plausible scenarios that could be envisioned. First, an ancestor that had a clear preference for one of the two sequences could have given rise to a lineage that reduced (and eventually flipped) its DNA binding preference. Alternatively, an ancestor that bound both sequences equally well could have given rise to one branch that tilted its DNA binding preference in one direction and another branch whose DNA binding preference tilted in the opposite direction. To empirically test these models, I used ancestral protein reconstruction [39,90,96] which allows reconstruction of the sequence and further synthesis of putative ancestor proteins at different nodes of a given phylogenetic tree.



**Figure 9** *CaTye7* shows a preference for the canonical E-box motif over the non-palindromic DNA.

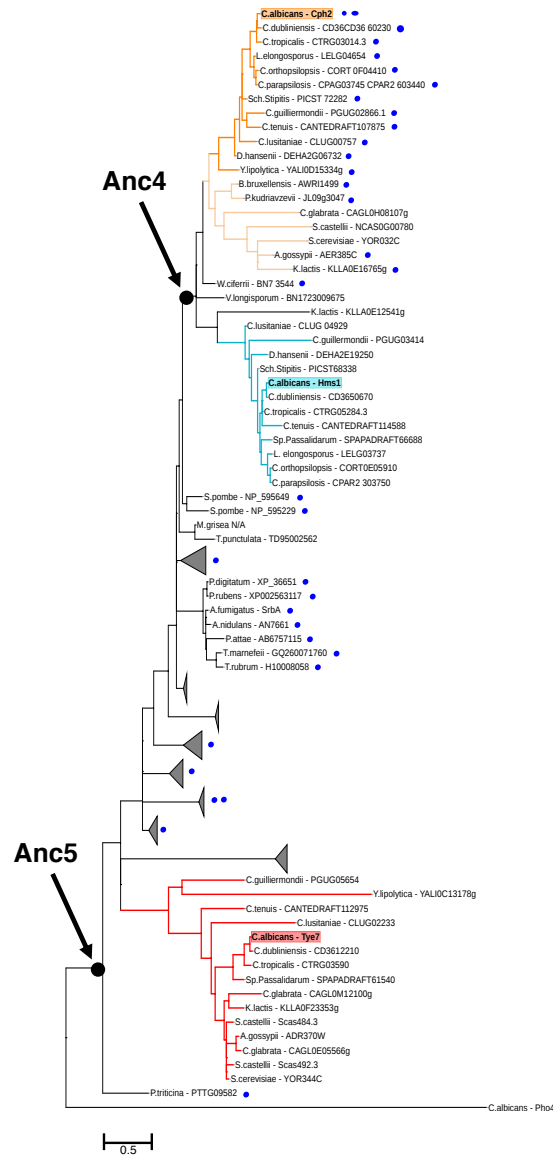
Gel shift competition assays with the purified DNA binding domain of *CaTye7p*. The  $\text{P}^{32}$ -radiolabeled DNA (and the unlabelled competitors) contained either the non-palindromic binding site (top panel) or a palindromic E-box sequence (bottom panel). The quantification of competition assays and best-fit curves are included on the right side of each gel. *Tye7p* shows a slight, but consistent, preference for the canonical E-box motif.

I inferred the amino acid sequences of the putative ancestors at two selected nodes of the fungal SREBP phylogeny (Figure 10 and 11A), then expressed and purified the predicted ancestral proteins. The first node corresponding to the named Anc4 is the common ancestor of the branches that gave rise to the *C. albicans* Hms1 and Cph2 lineages. The second reconstructed ancestor is named Anc5 and corresponds to the “oldest” SREBP ancestor of the phylogeny.

The reconstructed ancestors (Anc4 and Anc5) are unique protein sequences. Anc4 shares 67.9, 57.1 and 51.8% identity with Cph2, Hms1 and Tye7, respectively whereas the percentage of identity of Anc5 is 66.1, 62.5 and 58.9%. The alignment in Figure 11A shows the sequences of both ancestor 4 and ancestor 5 as well as the sequences of the DNA binding domains of *C. albicans*' Hms1 and Tye7 for comparison.

I then tested the ability of the ancestor proteins to bind to DNA in EMSAs. While a higher amount of ancestor protein was required to bind to the DNA sequences and produce a “shift,” the protein-DNA interactions were still sequence-specific (Figure 11B). I then carried out *in vitro* competition assays to determine the preference of the purified ancestral proteins for either the palindromic or the non-palindromic sequence. As shown in Figure 11C, the “oldest” ancestor (Anc5) displayed only a slight preference for the non-palindromic sequence whereas the more “recent” ancestor (Anc4) exhibited a stronger preference for the same sequence (non-palindromic site over the canonical E-box). Compared to the “oldest” ancestor, Anc4's DNA binding preference is a step closer to that displayed by *CaHms1p*. However, the preference exhibited by *CaHms1p* towards the non-palindromic sequence is still about an order of magnitude higher than in Anc4 (Figure 12).

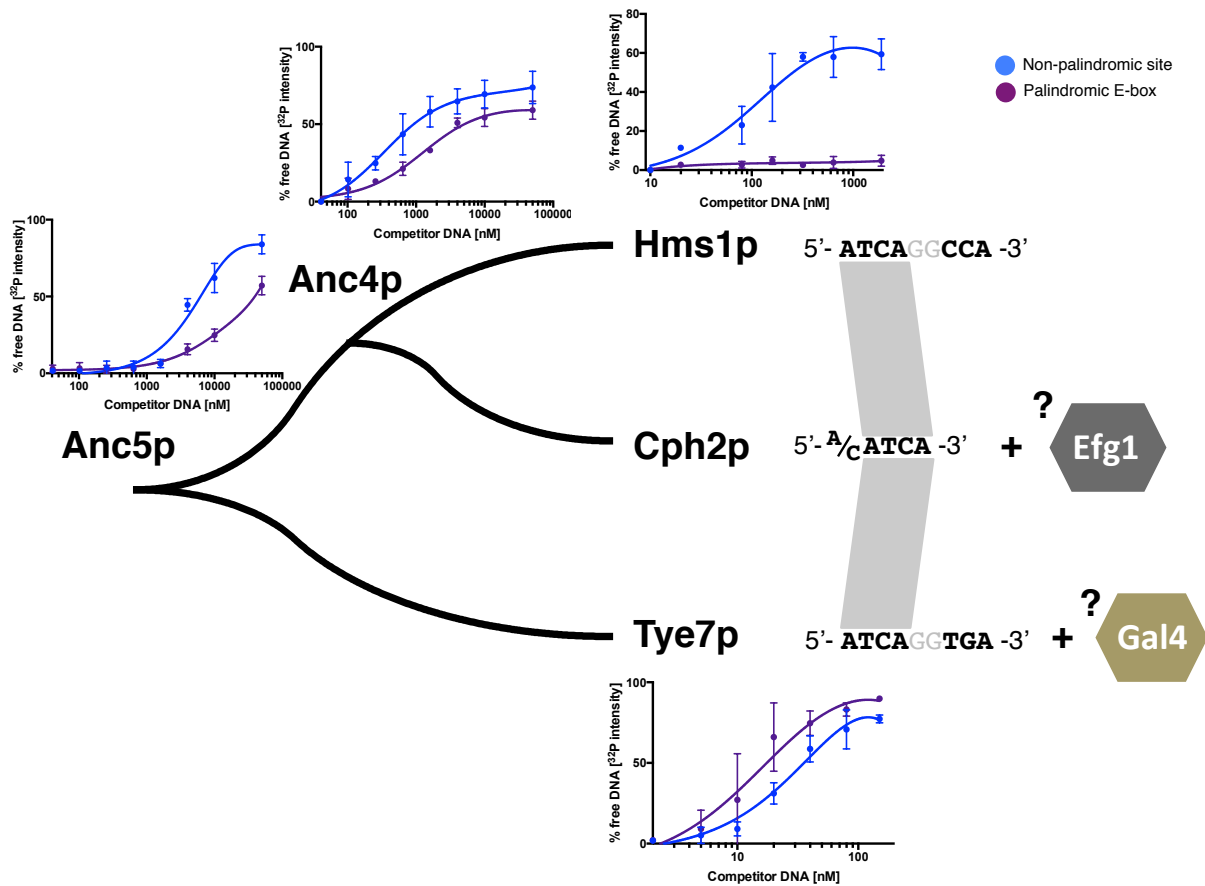
These findings support the notion that two extant branches of fungal SREBPs, which are represented by Hms1p and Tye7p, followed divergent paths after separating from their last common ancestor: The Hms1 lineage enhanced the ancestor's initial preference for the non-palindromic sequence whereas the Tye7 lineage reduced, and eventually flipped, the DNA-binding preference of the ancestor (Figure 12).



**Figure 10 Ancestral protein reconstruction at two selected nodes of the fungal SREBP phylogeny.**

Shown is the phylogenetic reconstruction of fungal SREBPs. Two likely ancestors at the indicated nodes were reconstructed using Phylobot software [90]. Anc4 is the likely ancestor of the branches that gave rise to *CaCph2* (orange) and *CaHms1* (cyan). Anc5 is the likely ancestor of the branches that gave rise to all three *C. albicans* SREBPs.





**Figure 12 Divergence of DNA binding preferences in fungal SREBPs.**

Diagram depicting the DNA binding preferences in extant and ancestral fungal SREBPs. Shown are the quantifications of *in vitro* DNA binding competition assays carried out with the two ancestor proteins, *CaHms1p* and *CaTye7p*. The more separation between the blue and purple lines indicates clearer preference for one of the two sequences. Notice that Hms1p shows strong preference for the non-palindromic sequence (blue) whereas Tye7 displays slight preference for the palindromic E-box (purple). To the right are shown the DNA sequences that each protein binds *in vivo*. Co-factors such as Efg1p and Gal4p may contribute, at least in part, to the specific binding *in vivo* of Cph2p and Tye7p.



### 2.1.5. Discussion

In this chapter I have studied the DNA binding preferences of the SREBP family of transcription regulators in fungi and the mechanisms by which they diversified over evolutionary timescales. I generated a comprehensive phylogenetic reconstruction, based on the DNA binding domain of the proteins, which showed that the genomes of the majority of fungi encode one or two SREBPs. Strikingly, in the ascomycete yeasts, some *Candida* species encoded three SREBPs. In the case of *Candida albicans*, the three SREBPs were found spread in distinct branches of the phylogeny (Figure 3). The basic-helix-loop-helix architecture of their DNA binding domain allows SREBPs to bind the canonical E-box motif that other bHLH proteins bind to. All SREBPs, however, contain a distinctive tyrosine residue within the first helix of the DNA binding domain (where all other bHLH have an arginine) that confers the ability to bind to a non-palindromic motif in addition to the palindromic E-box *in vitro* [59]. The most well studied SREBP, the human SREBP1, is able to bind both, the palindromic and non-palindromic DNA sequences *in vitro* [58]. Nonetheless, ChIP experiments revealed that *in vivo* the protein binds the palindromic E-box only [56]. Therefore, the significance of such “dual” binding, if any, is not understood.

In this chapter I have scanned the DNA binding affinities of several fungal SREBPs representative of the different branches of the phylogeny. In *C. albicans* the three SREBPs displayed distinct, and for the most part, non-overlapping binding profiles in comprehensive *in vitro* binding experiments. *CaTye7p*, like the canonical human SREBP, bound preferentially to instances of the palindromic E-box motif. *CaHms1p* displayed a strong preference for versions of the non-palindromic DNA and *CaCph2p* showed an intermediate binding profile binding to DNA sequences resembling both motifs (Figure 4). The preference for the non-palindromic site was not exclusive to *CaHms1p* but could also be extended to the orthologous proteins in other members of the ascomycetes like *C. parapsilosis* and also SREBPs outside this clade, e.g. *A. fumigatus* SrbAp. In the case of *CaHms1p* the affinity could be traced back to amino acids of the first helix and loop of the DNA binding domain of the protein (Figure 8).

The data derived from *in vitro* tests, where only the DNA binding domain of the proteins was purified, largely agreed with the *in vivo* results from the ChIP experiments. This observation suggests that, to a large extent, it is an intrinsic ability of the proteins to

distinguish and have different preferences for the palindromic or non-palindromic sequences (Figure 5H). However, despite the differences of affinity seen among the three *C. albicans* SREBPs, extra co-factors may play a role *in vivo*. This notion is supported by the fact that in the case of *CaCph2p* there was a co-occurrence of the Cph2p motif with the motif recognised by another regulator, Efg1p (Figure 5F), and they seem to co-regulate some of their target genes *in vivo* (Figure 5G).

Thus, the *C. albicans* SREBPs have distinct DNA binding affinities. Ancestral protein reconstruction was used to shed light on the question of how the different affinities originated. The experiments revealed that such diversification was resolved during the evolution of the family with one branch tilting the preference towards one of the DNA motifs largely through amino acid changes in the same protein whereas another branch tilted the preference towards the second motif.

Taken together, the results suggest that different mechanisms likely contributed to the diversification of the DNA binding preferences within this family. On the one hand, some regulators enhanced the previous affinity for the non-palindromic site, which is the case of *CaHms1p*, *CpHms1p* and *A/SrbAp*. In some cases, like for *CaHms1p*, to the point where the regulator almost lost the ability to bind the palindromic E-box. On the other hand, some proteins decreased and eventually flipped the initial ancestral preference for the non-palindromic DNA towards the palindromic site, which is the case for *CaTye7p*. Finally, *CaCph2p* seems not to have a clear preference for either of the two evaluated sequences because it recognises only the first half-site of the motif (which is shared by the palindromic and non-palindromic sites). The DNA binding specificity of this protein *in vivo* may be established through interactions with co-factors (Figure 12).

## **2.2. Diversification of the biological functions controlled by the SREBP family of transcription regulators**

### **2.2.1. Summary**

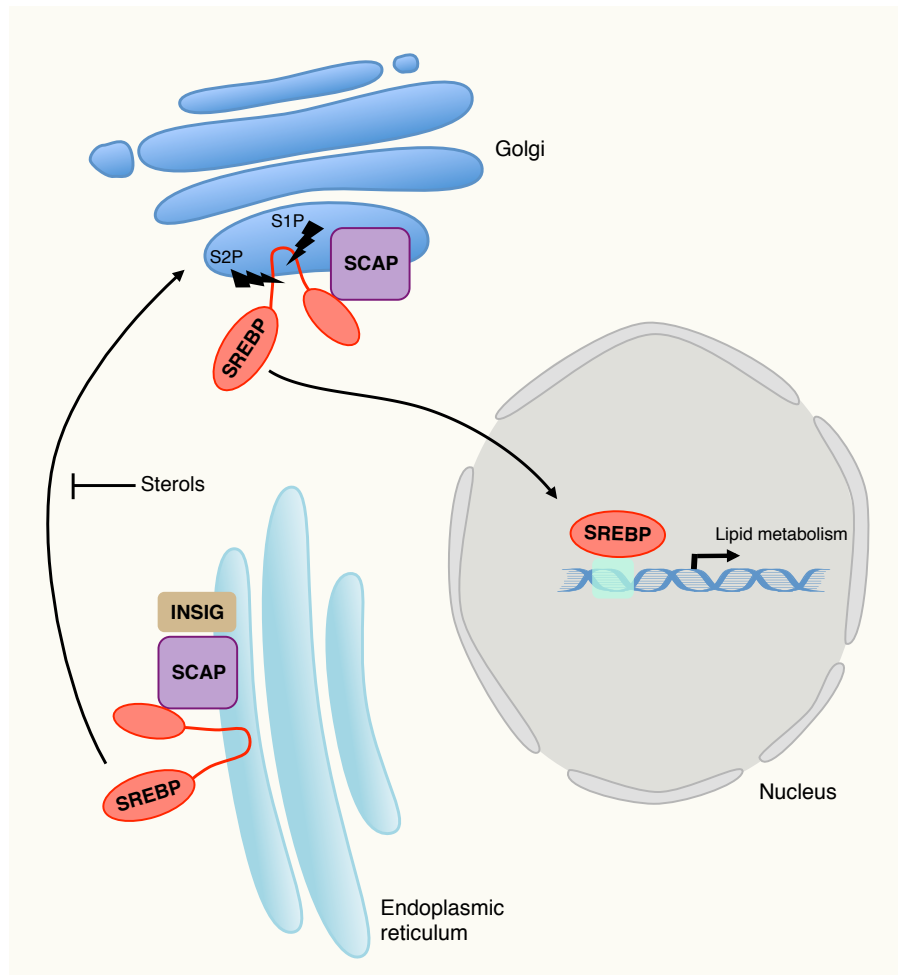
The Sterol Regulatory Element Binding Proteins (SREBPs) are a family of transcription regulators widely conserved among eukaryotes ranging from fungi to humans. As their name implies, these proteins regulate the expression of genes related to lipid biosynthesis in most organisms. In this thesis, I have studied SREBPs in the fungal lineage that encompasses the human associated yeast *Candida albicans* which harbours three SREBP proteins. In the first chapter of this thesis, I described that *C. albicans*' SREBPs diversified in their DNA binding affinities. All three proteins contribute to the ability of *C. albicans* to colonise the mammalian gut, but, unlike canonical SREBPs, they do so by regulating cellular processes unrelated to lipid biosynthesis. In this second chapter, I report that two SREBPs in the human commensal yeast *C. albicans* drive a transcriptional cascade that inhibits a morphological switch of the fungus under anaerobic conditions. Preventing this morphological transition enhances *C. albicans*' colonisation of the mammalian intestine – the fungus' natural niche. Therefore, the regulatory changes in the transcriptional network governed by the SREBPs were likely key in enabling this fungal lineage to associate with animals.

## 2.2.2. Introduction

### 2.2.2.1. SREBPs regulate the expression of sterol biosynthesis genes in most eukaryotes

An essential requirement for the maintenance of cell membrane structure and fluidity is the tight regulation of sterol lipids. For this reason, it is important to understand how cells maintain and regulate their sterol levels. SREBPs are basic-helix-loop-helix (bHLH) transcription regulators that play a crucial role in this process. In the early 90s the first described SREBP was isolated from rat liver and was characterised as a nuclear protein with the ability to bind a sterol regulatory element (DNA sequence) upstream of the promoter of the low density lipoprotein (LDL) receptor [49]. In mammals, when cellular levels of sterol are low, the LDL receptor is transcribed leading to an increase of cellular LDL uptake. SREBP1 (name given to the rat SREBP) is the protein in charge of regulating the levels of expression of the LDL receptor gene by binding to a specific *cis*-regulatory sequence (which researchers at the time termed SRE1) located in its promoter [49,97].

SREBPs are extensively distributed among eukaryotes ranging from humans to unicellular fungi. In mammals, SREBPs are synthesised as inactive precursors and are anchored to the membrane of the endoplasmic reticulum (ER) via two transmembrane domains which result in both the N- and C-termini facing the cytosol [98]. There, SREBPs are bound by a sterol-sensing protein named Scap (SREBP cleavage activating protein) that mediates SREBP's activation. When cellular sterol levels are high, cholesterol is bound by Scap triggering a conformational change that allows Scap to interact with INSIG, an ER-resident protein. This complex retains SREBPs inactive in the ER membrane. However, when cholesterol is depleted, the Scap-INSIG complex disassociates and vesicles mediate the translocation of the Scap-SREBP complex to the Golgi apparatus where site-1 serine proteases and site-2 metalloproteases cleave and release SREBP from the membrane. The N-terminus of SREBP is released into the cytosol and then shuttles to the nucleus to regulate transcription (Figure 13) [99–101].



**Figure 13 Mammalian SREBP pathway.**

In response to low levels of cellular sterols, SREBP dissociate from the Scap-INSIG complex that anchored to the ER membrane and is translocated to the membrane of the Golgi apparatus, cleaved and shuttled to the nucleus. Once in the nucleus it binds specifically to the promoter regions of genes involved in lipid metabolism to promote their transcription.

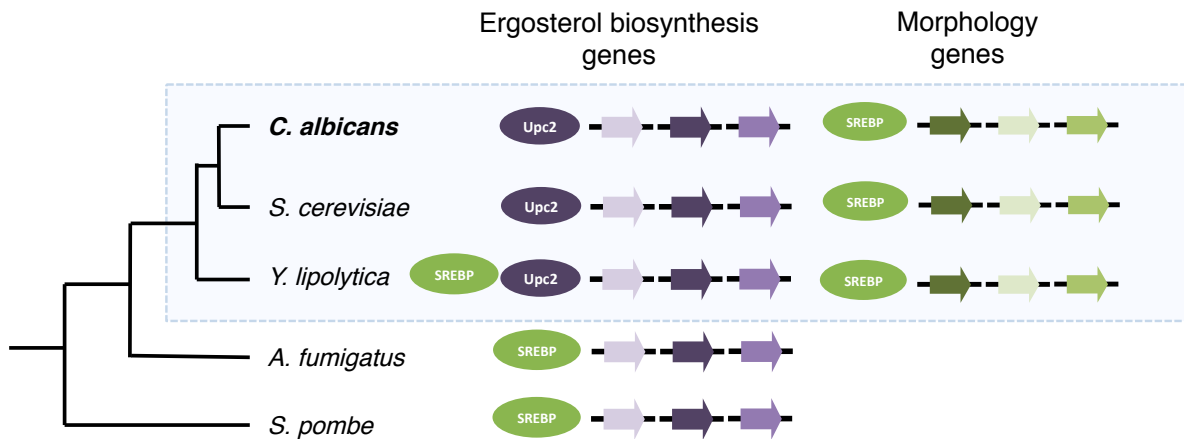
In addition to higher eukaryotes, SREBP family members are also broadly distributed in fungi where they have been found to underlie similar sterol related functions. For instance, in the fission yeast *Schizosaccharomyces pombe* the SREBP protein named Sre1 is also synthesised as an inactive precursor and it is cleaved in response to low levels of ergosterol, the fungal equivalent of cholesterol [102]. Furthermore, *S. pombe* harbours homologs of the mammalian Scap and INSIG proteins. In this organism, SREBPs regulate genes required for sterol biosynthesis and its absence leads to the inability of the cells to maintain adequate sterol levels. SREBPs carry out similar functions in other fungi such as the human pathogen *Cryptococcus neoformans* [102]. Thus, the SREBP pathway is crucial in regulating cellular sterol homeostasis and it is conserved in several fungi [91,99,103].

### 2.2.2.2. SREBPs have adopted different functions in yeasts

As explained above, the main function associated with most SREBPs is the regulation of sterol biosynthesis genes. This statement holds true for several organisms such as humans [56,97], mice [49], the fruit fly *D. melanogaster* [55], the nematode *C. elegans* [104] and unicellular fungal species like *S. pombe* [103] or *C. neoformans* [102]. In earlier chapters of this thesis (section 2.1.4.1) I have described that in fungi most species encode one or two SREBPs in their genomes. However, in some lineages like the *Candida* clade of the ascomycete yeasts (Saccharomycotina), the family has expanded raising questions about their role in these organisms. In the ascomycete yeasts (where the family has expanded), the regulation of sterol biosynthesis genes is not controlled by SREBPs but by a transcription regulator named Upc2p, which belongs to the zinc finger family. In fact, in this lineage SREBPs have been reported to control a process completely unrelated to lipid metabolism: they have been described to regulate the expression of genes related to the yeast-to-hyphae morphological switch [105].

Interestingly, in the yeast *Yarrowia lipolytica*, which lies at the base of the subphylum Saccharomycotina and is considered an ancestor of the group, the ergosterol biosynthesis genes are regulated partly by both, SREBPs and Upc2p. This observation has

been interpreted as a transition of the SREBP proteins from regulators of sterol synthesis to morphology in the Saccharomycotina lineage (Figure 14).



**Figure 14 SREBPs regulate a morphological switch in the Saccharomycotina clade.**

Cladogram depicting the phylogenetic relationships between several fungal species. Organisms belonging to the Saccharomycotina are represented inside the blue rectangle. Upc2p replaced SREBPs in the regulation of ergosterol genes in the Saccharomycotina. In *Y. lipolytica*, however, SREBPs retained a role in ergosterol metabolism. SREBPs govern the yeast-to-filament transition in Saccharomycotina.



### 2.2.2.3. *Candida albicans* morphologies: yeast and filamentous forms

The ascomycete yeast *Candida albicans* is a polymorphic fungus [106]. Through a morphogenesis process it can switch between different forms depending on environmental cues. Some of the best-studied morphologies (and the ones that are relevant to this dissertation) are the yeast and the filamentous forms (hyphae and pseudohyphae). Cells in the yeast form have a round or oval morphology similar to that of *S. cerevisiae* [106,107]. This cell type is considered to be innocuous for the host as yeast cells have been found residing in the mammalian gut when *C. albicans* grows as a commensal [108]. When found in its filamentous form, *C. albicans* cells form long tube-like structures named hyphae. This morphology has been associated with pathogenesis because hyphae are invasive [109] and express several virulence factors (like adhesins) and tissue-degrading enzymes (for instance secreted aspartyl proteases) not produced by yeast cells. Moreover, hyphae have been found to induce endocytosis and posterior damage in human epithelial cells triggering host pro-inflammatory responses [110,111]. The ability to switch between yeast and hyphal morphologies is required for virulence and biofilm formation [112–114].

Several environmental cues trigger the morphology switch in *C. albicans* by activating different signalling pathways. For instance, temperature of 37°C (the temperature of the mammalian host), CO<sub>2</sub> levels, alkaline pH or the addition of serum to the medium can affect morphology by activating the cyclic AMP protein kinase A (cAMP-PKA) signalling pathway. The PKA complex can activate key TRs that promote the expression of filamentation genes [115,116]. Cell damage as well as oxidative and osmotic stress influence morphology also through kinase signalling pathways that in turn promote the expression of hyphal-specific genes [117,118]. Finally, other signals such as oxygen or nutrients can repress hyphal formation [119]. Thus, morphological transitions between the different cell types are controlled at a transcriptional level. One of the key regulators of the morphological transitions is Efg1p [120]. This regulator promotes the switch from yeast to hyphae when cells are exposed to elevated levels of CO<sub>2</sub>, serum or starvation [120] but it can also repress the filamentous form when the cells are grown embedded in agar [121].

#### 2.2.2.4. Anaerobiosis in *C. albicans*

*C. albicans* lives as a commensal in association with humans and other animals. Some of the niches the fungus is able to colonise in the human body are superficial (*i.e.* the skin or the oral cavity) where oxygen is available at high levels. On the contrary, in other niches like the gastrointestinal tract the fungus faces an environment largely devoid of oxygen. The levels of oxygen can be generally defined in three terms: (*i*) anaerobic or anoxic, to refer to an environment with a near complete absence of oxygen (*ii*) hypoxic, to describe an environment with a reduction in available oxygen compared to atmospheric levels and (*iii*) normoxic, to refer to atmospheric levels of oxygen, generally 21%.

The response to hypoxia in *C. albicans* is complex and not fully understood. In other fungi like *S. pombe*, the SREBP named Sre1p functions as an oxygen sensor by monitoring oxygen-dependent sterol synthesis making levels of sterol an indirect measurement of oxygen availability [91,103]. Decreasing levels of oxygen are detected by the reduction of sterol synthesis which requires 12 molecules of oxygen to convert one squalene in ergosterol [91]. In response to low levels of sterol, a homolog of the human Scap protease cleaves Sre1p (an SREBP protein) which is the main regulator of the hypoxic response in *S. pombe* [91,103]. In *C. albicans* the response to hypoxia is also associated with ergosterol synthesis but unlike *S. pombe*, in *C. albicans* (which has no known homologs of Scap) the response to sterols is mediated by Upc2p [122].

Other regulators that also play an important role during the response to hypoxia, mainly affecting the morphology of the fungus, are Efg1p [123] and Ace2p [124]. Thus, alterations in the levels of available oxygen affect the morphology of *C. albicans*. Studies have reported that hypoxia induces the switch of the fungus from yeast to hyphae; this phenotypic transition has been associated with invasion and virulence [124–126]. However, in the GI tract – where some parts like the colon or cecum have less than 1 mm Hg of oxygen available, making it a largely anaerobic environment rather than hypoxic [127,128] – the predominant morphology is the yeast form, which seems to allow *C. albicans* to reside and grow in this niche undetected by the immune system [108].

### 2.2.3. Materials and methods

#### 2.2.3.1. Transcriptome analyses

The *C. albicans* reference strain and the *hms1* and *cph2* deletion mutants were grown in Todd-Hewitt broth in an anaerobic chamber at 37°C for 24 hours. The gas composition in the anaerobic chamber (Don Whitley Scientific) was 10% hydrogen, 10% carbon dioxide and 80% nitrogen. The medium was placed in the anaerobic chamber at least two days before inoculation to remove any oxygen traces. Two independent replicates were used for the analysis. Total RNA extraction and cDNA synthesis was performed as previously described in section 2.1.3.5. Between 63 and 91 million reads per sample were obtained. Prior to mapping, Illumina adaptor sequences were removed from the reads using Trimmomatic v0.36-5 [129] as follows:

```
$ trimmomatic SE -phred33 <reads.fastq> <reads_trimmed.fastq>
  ILLUMINACLIP: ./TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3
  SLIDINGWINDOW:4:15
```

Trimmed reads were then aligned to the *C. albicans* genome using STAR v2.5.2b [130]. First, I generated a reference genome using the *C. albicans* transcriptome (assembly 21) available at the Candida Genome Database ([www.candidagenome.org](http://www.candidagenome.org)):

```
$ star --runMode genomeGenerate --runThreadN 3 --genomeDir
  ./reference_genome --genomeFastaFiles
  ./reference/<transcriptome_sequence.fasta> --sjdbGTFfile
  ./<transcriptome_annotation.gtf> --sjdbOverhang 49
```

Trimmed reads were then aligned to the reference genome:

```
$ star --runThreadN 3 --genomeDir ../reference_genome --readFilesIn
  <reads_trimmed.fastq> --outFileNamePrefix
  ../output/bam/<reads_trimmed.bam> --outSAMtype BAM
  SortedByCoordinate --outSAMstrandField intronMotif
```

Reads assigned to genome features were counted from the BAM files using HTSeq [131]:

```
$ htseq-count -s no -f bam -r pos <reads_trimmed.bam>
  ../../reference/ <transcriptome_annotation.gtf> >
  ../../counts/<reads.counts>
```

More than 97% of reads of each sample were uniquely aligned to the *C. albicans* genome. Read counts were loaded into R (v3.3.2) and analysed with the DESeq2 [132] package (v1.14.1). With our depth of sequencing, significant numbers of reads were detected for ~6,100 annotated ORFs. Cytoscape [78] (v3.4) was used to visualise the data and generate the network graphs. The significance of the overlap between the differentially expressed genes from the RNA-seq datasets was estimated using the hypergeometric test. The Gene Ontology Term Finder of the Candida Genome Database ([www.candidagenome.org](http://www.candidagenome.org)) was used to identify enriched processes in our RNA-seq dataset.

The RNA-seq datasets from reference strain and *cph2* and *hms1* mutants grown in anaerobic conditions can be found in the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE118414. Datasets for the remaining samples mentioned in chapter 2.2.4.3 (RNA-seq of cells under aerobic vs. anaerobic conditions) were generated by other members of the Pérez Lab.

### 2.2.3.2. Cell morphology determination

Overnight *C. albicans* cultures (in Todd-Hewitt broth at 30°C) were diluted to OD<sub>600</sub> ~0.1 in fresh Todd-Hewitt broth and incubated at 37°C under anaerobic conditions for 24 hours. The medium used to dilute the overnight culture had been pre-incubated in an anaerobic chamber for 48 hours to achieve complete anaerobiosis. After the 24-hour period of growth, cells were washed with sterile PBS and fixed in glass slides for morphology evaluation under the microscope.

### 2.2.3.3. Protein immunostaining

Two *C. albicans* strains in which the *CPH2* gene was Myc-tagged at different locations were employed for immunostaining experiments (see genotypes in Appendix Table 7). In the first strain (JCP\_880) a Myc-tag was incorporated at the C-terminal of *CPH2*, immediately after the DNA binding domain and thus replacing the transmembrane domain and rest of the protein. As previously described in section 2.1.3.4, this modification rendered the protein constitutively active. The expression of Cph2-Myc was under the control of the endogenous *CPH2* promoter. The second strain was provided by Dr. Haoping Liu [95]). Briefly, this strain harbours an N-terminally tagged version of one *CPH2* allele, which is under the control of its own promoter. This construct is inserted into the *ADE2* locus. By tagging the N-terminal of the protein both the cleaved and full-length fractions of *CPH2* could be visualised. Additionally, a reference strain with no Myc-tag in its genome was used as a negative control and a strain harbouring a Myc-tagged version of *SHE3* (Myc-*SHE3*) served as a positive control.

Cells were grown in YPD broth at 30°C to an OD<sub>600</sub> of 0.5 and then washed with 1x PBS before resuspending them in 4% paraformaldehyde during 1 hour at room temperature in order to fix them. Cells were centrifuged and washed with ice cold buffer B to remove any traces of paraformaldehyde. In order to remove the cell walls and generate spheroplasts, cells were resuspended in spheroplasting buffer (prepared fresh) for 30 minutes at 30°C with gently inversion every 5 minutes. Spheroplasts were then centrifuged (at a low speed of 3500 rpm to minimise damage) and washed with ice cold buffer B before placing them onto 8 well chambers (Ibidi). To achieve attachment to the well chamber, the spheroplasts were incubated for at least 1 hour at 4°C then washed with cold buffer B and dehydrated with 70% ethanol overnight at -20°C. Once cells were attached and dehydrated the ethanol was removed and the cells were rehydrated by adding 2× SSC buffer and then blocked with 5 % milk powder in 1× TBS for 1 hour at room temperature. Next, the cells were stained with 1:300 mouse anti-c-Myc antibody (Sigma) in blocking buffer for two hours, washed three times with 1× PBS and incubated for 1 hour with 1:200 Alexa-fluor-488 anti-mouse secondary antibody (VWR) and DAPI in blocking buffer at room temperature and in the darkness. Finally, cells were washed three times

with 1× PBS before being imaged with a confocal microscope (Leica). Images were processed using ImageJ (v. 2.0)

#### Spheroplasting buffer

1,8 ml     1× Buffer B  
5 µl       β-Mercaptoethanol  
390 µl     H<sub>2</sub>O  
40 µl       Lyticase (25.000 U)

#### 1× Buffer B

1,2 M       Sorbitol

Adjusted to 1L with 100mM potassium phosphate buffer pH 7.5

#### 100 mM Potassium phosphate buffer (pH 7.5)

83,4 ml    1M K<sub>2</sub>HPO<sub>4</sub>  
16.6 ml    1M KH<sub>2</sub>PO<sub>4</sub>

#### 20× SSC (pH 7.0)

3 M         NaCl  
300 mM    Na<sub>3</sub>C<sub>6</sub>H<sub>5</sub>O<sub>7</sub>

#### **2.2.3.4. Western blot analysis**

A *C. albicans* strain encoding an N-terminal Myc-tagged Cph2 protein (JCP\_960, see genotype in Appendix Table 7, for further information about this strain see section 2.2.3.3) was grown in either minimal (SD) or rich medium (Todd-Hewitt or YPD broth). Two to three colonies were taken from YPD agar plates and resuspended in 3 ml of either SD, YPD or TH broth. This medium containing the resuspended colonies was then used to inoculate a 50 ml culture through a 1:200 dilution in rich media and 1:100 in minimal medium. The cultures were grown at 37°C on a shaker until OD<sub>600</sub> reached values of 0.5 – 0.6. Sample preparation was carried out as follows: 2-3 ml of cell culture were washed

with 1 ml of water, centrifuged and resuspended in 100  $\mu$ l of water. Next, 100  $\mu$ l of 2M NaOH were added and the mix was incubated for 10 minutes at room temperature. Cells were then centrifuged down, resuspended in 50  $\mu$ l of SDS sample buffer (0.06M Tris-HCl, 5% glycerol, 2% SDS, 4%  $\beta$ -mercaptoethanol, 0.0025% bromophenol blue) and boiled at 100°C for 5 minutes. Samples were resolved by SDS polyacrylamide gel electrophoresis (PAGE) for 1.5 hours at 140 V. The protein samples were then blotted into a PVDF membrane (previously equilibrated with methanol) by wet transfer for 1.5 hours at 30V in transfer buffer. The blotted membrane was blocked for an hour using 1 $\times$ TBS containing 5% of milk powder and then hybridised with 1:4800 mouse anti-c-Myc primary antibody (Sigma), diluted in 1 $\times$ TBS containing 5% (w/v) of milk powder, overnight on a shaker at 4°C. The membrane was washed with 1 $\times$  TBST and hybridised with 1:12000 secondary antibody conjugated to horseradish peroxidase (HRP) for 2 hours at room temperature. Chemoluminescence substrate solutions (Thermo Scientific) were used before imaging.

#### Transfer buffer

14.5 g	Glycine
3 g	Tris base
100 ml	methanol
400 ml	H <sub>2</sub> O

#### TBS buffer (10 $\times$ )

24.11 g	Tris base
72.6 g	NaCl

pH = 7.4 adjusted with HCl 1l H<sub>2</sub>O

#### TBST buffer

1 $\times$	TBS
0.1% (v/v)	Tween 20

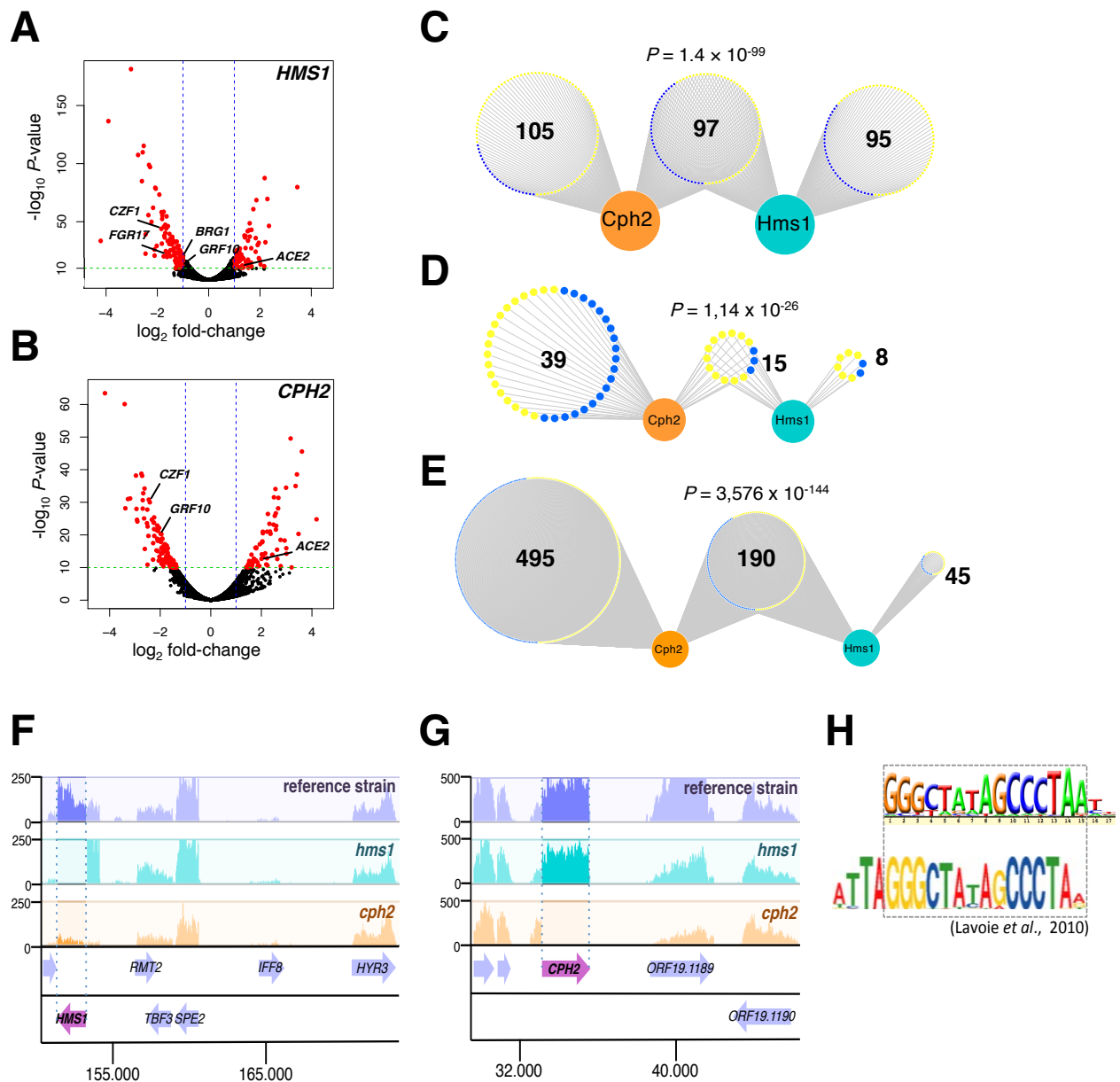
## 2.2.4. Results

### 2.2.4.1. The *C. albicans* SREBPs Hms1 and Cph2 form a regulatory cascade

The *C. albicans* genome harbours three SREBPs whereas most other organisms have only one or two. This raises the question of what functions they perform in this particular fungus. *CaTye7p* is the best studied of the three and has been implicated in glycolysis and sugar metabolism [94]. However, the function(s) of *CaHms1p* and *CaCph2p* remain(s) less clear. The SREBPs in other species, and the *CaTye7*, have in common that they regulate cellular processes sensitive to oxygen [94,102,133]. I hypothesised, then, that the other two *C. albicans* SREBPs might play a role when the fungus proliferates in a niche largely devoid of oxygen. To identify the repertoire of target genes regulated by *HMS1* and *CPH2*, I performed transcriptome analyses (RNA sequencing) of the wild-type reference strain and isogenic *cph2* or *hms1* deletion mutants grown in an anaerobic chamber at 37°C (the temperature of the mammalian host).

Overall, the RNA-Seq experiment revealed 202 and 192 protein-coding transcripts whose expression was dependent on *CPH2* and *HMS1*, respectively ( $-\log_{10} P > 10$  and expression changes  $>2$ -fold; Figure 15A and B; 685 and 235 targets at  $P < 0.001$  and expression changes  $>2$ -fold). There was a significant overlap in targets of regulation between *CPH2* and *HMS1* ( $P = 1.36 \times 10^{-99}$ ) (Figure 15C) implying that these two SREBPs form a regulatory cascade. The overlap was significantly higher than expected by chance even if more (Figure 15D) or less (Figure 15E) stringent thresholds were applied. Moreover, all genes whose expression was dependent on both TRs changed expression in the same direction; that is, if they were up regulated by *HMS1* they were also up-regulated by *CPH2*. Consistent with the idea of these two proteins forming a regulatory cascade, I found that *HMS1* expression was dependent on *CPH2* (but not vice versa) (Figure 15F). *Cph2p* binds *in vivo* to the intergenic region upstream of *HMS1* (Figure 15G and [95]) further supporting a direct regulatory link between these two factors.





**Figure 15** The *C. albicans* SREBPs *CPH2* and *HMS1* form a regulatory cascade.

(A, B) Identification of transcripts regulated by HMS1 and CPH2. Total RNA was extracted from wild type, *hms1* and *cph2* deletion mutant strains grown at 37°C under anaerobic conditions. Shown are volcano plots where each dot represents one transcript. Red dots represent transcripts that pass the threshold of expression changes > 2-fold (x axis) and  $-\log_{10} P$ -value < 10 (y axis). Several regulators of filamentation are marked (*ACE2*, *BRG1*, *CZF1*, *FGR17* and *GRF10*). (C, D, E) Overlap of targets of regulation between Cph2p and Hms1p. Up-regulated genes are shown in yellow and down-regulated genes in blue. The significance of the overlap was calculated using the hypergeometric distribution. The

thresholds used to construct the networks were the following:  $\log_2$  fold change  $> |1|$   $-\log_{10}$   $P$ -value  $> 10$  in C,  $\log_2$  fold change  $> |2|$   $-\log_{10}$   $P$ -value  $> 20$  in D,  $\log_2$  fold change  $> |1|$   $P$ -value  $< 0,001$  in E. **(F, G)** The expression of *HMS1* is dependent on *CPH2* but no vice-versa. Representative segments of RNA-seq tracks for the wild-type reference strain (purple), *hms1* (blue track) and *cph2* (orange track) mutants. The levels of *HMS1* transcript are dramatically reduced in the *cph2* deletion strain. **(H)** Cph2p induces ribosome biogenesis through Tbf1p. Motif search analyses on the promoter regions of genes regulated by *CPH2* only revealed an overrepresented motif (top) that is almost identical to the sequence recognised by Tbf1p (bottom).

Gene Ontology (GO) analysis of the differentially expressed genes revealed filamentous growth ( $P = 1.7 \times 10^{-4}$ ), pathogenesis ( $P = 1.44 \times 10^{-7}$ ) and biofilm formation ( $P = 1.82 \times 10^{-13}$ ) as cellular processes or functions enriched in the dataset (it should be noted that over 50% of the genes in the dataset are annotated as having unknown functions). Indeed, the transcript levels of several well-established regulators of yeast-to-filament transition, *e.g.* *CZF1*, *GRF10* and *ACE2* [124,134,135], appeared to be under control of both *HMS1* and *CPH2*. The direction of the change in expression in *CZF1*, *GRF10* and *ACE2* suggested that, under the environmental condition evaluated, both *HMS1* and *CPH2* work by preventing filamentation. In other growth conditions, *HMS1* and *CPH2* have been associated with the opposite phenotype (i.e. promoting the yeast-to-filament transition) [136,137].

Additionally, GO analysis also showed an overrepresentation of genes related to ribosome biosynthesis and nitrogen compound biosynthesis ( $P = 1.12 \times 10^{-14}$ ). This was striking because neither Cph2p nor Hms1p had been reported to play a role in ribosome biosynthesis before. As shown in the overlaps of Figure 15C-E, *CPH2* has more targets of regulation on its own compared to *HMS1*. Further analysis revealed that the ribosome biosynthesis genes were all up regulated and controlled only by *CPH2*. These data suggest that either Cph2p itself bound upstream of ribosome related genes or alternatively, that Cph2p affected the expression of another TR which would bind to the ribosome genes causing downstream effects. To distinguish between these two possibilities, I scanned the sequences of promoter regions of the genes regulated by *CPH2* only (threshold:  $\log_2$  fold change  $> |1|$  and  $P < 0.001$ ) to find overrepresented sequences. The resulting motif is shown in Figure 15H (motif on top) and was nearly an exact match to the motif that had been previously derived for Tbf1 (*ORF19.801*) (Figure 15H bottom motif). Tbf1 is a well-studied TR that controls ribosomal gene expression in *C. albicans* [31]. In fact, the RNA-seq experiments also revealed that the levels of *TBF1* transcript changed significantly in the *cph2* deletion mutant (with a  $\log_2$  fold-change expression of 0.8 and  $P = 0.00027$ ). These results suggest that the large fraction of ribosome biosynthesis genes that is differentially expressed in the *cph2* deletion mutant is likely due to downstream effects caused by *TBF1*.

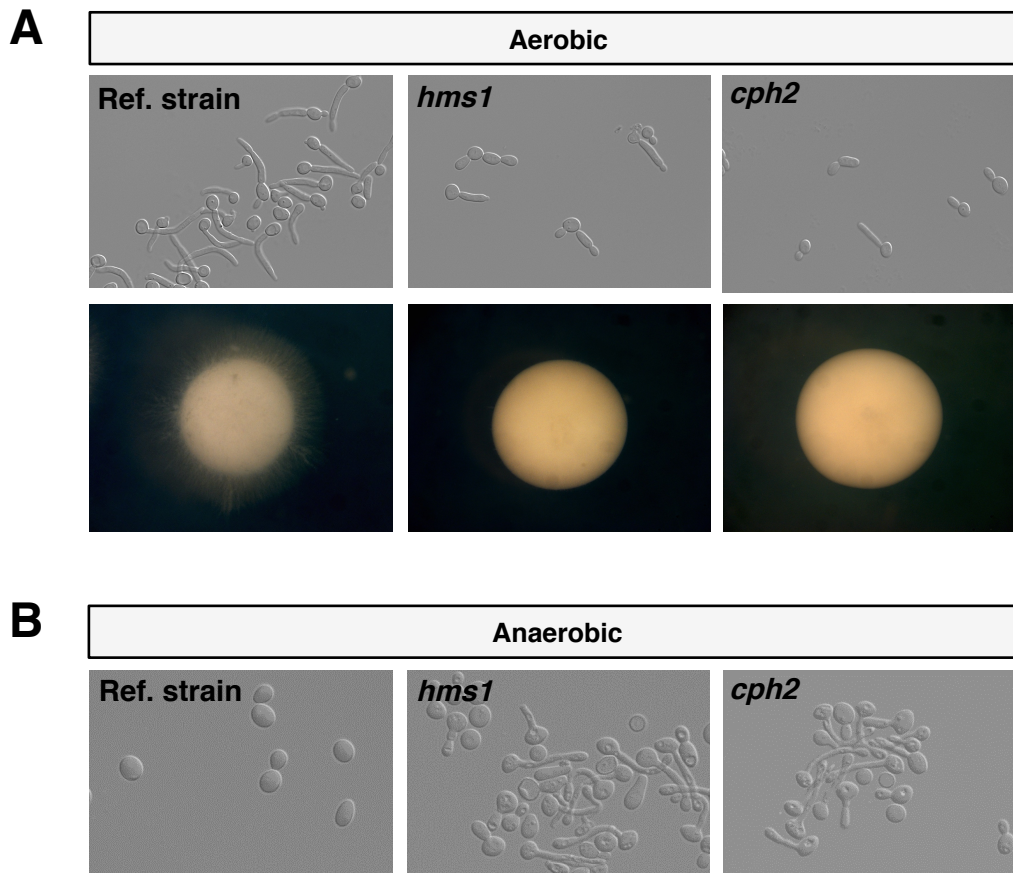
#### 2.2.4.2. SREBPs regulate morphology in *C. albicans*

*C. albicans* is a polymorphic fungus that can grow in yeast and hyphal forms. Multiple cues such as mammalian body temperature (37°C), serum, O<sub>2</sub> tension, pH as well as the presence of some bacterial species have been reported to trigger hyphal formation [138]. The morphology of the fungus impacts greatly its interplay with the host (*i.e.* yeast and hyphae induce different immune responses [107]). Transcriptome experiments (see chapter 2.2.4.1) revealed that *CPH2* and *HMS1* regulate the expression of several regulators of filamentation suggesting a role of these two SREBPs in morphology. In earlier reports, experiments performed under aerobic conditions revealed that *hms1* and *cph2* deletion mutants were unable to form hyphae under conditions where the reference strain would [136,137]. To verify this, I examined the morphology of cells and of colonies embedded in agar at 37°C (a temperature that induces filamentation of *C. albicans* wild-type strains). In agreement with previous studies, both *hms1* and *cph2* deletion mutant strains showed a lower degree of filamentation than the reference strain under aerobic conditions (Figure 16A).

According to the RNA-seq data presented above, *HMS1* and *CPH2* control several regulators of morphology. The RNA-seq experiment was performed on cells that had grown in a completely anaerobic environment. The direction of change in expression of filamentation-related TRs indicated that *HMS1* and *CPH2* would be defective in filamentation. To establish whether indeed these two regulators function as predicted by the RNA-seq experiment, I examined the morphology of cells in anaerobic conditions at 37°C in Todd-Hewitt broth. This setup resembles the anaerobic environment found in the mammalian GI tract and supports the growth of other bacterial species that cannot grow in conventional media. Microscopy analyses demonstrated that under these conditions the wild type reference strain predominantly (>99% of cells) adopted the oval-shaped yeast form whereas the *cph2* and *hms1* deletion mutants were starting to form hyphae (Figure 16B).

It has recently been shown that filamentation in *C. albicans* is detrimental for intestinal colonisation [108]. Since the *hms1* or *cph2* deletion mutant strains are impaired in their ability to persist in the murine gut [86,139], the fact that these strains fail to retain

the yeast form in anaerobic conditions suggests that *HMS1* and *CPH2* may promote gut colonisation, at least in part, by preventing the yeast-to-filament morphology transition.



**Figure 16** The *C. albicans* SREBPs *CPH2* and *HMS1* affect morphology.

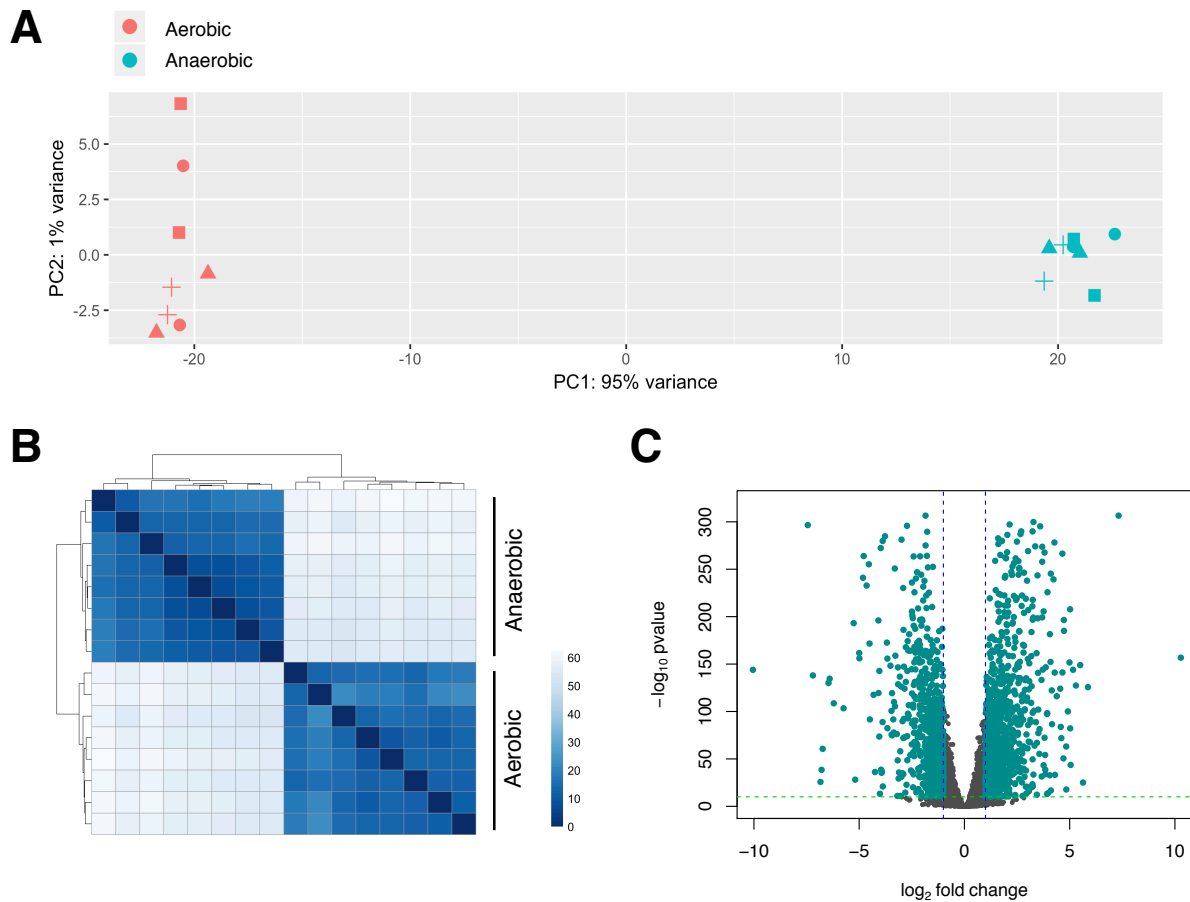
(A) Cell and colony morphologies of cells grown in standard YPD medium at 37°C (top panel) and then plated in YPD agar (bottom panel). Both *hms1* and *cph2* deletion mutants form less filaments than the reference strain in normoxia. (B) The wild-type reference strain, *cph2* and *hms1* mutants were grown in Todd-Hewitt broth at 37°C under anaerobiosis for 24h. The morphology of cells was examined by microscopy. Under anaerobic conditions both *HMS1* and *CPH2* prevent filamentation of *C. albicans* cells. Examination of colony morphology under anaerobic conditions was not possible due to the extreme low growth of the cells and small colony size.

### 2.2.4.3. Anaerobic conditions induce changes in the *C. albicans* cell wall

As shown in the previous section, the presence or absence of oxygen in the environment is a critical factor that influences the morphology of *C. albicans*. Both *hms1* and *cph2* deletion mutant strains were less filamentous than the reference strain under aerobic conditions and yet produced more filaments than the wild type reference strain when grown in anaerobiosis. Thus, changing levels of oxygen available lead to opposite morphology phenotypes in two SREBP deletion mutants.

Two datasets reporting the changes in the transcriptome of *C. albicans* cells under hypoxic conditions have been generated to date [122,123]. However, there is no reported dataset describing changes in the fungus' transcriptome under anaerobic conditions. Thus, given the importance that oxygen has in the biology of *C. albicans* I sought to investigate how complete depletion of oxygen impacts the transcriptome of *C. albicans* cells. To this end, I compared the transcriptome of cells grown under aerobic versus cells grown under anaerobic conditions. The complete RNA-seq dataset consisted of eight samples (including reference strains and several other TR deletion mutants cultured at either 30 or 37°C) grown under normoxic conditions and another eight samples (which included reference strains, *cph2* and *hms1* deletion mutants among other TR deletion mutants) grown inside an anaerobic chamber.

Principal component analysis (PCA) showed that, despite significant biological differences between the strains (whether they were reference or were depleted of a certain regulator) and the temperature the cultures were grown at, those grown in anaerobic conditions clustered as a group clearly separated from those grown in aerobic conditions (Figure 17A and B). Thus, these data suggest that the main driver of differences between the transcriptome of the 16 evaluated samples was the level of available oxygen present in the environment.



**Figure 17 Anaerobic conditions induce massive changes in the transcriptome of *C. albicans*.**

(A) Total RNA was extracted from 16 samples (with different genetic backgrounds) 8 of them were cells grown under aerobic conditions (in red) and another 8 under anaerobic conditions (in blue). The principal component analysis shows that cells grown under aerobic conditions clearly cluster separately from those grown under anaerobic conditions. Different shapes represent different batches. (B) Hierarchical heat map where each row (and column) represents one sample. Shades of blue indicate proximity of the samples. In agreement with the PCA (shown in A) the main driver of differences between the samples is the amount of available oxygen in the environment. (C) Volcano plot showing the comparison of samples grown under anaerobic vs. aerobic conditions. Each dot represents one transcript. The transcripts that pass the threshold of expression changes > 2-fold and  $-\log_{10} P\text{-value} > 10$  are marked in turquoise.



Overall, the RNA-seq experiment revealed 1957 genes differentially expressed (fold-changes > 2 and  $-\log_{10} P$ -value > 10) (Figure 17C) between samples grown under anaerobic when compared to aerobic conditions. Given the elevated number of genes whose expression changed significantly and the high magnitude of such changes I focused only on the strongest signal for further analysis.

Applying a more stringent threshold to the dataset ( $\log_2$  fold expression changes > |4|) revealed 92 genes whose expression changed significantly under anaerobic conditions (Appendix Table 8), 46 of the transcripts were up- and 46 down-regulated. Gene Ontology (GO) analysis showed that the most significant function enriched in the dataset was cell surface ( $P = 5.55 \times 10^5$ ). It should be noted that approximately 50% of the genes had unknown functions. Consistent with the GO results, I found several GPI-anchored proteins: *PGA13*, *PGA10*, *PGA31*, *RBR1*, *PGA23*, *PGA45*, *PGA34*, *PGA26*, *RDH3*, *PGA38* and *FGR41*. Additionally, chitinases *CHT2* and *CHT3*, which are a major component of the cell wall, as well as the chitin synthase *CHS1* were amongst the most down-regulated transcripts (with  $\log_2$  fold expression changes of -4.5, -6.8 and -4.6, respectively). At this high threshold cell cycle related genes were also amongst the most down-regulated transcripts: the histone *HTB1* and the putative histones *HHF22*, *HHT21* and *HHT2* (with  $\log_2$  fold expression changes of -4.3, -5.7, -5.9 and -6.3, respectively) and nucleosome assembly was the most enriched process in GO analysis ( $P = 0.007$ ) of only downregulated genes. The downregulation of histones suggests a likely stop in the cell cycle and growth which is in agreement with the fact that in the anaerobic chamber the cells were never able to grow to an  $OD_{600}$  higher than 0.3 – 0.4. Two transcription regulators also appeared downregulated in the dataset: *WOR2*, a zinc cluster TR that controls the morphological switch from white to opaque [84] and *YOX1*, a gene involved in cell cycle which normally has its peak of expression at G1/S phase [140] (the  $\log_2$  fold expression changes was -4.0 for both regulators).

Interestingly, some of the anaerobic-specific genes related to cell wall and biological adhesion were also amongst the most significantly up- or down-regulated transcripts in the *cph2* deletion mutant strain (but not *hms1*). These anaerobic- and Cph2p-specific genes include the GPI-anchored cell wall proteins *PGA23*, *RBR1* and *RHD3*, as well as *PBR1* a protein required for cell cohesion and adhesion.

In sum, the data suggest that anaerobiosis induces great changes in the transcriptome of *C. albicans* and strongly affects the composition of its cell wall. Furthermore, the expression of some of the anaerobic-specific genes is regulated by the SREBP Cph2p. Thus, Cph2p, which contributes to host colonisation by regulating morphology together with *HMS1*, may additionally have a role in the cell wall modifications that occur when *C. albicans* cells grow under anaerobic conditions.

#### **2.2.4.4. Intracellular localisation of the Cph2 protein in *C. albicans***

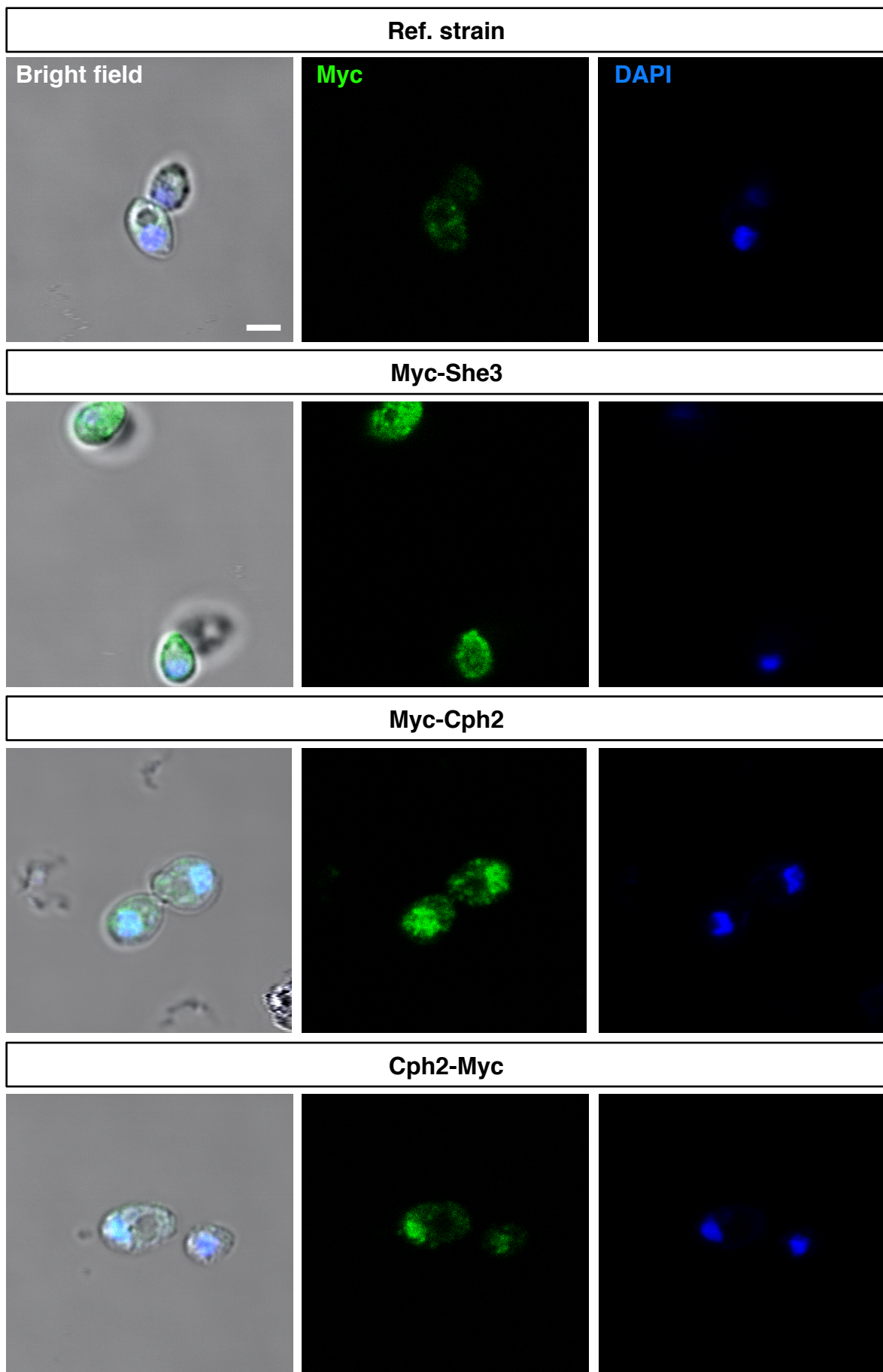
A distinctive feature of the SREBP family is the presence of a transmembrane domain (or more than one) that anchors the protein to an internal membrane keeping the transcription regulator inactive (i.e. away from the nucleus). Upon selective proteolytic cleavage (which is triggered by sterol levels in most SREBPs) the N-terminal portion of the protein is shuttled to the nucleus where it can bind to DNA and promote or repress transcription [101]. With the exception of a few ascomycetes and some *Aspergillus* species, the majority of fungal SREBPs possess at least one predicted transmembrane domain (Figure 3). In the case of *C. albicans*, which has three SREBPs, only one of them, Cph2p, harbours two putative transmembrane domains in its sequence. An earlier study reported that *CaCph2p* localised to an intracellular compartment that is consistent with the endoplasmic reticulum (ER) [95]. The shuttle of the protein to the nucleus, however, has not been established (presumably because the cues that induce Cph2p proteolysis are unknown).

To address this question, I carried out protein immunostaining experiments using *C. albicans* strains encoding either active or inactive Myc-tagged Cph2 proteins (see section 2.2.3.3 of materials and methods and Appendix Table 7 for genotype). On the one hand, to visualise the active form of the protein I used the strain harbouring the C-terminally tagged version of Cph2p (JCP\_880 also used for ChIP experiments section 2.1.3.4), which harbours a Myc tag immediately after the DNA binding domain and does not contain transmembrane domains presumably rendering the protein constitutively active. On the other hand, in order to visualise the inactive form of Cph2p (or at least a

fraction of it) I employed a strain harbouring an N-terminally tagged version of Cph2p (JCP\_960) that allows the detection of the full-length protein (which would be anchored to membranes and therefore inactive) as well as the cleaved and active form of the protein. In these experiments, cells were grown in standard yeast growth conditions (YPD medium) at 30°C. The Myc antibody is unable to penetrate the cell wall of *C. albicans* therefore before immunostaining the cell wall was digested enzymatically to achieve the generation of spheroplasts, which were then stained with an anti-Myc antibody and visualised under the confocal microscope.

On the one hand, the constitutively active C-terminally tagged version of Cph2p (green signal) was found co-localising almost exclusively with the nucleus (stained in blue), indicating that after proteolytic cleavage the protein is shuttled to this organelle where it binds to DNA. On the other hand, the N-terminally Myc-tagged version of Cph2p (which allowed the visualisation of the cleaved and the membrane bound fractions) was found localising inside the nucleus as well as in structures surrounding the nucleus. The fraction of protein co-localising with the nucleus presumably corresponds to the cleaved and active form of the protein. However, the fraction co-localising with structures surrounding the nucleus may correspond to the full-length Cph2p that is anchored to the membrane of the endoplasmic reticulum (Figure 18).

These data suggest that in standard yeast growing conditions a fraction of the Cph2 protein is constitutively being cleaved, activated and shuttled into the nucleus of the cells. The full-length inactive fraction (harbouring the transmembrane domains) remains anchored to the membrane of the endoplasmic reticulum. However, further experiments (potentially using an ER-specific dye) should be performed in order to verify that the structures/organelles surrounding the nucleus the full-length Cph2 protein co-localises with are endoplasmic reticulum.



**Figure 18 The inactive form of Cph2p localises in the membrane of the ER whereas the active form is restricted to the nucleus.**

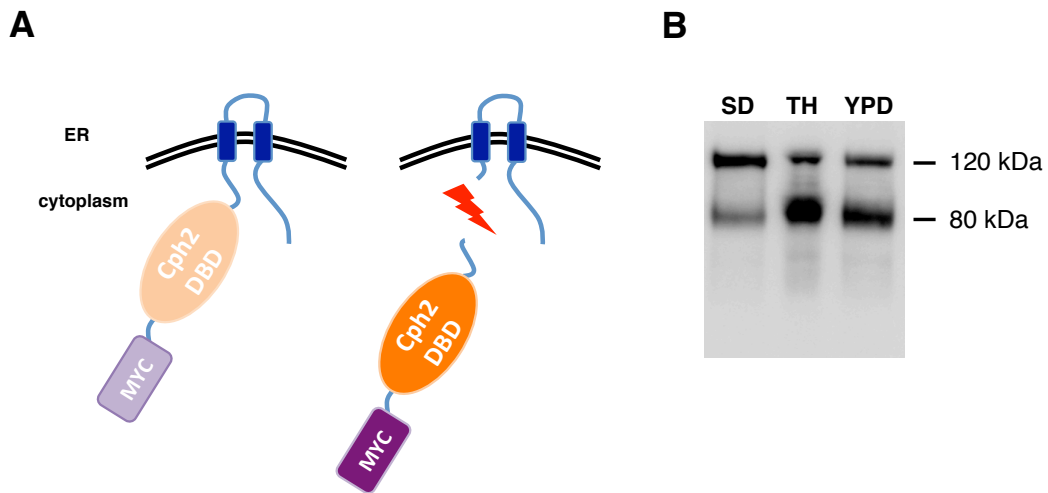
Cells were grown in standard YPD medium and harvested at log phase to generate spheroplasts that were stained with an anti-Myc antibody. Signal from the Myc antibody is shown in green whereas the nuclei are stained with DAPI and shown in blue. Reference strain showing background Myc signal was used as a negative control (first panel). A Myc-tagged She3p, which localises in the cytoplasm, was used as a positive control (second panel). N-terminally tagged Cph2p (third panel) was presumably localised in the membrane of the endoplasmic reticulum (inactive form) as well as in the nucleus (active protein). Constitutively active Cph2p (fourth panel) co-localised almost exclusively with the nucleus. Scale bar represents 2.5 $\mu$ m. All pictures were taken with the same microscope settings. Equal posterior image modification settings were used to adjust all the images.

#### 2.2.4.5. The signal that triggers the cleavage and release of Cph2p from the membrane differs from most SREBPs

Typically, in eukaryotic cells SREBPs are activated in response to fluctuations in the levels of intracellular sterol (cholesterol in the case of mammals and ergosterol in fungi) [56,97,98]. However, in some fungi such as *C. neoformans*, *A. fumigatus* and *S. pombe*, it is changes in the levels of oxygen what triggers the cleavage of SREBPs from the membrane [91,102,103,141]. The *C. albicans* Cph2 protein is the only one the fungus' SREBP orthologs to harbour transmembrane domains and therefore requires proteolytic cleavage prior to its activation. However, the signal that triggers the cleavage of Cph2p and its release from the membrane remains unknown. Previous reports described that the cleavage of CaCph2p is affected neither by levels of sterols nor hypoxic conditions [95]. Here I sought to determine the environmental cue(s) that trigger the proteolytic cleavage of this protein. The *C. albicans* strain used to answer this question harbours a Myc-tag at the N-terminal of the protein (Figure 19A) (JCP\_960 described in [95] and Appendix Table 7 and used in the previous chapter 2.2.3.3 for immunostaining experiments). When the expression of the Myc-Cph2 protein was analysed by western blot experiments two bands emerged: (i) a band of high molecular weight (approximately 120 kDa) corresponding to the predicted full-length size of the protein and (ii) a second band of lower molecular weight (80 kDa), which likely corresponded to the cleaved version that is released to the cytoplasm (Figure 19B).

Transcriptome analyses performed with the *cph2* deletion mutant strain (experiment described in chapter 2.2.4.1) revealed that a series of amino acid permeases that mediate nutrient uptake (*ALP1*, *CAN1*, *CAN2*, *GAP5*, *GAP6* and *MUP1*) were differentially expressed suggesting that the presence or absence of certain amino acids or other nutrients in the environment could play a role in the activation process of Cph2p. Therefore, I used western blot analyses to detect the expression of Myc-Cph2p when grown in rich and minimal media (SD: synthetic complete medium containing dextrose) in either aerobic or anaerobic conditions. The advantage of the Myc-Cph2p construct is that it allowed direct comparison between the fractions of the cleaved versus the full-length protein within the same sample. I found that the levels of oxygen (aerobic versus anaerobic conditions) did not affect the activation levels of Myc-Cph2p (data not shown).

However, western blot analyses showed that when the cells were grown in minimal medium (SD) they displayed a lower proportion of cleaved “active” Myc-Cph2 protein than cells that had grown in rich media (YPD or TH broth) (Figure 19B). The minimal media used in these experiments consisted of a nitrogen source, glucose, amino acids and water. In the case of YPD and TH, the addition of yeast extract, peptone and salts (as well as peptic digest of animal tissue in the case of TH), makes these media much more nutritious than SD. Thus, the findings suggest that the levels of extra nutrients, or perhaps different nitrogen sources (coming from peptides or proteins in the rich media) affect the proteolytic cleavage of *C. albicans* Cph2p from the membrane.



**Figure 19 Changing levels of nutrients in the medium trigger the proteolytic cleavage of the *CaCph2p*.**

(A) Schematic representation of the N-terminally Myc-tagged version of Cph2p used for western blot experiments. The inactive form of the full-length protein (left) is anchored to the membrane of the endoplasmic reticulum with the N-terminal part of the protein facing the cytoplasm. After proteolytic cleavage (right) the N-terminus is transported to the nucleus where it regulates transcription. (B) Cells were grown in minimal (SD) or rich media (TH and YPD) and harvested in logarithmic phase for western blot analysis. The same number of cells was used in each of the lanes. The 120 kDa band corresponds to the full length Cph2p anchored to the membrane whereas the smaller 80 kDa represents the cleaved and active form of the protein. Note that the lane corresponding to cells grown in minimal media (SD) shows a lower proportion of active Cph2p than the lanes corresponding to rich media (YPD and TH).



### 2.2.5. Discussion

In this second chapter of my dissertation I have explored the cellular functions governed by the SREBPs in the commensal fungus *C. albicans*. Tye7p, one of the *C. albicans* SREBPs, has previously been characterised and shown to be a major regulator of glycolytic genes [94]. I therefore focused on the remaining two SREBPs, CaCph2p and CaHms1p, for the analysis.

Transcriptome analysis carried out with *hms1* and *cph2* deletion mutants, under conditions that mimic the *in vivo* environment (when the fungus grows inside the GI tract of warm-blooded animals), revealed that these two proteins form a regulatory cascade. The idea that Hms1p and Cph2p form a regulatory cascade is further supported by other findings: (i) at different thresholds (more or less stringent) the targets of regulation of Hms1p and Cph2p overlapped significantly, (ii) all genes regulated by both proteins changed their expression in the same direction, (iii) deletion of *cph2* resulted in a significant decrease in the *HMS1* transcript; deletion of *hms1*, by contrast, did not affect *CPH2* expression; and (iv), *HMS1* is a direct target of regulation of Cph2p (Figure 15). This suggests that the transcriptional cascade starts with Cph2p as main regulator, which in turn activates the expression of Hms1p triggering further downstream effects that influence the morphology of the fungus (Figure 20).

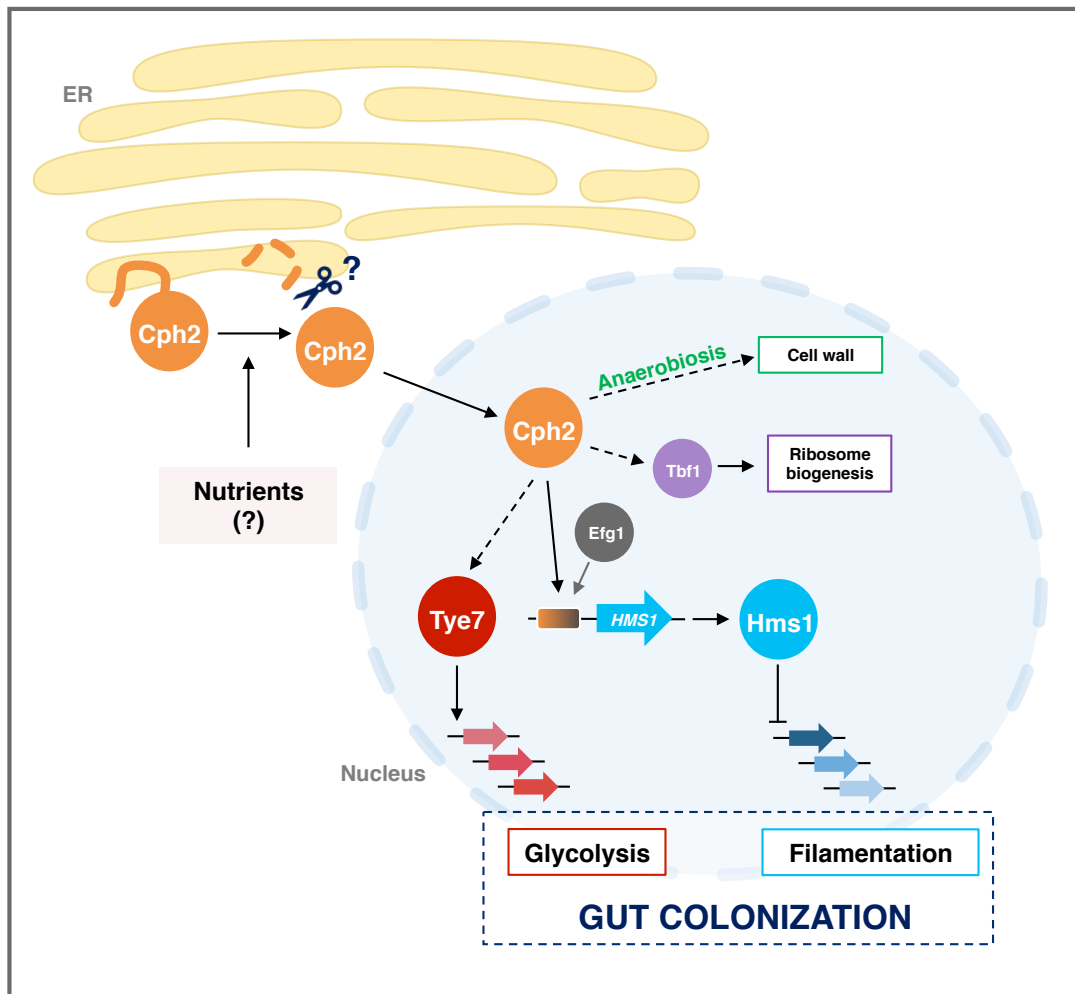
In the fungus *Yarrowia lipolytica*, which is considered an ancestor of the Saccharomycotina subphylum, the SREBP YISre1p has been shown to be required for the yeast-to-hyphae morphological switch under hypoxic conditions [105]. Inspection of the morphology of the two *C. albicans* SREBP mutants showed that, just like in *Y. lipolytica*, the SREBPs have a role in determining the morphology of *C. albicans* under anaerobic conditions. However, CaHms1p and CaCph2p were needed to prevent the switch from yeast to filamentous form, which is the opposite when compared to *Y. lipolytica*. It is known that filamentation in the GI tract is detrimental for *C. albicans* colonisation [108]. Furthermore, both *hms1* and *cph2* deletion mutants have been reported to have reduced fitness in the murine GI tract [86,137]. The data presented in this chapter suggest that the reduced fitness of the two SREBP mutants *in vivo* could be explained by their inability to maintain the yeast form of the fungus.

The GI tract of warm-blooded animals is a natural niche of *C. albicans*. It is an environment largely devoid of oxygen and thus, oxygen availability is an important aspect in the biology of *C. albicans*. In this chapter I sought to investigate the changes that the transcriptome of the fungus undergoes when the cells are subjected to an environment where oxygen has been depleted. I report that anaerobiosis induced great changes in the transcriptome of *C. albicans* cells. It triggered a remodelling of the cell wall by affecting the expression levels of crucial components of the cell wall such as chitinases and GPI-anchored proteins. Two datasets describing changes in the transcriptome of *C. albicans* induced by hypoxic conditions have been published to date [122,142]. It is important to notice that the overlap between the published datasets and the 92-transcript list reported in this study is minimal suggesting that *C. albicans* cells have a very different response to hypoxia (low levels of oxygen) when compared to anaerobiosis (complete depletion of oxygen). At least two of the *C. albicans* SREBPs, Cph2p and Hms1p, play a role under anaerobic conditions. Interestingly, I identified a subset of genes whose expression was significantly affected under anaerobic conditions as well as in the *cph2* deletion mutant. This finding indicates that some of the anaerobic-specific genes identified in the transcriptome analysis were also targets of regulation (may be direct or indirect) of CaCph2p. Some of these anaerobic and Cph2p-specific targets were involved in cell wall structure and adhesion. Thus, this suggests that besides controlling morphology together with Hms1p, Cph2p may have an additional role in controlling cell wall structure when the fungus grows in an anaerobic environment. However, further investigation is required to study the mechanisms by which the anaerobic-specific genes controlled by *CPH2* affect cell wall under anaerobic conditions. These could potentially involve the generation of deletion mutants of those anaerobic and Cph2p-specific genes followed by assessment of their phenotype in morphology or adhesion to tissue under anaerobic conditions.

CaCph2p is the only SREBP from the fungus to harbour putative transmembrane domains [95]. The results from immunoprecipitation experiments presented in this chapter suggest that Cph2p normally resides anchored to the membrane of the endoplasmic reticulum where it remains inactive (Figure 18). Upon selective proteolytic cleavage the protein is shuttled to the nucleus to execute its function. The signal(s) triggering such cleavage have remained unknown until now. Here I show that levels of available nutrients in the medium trigger the proteolytic cleavage and presumably its

activation (Figure 19). This is in stark contrast with all other SREBPs studied to date, for *CaCph2p* it is not levels of sterol or oxygen what induces activation but rather the presence of nutrients. However, the identity of the protease (or proteases) that cleaves the protein remains to be elucidated.

In sum, in this second chapter I have reported that together with the diversification in DNA binding preferences, the SREBPs in the human commensal fungus *C. albicans* also diversified their cellular functions. At least two of the three *C. albicans* SREBPs form a regulatory cascade that controls the morphology switch from yeast to hyphae preventing the fungus from filamenting under anaerobic conditions. The control of morphology under such conditions is relevant for the biology of *C. albicans* since filamenting strains in the GI tract have a reduced fitness [108]. Anaerobiosis induces great changes in the transcriptome of the fungus causing a rearrangement in *C. albicans*' cell wall. Furthermore, I describe that the *Cph2p* protein normally resides in the ER membrane and shuttles to the nucleus in response to changes in the levels of available nutrients present in the environment.



**Figure 20 Overview of the function of SREBP proteins in *C. albicans*.**

Cph2p resides anchored to the endoplasmic reticulum and after proteolytic cleavage by a yet unknown protease(s), which is triggered by the presence of specific nutrients in the medium, it is shuttled to the nucleus. Once in the nucleus it directly regulates the expression of *HMS1* by binding to its promoter in cooperation with Efg1p and prevents *C. albicans* from forming filaments under anaerobic conditions. Additionally, Cph2p might indirectly influence the expression of *TYE7*, which is involved in glycolysis. The Cph2p-Hms1p cascade as well as the Tye7p function are required for colonisation of the mammalian gut [86,94,137].

### 3. Discussion

In this dissertation I have explored the mechanisms driving the functional diversification of a eukaryotic transcription regulator family, the SREBPs. In the ascomycete yeasts, the genome of several *Candida* species encodes three SREBPs. Previous work has shown that in this group of fungi, transcription of ergosterol biosynthesis genes – the main function associated with the family in most organisms – is regulated by proteins unrelated to SREBPs. These observations implied that this family diversified their function in the ascomycete yeasts, i.e. that the proteins adopted other genes and cellular processes as their main targets of regulation. In this study I found that, concomitant with a diversification of the cellular functions governed by the SREBPs, these proteins underwent significant changes in their DNA binding specificities. Several lines of evidence support this conclusion. First, phylogenetic reconstruction of the SREBP family based on the DNA binding domain of the proteins revealed that each one of the three *Candida* SREBPs belongs to a different branch of the family tree (Figure 3), a pattern consistent with these proteins being non-redundant. Second, the three *Candida* SREBPs displayed, to a significant extent, non-overlapping patterns of binding to a comprehensive library of DNA sequences (Figure 4). And third, only one of the SREBPs in *Candida*, Tye7p, bound to the palindromic E-box motif, which is recognised by most bHLH proteins (Figure 4 and 9). In contrast, the *Candida* Hms1p branch exhibited a strong preference for a non-palindromic DNA sequence whereas the third *Candida* SREBP, Cph2p, bound *in vivo* to a sequence consisting of only a half-site motif but likely in cooperation with other co-factors such as Efg1 (Figure 4 and 5). Basic HLH proteins typically bind to DNA as homodimers where each of the monomers recognises one half site of the DNA motif (Figure 1A). Thus, the ChIP-derived motif for Cph2p (which consists of one half-site) suggests that under these conditions the protein might bind to DNA as a monomer. Interestingly, *in vitro* MITOMI experiments showed that Cph2p bound to a motif composed of two half-sites that resembled both E-box and non-E-box DNA sequences (Figure 4C) therefore suggesting a dimer conformation. Previous studies have reported that certain TRs, like the mammalian NFATp, bind to DNA as either monomers or dimers *in vivo* and depending on this conformation, trigger different biological effects [143]. Thus, it is possible that Cph2p

might have the potential to bind to DNA as a homodimer in *in vitro* conditions but find it specific targets of regulation in an *in vivo* context by binding as a monomer in cooperation with other regulatory proteins.

The archetype and most studied member of the SREBP family, the human SREBP1, exhibits dual DNA binding specificity *in vitro*: It can bind the palindromic E-box (5'-CANNTG-3') generally recognised by bHLH proteins as well as a non-palindromic sequence (5'-TCANNCCA-3') [58]. The protein, however, binds *in vivo* to the palindromic E-box as revealed in a ChIP experiment [56]. The results presented in this thesis indicate that the branch of fungal SREBPs represented by the *C. albicans* Tye7 protein shares the same DNA binding features with the human SREBP1. That is, the *C. albicans* protein displayed the same dual DNA binding specificity *in vitro* (Figure 9) but bound *in vivo* preferentially to the palindromic E-box variant (Figure 4 and [94]). Their similarities in DNA binding profile are in stark contrast to the divergent cellular functions that they govern: While the human SREBP1 regulates the expression of sterol biosynthesis genes, the Tye7 protein controls the expression of sugar acquisition and sugar metabolism genes [94] (Figure 20). Furthermore, the former harbours the transmembrane domains that are a feature of the family [144] whereas Tye7p has no traces of any transmembrane domain in their sequences (Figure 3). Thus, the Tye7 branch of fungal SREBPs shares the human SREBP1's DNA binding features despite the distinct roles that each protein plays in their organisms.

The dual DNA binding ability of the human SREBP1 has been traced back to a tyrosine residue in the DNA binding domain of the protein [58]. Most bHLH proteins have a conserved arginine in this position (instead of the tyrosine) (Figure 3A). The arginine residue forms a stabilising salt bridge with a conserved glutamate nearby; such structure underlies, at least in part, the protein-DNA contacts with the canonical E-box [59]. This salt bridge cannot be formed when the tyrosine is present, conferring conformational plasticity to accommodate protein-DNA contacts with the non-palindromic sequence (5'-TCANNCCA-3') besides the palindromic E-box [59]. All the proteins included in this study (Figure 3) harbour the tyrosine residue. Nevertheless, in contrast to the DNA binding pattern displayed by the *C. albicans* Tye7p and human SREBP1, the branch (or branches) of the fungal SREBPs represented by the *C. albicans* and *C. parapsilosis* Hms1p and the *A. fumigatus* SrbAp exhibited a marked preference, both *in vitro* and *in vivo*, for the

alternative, non-palindromic sequence (5'-TCANNCCA-3'). These results suggest that the tyrosine residue that is the hallmark of SREBPs enables alternate binding specificity in addition to dual DNA binding (the latter is what has been reported in the human SREBP1).

The DNA binding assays with purified *CaHms1* and *AfSrbA* proteins demonstrate that their preference to bind the non-palindromic sequence (5'-TCANNCCA-3') over the palindromic E-box is an intrinsic property of the proteins (Figure 6 and 7). Amino acid residues in the first helix and the loop region of the *CaHms1p* were necessary to confer the specificity towards the non-palindromic sequence (Figure 8). Based on crystal structures of various bHLH proteins, the amino acid residues making direct contact with DNA are located within the basic region and first helix of the DNA binding domain [52,54,57,59]: A glutamine and an arginine residues in the first helix and a histidine in the basic region make direct contacts with the bases that comprise the E-box [52]. These three amino acids are fully conserved throughout the fungal SREBPs included in this study (Appendix Table 5). Thus, the changes in DNA binding specificity that we identify in the SREBP family cannot be due to variation in any of these positions. In bHLH regulators such as the *S. cerevisiae* Pho4, residues at the boundaries of the loop region (*i.e.* towards the end of the first and beginning of the second helices) are known to interact to stabilise the overall structure [52]. I speculate that at least some of the amino acids underlying the change in DNA specificity in Hms1p may, similarly, be involved in “stabilising” the structure rather than in making direct contacts with DNA. The fact that the first helix and the loop region were necessary imply that more than one single amino acid change is responsible for the switch in specificity. This finding is consistent with the observation that cumulative amino acid changes—which often must occur in a particular order—are usually responsible for the modifications in protein function that occur during evolution [41,145,146]. Furthermore, these results suggest that “disordered” regions of a protein’s DNA binding domain, such as the loop region in bHLH proteins, may also influence DNA binding specificity.

In contrast to *CaHms1p* and *AfSrbAp*, it is apparent that for other SREBPs, such as *CaTye7p* and human SREBP1, additional factors may contribute to their *in vivo* DNA binding specificity. I hypothesise that binding to target promoters with co-factors may be one such determinant. It has been shown, for example, that the human SREBP1 cooperates *in vivo* extensively with the co-factors NFY and SP1 [56]. The *C. albicans* Tye7 protein has

also been shown to bind to many promoters together with another protein, Gal4 [94]; indeed, both regulators Tye7 and Gal4 are needed to control the expression of glycolysis genes in this species [29,94].

The prototypical SREBPs, and the human SREBP1, regulate sterol biosynthesis. However, in the Saccharomycotina lineage, these proteins have transitioned to govern morphology instead. The regulation of sterol biosynthesis genes was handed over to a protein that does not belong to the SREBP family [105].

The results reported in this thesis show that in *C. albicans* at least two SREBPs, Cph2p and Hms1p, regulate a morphology switch from yeast to hyphae under anaerobic conditions; they prevent the fungus from developing filaments (Figure 16). The regulation of this morphology transition is biologically relevant for *C. albicans*: the GI tract is an environment largely devoid of oxygen where *C. albicans* resides as a commensal. The different morphologies of the fungus have different properties which impact on the interplay with the host [147,148]. Hyphae have been associated with invasion of tissue and cell damage [147,148] whereas the yeast form seems to have a significant fitness advantage in the gut [108]. By promoting the yeast form of the fungus, the SREBPs may contribute to the persistence of *C. albicans* in its commensal form in the mammalian gut. Thus, the SREBPs played a key role in the regulation of a morphological switch (Figure 16) or in sugar metabolism [94] in *C. albicans*; therefore, the diversification in DNA binding specificities was likely key in the SREBPs' expansion of targets of regulations in the *Candida* lineage.

The transcriptome experiments presented in this thesis indicate that CaHms1p and CaCph2p form a regulatory cascade in which *HMS1* is a direct target of regulation of Cph2p (Figure 15). Additionally, Cph2p is the only SREBP in *C. albicans* that harbours transmembrane domains. Surprisingly, the stimulus that appears to regulate its activation is the availability of nutrients, which differs from all other SREBPs (Figure 19). Hms1p, on the other hand, lacks transmembrane domains; hence, the activity of this protein most likely responds to different intra- or extra-cellular signals that remain to be elucidated. The Cph2-Hms1 regulatory cascade could expand the repertoire of stimuli that feed into the circuit to control yeast-to-filament transition [19,149,150].



The fungal cell wall is crucial for maintaining osmotic homeostasis, morphogenesis and for protecting fungal cells against environmental stresses [151]. Moreover, in pathogenic fungi, the cell wall plays a key role in the interactions between the fungus and its host by contributing to adhere to tissue and modulate the host's immune response [152]. In the particular case of *C. albicans*, regulation of cell wall functions have been shown to be significant contributors to the fitness in the mammalian GI tract [45,153,154]. Here I describe that when *C. albicans* cells grow under anaerobic conditions (which resemble the gut environment) the levels of major components of the cell wall, such as chitinases and GPI-anchored proteins [155] change significantly suggesting that they are crucial elements for the survival of the fungus in such environment (Appendix Table 8). Interestingly, a subset of anaerobic specific genes was under the control of the *CaCph2p* expanding the role of the SREBP family beyond morphology (*Cph2p-Hms1p* cascade) and carbohydrate metabolism (*Tye7p*). However, the exact role in which cell wall modifications impact on the fitness of *C. albicans* in the GI tract (whether they influence the host defence mechanisms or play a role in interactions with other commensal microbes) remains unclear [45].

In sum, the data presented in this thesis demonstrate that the fungal SREBPs comprise several branches that differ from one another in their DNA binding preferences and in the biological processes that they regulate (Figure 3 and 20). A key element in the diversification of the family appears to be the intrinsic structure of the DNA binding domain of the SREBPs, which allows these proteins to adopt two distinct conformations and therefore recognise at least two different DNA sequences. Ancestral protein reconstruction experiments indicate that this promiscuous state was resolved during evolution of the family: One branch tilted the preference towards one of the DNA motifs largely through amino acid changes in the same protein whereas another branch tilted the preference towards the second DNA motif. The diversification in their DNA binding preferences likely enabled the SREBPs to expand and regulate diverse cellular processes in fungi.

## **APPENDIX**

**Appendix Table 2 Oligonucleotides used in MITOMI experiments**

Oligo Number	Oligo Sequence
1	GCAATTGCAATGCAAAAAACGTAGCGCAGGTGTGAGTACCCTCGTACGGTTA
2	ACGGTTAAGACTAGCGCCAGCATAGACCGGAGGCGTAAACACTGCATCCGCC
3	ATCCGCCCGCCTTCAATGCGGCCTGAACGAGGGGCCAGATGTGGGTGGTCTGT
4	TGGTCGTTTCTGCTATACTTAAGGAGGAATACGACACACGTCCGCCACAAGG
5	CACAAGGTCATCAAGCTACCCACCTAGCACAGTGCCGGCTGTCCGGCTATGG
6	CGTATGGCAGATCGTCGGGCTGTAAAGGTGTAGTTCAACAGTTTTCTGTGAA
7	CTGTGAACACCGTGCACAGGGCTAAGATGGACTGCCCCACGAGTGAAGGAGG
8	AAGGAGGGCGACACGGTCTGTCAGGATGGTGGACAAGTAATTCGATAACA
9	GCTAACACCGGCCAGGCAAAGCGTGTTTAACGCTAGCATCATATTTGAAGG
10	TTGAAGGTTAGTGTGACTTCGCCCCCTGTGGAGGGGCTACTGATTTACGTA
11	TTACGTAATTACGTCGTCCGCACGGGAGAAAAAAGTGGGCAGTAGACATGA
12	GACATGAAGCGCGTTATGCATGCATCCTACCCTCTGAAAGTCTACAGCGG
13	ACAGCGGTATTGTAACCAGCTATTGCTTAAGTGGAGATTTCAATAAGTAAT
14	AAGTAATAACGTAAGTCTGCCACCTAAGTACACGGAATCTAACCGGTACCC
15	GGTACCGTCTTACAGTCCCGGAAAGCGGTACAGGACGCCGATACTCGCGG
16	TCCGCGGTCTAATCCGTAATGAGTGGTAGCTGTTCTTAGAACCTCTCTG
17	CTCTCGCCAGCCCTAGTTAAGGTCTAAATCTCGCATTTTGCCAATACAATC
18	TACAATCTAGGACACATTTAGAAGGCCATATCCTCGTTGGCCTACTGTTAGC
19	TGTTAGCCAACCTGCGTTCGAAAGTTCACCGTTGGAATTATAAGCTAGATGG
20	TAGATGGTGTGACAGCCAGAAATCGGGCAACGCACGCTCGGGCGTAATCGCG
21	AATGCGGGTGTGGCTATATGTTTAGGCTATAGCTATGAGGGACTCGCCCGTG
22	GCCCGTGGGAATCGCTGAGCCCGGTTGCACACACTCAACGGGAAGTGAGCA
23	GTGAGCAGGATGCGTTCAAGCCCGCGGCCAACTTCCGTAACCACTATTTCT
24	TATTTCTATACCGAATGCTCCTGGGTTCTTGACTCTATCGGCTTCCGAGGA
25	CCGAGGATCATACAGGCCAGAATTAAGATGAATCACTAGCCACTCTGCGG
26	TCCTGCGGGTCTCACACTCGATTCTGATACTACCTTACAGCAGATGACTGG
27	TGACTGGCTAGCGCACTAATCGCAGAGTGTCCGGTACGACTTCCGGTGTTC
28	GCTGTTCTAGACTATATTTGGAGACCGCATAGTCGATAGTATCTGGTTGGC
29	GGTTGGCGTGTGTCTAATGCCATGGATACCAGTTAGTCATTACAGTTCCAA
30	GTTCCAAATTGAACAACCTATAGGTTGGACATTTCTCCGGTTGGGGCTTAGTT
31	CTTAGTTACCATCGCCCGGTTTAAAGCGCGGATCATTACATGAGCGTGGAGA
32	GTGGAGAAGTACTTAAGTCGAGTATTTCCACAGCTCCGACGCGACGTGAGTA
33	GTGAGTAATACCATCTCCAATGGCCGAAGATCTGGGGATTCTATGTGCCAT
34	TGCCCCATATTCGCATAACGGCCTTGGTGAGCTTCATACTTTCGGACAATCG
35	ACAATCGAACCGCCCTTTACCGGGTATACCATGGTACCAAGCTGCTCCC
36	TGTCGCCCTCCGCGTATGATTGGATTCATGCCGACCCTGCCAAGACCGCC
37	ACCCGCCACGTCGTAGTATCGTTCAACGCTTATGTTCTGTAGAGGGTAACA
38	GGTAACATGCCTACTAGTACTACGGCCGTACGGGCATAAACTTTAACATTGC
39	ACATTGCTCTCCGAAGAGCGATGCAGACATAGTGTATCTATCGAGCGTTACC
40	CGTTACCGCTAGCGCTAAGTTGGTCCGTAGTCATGTCGTACGATCGTCAAT
41	GTGCAATGACTCTCCACAAGAGGTCACATTCGTGTACCCGGTACCCTGAT
42	CCCTGATAAATGCCAACGTCAAACTCCCGCTACGGCGAAAAATATTCATTT
43	CTCATTTAGCGCGTGTGAGGAGACGCTAGACGCATGCGCATCCCTTACCG
44	CTTACCGGGTTCATAATGATATCGCTAAACCACCTATGTGGTTGACAAAC
45	GACAAACTTCGCTTTTCACTCCTCTCGACCAATAATGGAGTGATCTCTCT
46	TCTCTAGAAAAACCAGGATGAGTTATCTCGGACCCTGTACGTATAGAGAAG
47	AGAGAAGTCCATATACAAAATTCATTCAGTCATGTTGTGCGACGGTCTCC
48	GTTCTCCAGTCAATTCCTGTGGCAAGTCACCCTACCCGCGATGGAGCGGCC
49	GGCGGCTCTTTGAGATCTAAGTCACGTGCGCTCCAAGGTGATAACGCCTTG
50	CGCCTTGTCCGTACGCCTAAAAGTTAGACTAACGGCGATCCACCGAAATTAT
51	AAATTATCTCCACCGCGGAAGCGCATCGATAGCCTTGATACGCCAGGTTAC
52	AGGTTACTAGTCCACTACGGTAGTACGAGCACTTGCTTTTCTCCCGCGGAG
53	CGGGAGATATCAACCTAAATGCCAATGTAAGGCGACGCGTGCAGTGGCG
54	AGTGCGGGTCAATTTAAGGATCTAAACGTCGATCAACCGCGCAACGAGCT
55	ACGAGCTAAGCCGACGGCCGATCGAGAGTATTATGCATCGAGGACCTAAAC
56	CCTAAACTAGATTATAATTATCCGTATTAAGTGACCATCTTCTACTCAGTTT
57	TCAGTTTTGCCCTGCCCGGCTTCCAGCTCGTGGTTAATCGCTAGCTCTAGA
58	CTCTAGAGCTAATCATTAGCATTTTCTCCGACCATCGGAGTTTTAAAC
59	TTTAAACCGAAGAAGGTAATAATACGTAACCTCCAGAGTCTTTCCCGGAAAC
60	CGAGAACCTCTTCCCTTTCTATACAACCTAAGTAGCGGAACTACGATTGAG
61	GATTGAGTCTTCAGCGTCTTCAGACCATCCGGAGCTAACAACTCGAAAGT
62	CGAAAGTCAATGCGATCCCGCAGCAGTTATAACCGGTTAACGTTAAATTTG
63	TAAATTGCTTTAAGAAGCCCTTAAAGCCAGGCGGTCTGGAGCTACGTGTTT
64	CGTGTCTTCTCGTTAGCAACAGTGTCTCGGTACATGTCTCGAACAACAGTAA
65	ACAGTAACGTCGGAAGCGATCAATGCTTCATCTCTCCGGCTACCATGAAG
66	CATGAAGTACGGCGCCGCGCGGCATTTGACTTGGTCTTACGAGAATTGGC

67	AATTGGCTACGAGGAGAGTGTGCTTTTATAAATATTTTCGATACAGACTGCCG
68	ACTGCCGGCATGCCGCTGATCCAAGCCAGTTGGGGGAACTCCAGTGGCACGA
69	GGCACGATTGCAAGGTCCGGAGGTTCTCGGGAGAGTTCGATCGCCAGTGGAT
70	AGTGGATTCTGTACCATTTTCGTACGTTAGAGGGCCGATATTTCACTGACCA
71	CTGACCACGACATATGCCCTGGTTTCCAGCAGAATCTGCGTGGACGCTTTAA
72	CCTTTAAGCCAAATGGGAGACTCCTCCGGCCCTAATACAGTGATACAAATCG
73	CAAATCGGCGACCATTGCATCCCCTGAATTACGAACTTAAACAGGTTCTGT
74	GTTCTGTGTCAAACGAAAGTTTCCGGACACTAGCGTGGTTGTACTACACA
75	CTACACAAAGGGCACCGCAATGCTGAGCACTGAATTCATGAACTCCGGCATT
76	CGGCATTGGCTCCAATTTATGTCCGCGCAGATTTTAAACAGCCACGTACAAT
77	GTACAATTCTGTTACAGACAGACCTATCACGCATGATGCACGCGATAAGCCG
78	TAAGCCGTCCCTTATTGTGCTATAATAAAGTTCGTCTACCTATCGGACATCCC
79	ACATCCCGATGGATATTGGACTGGATAGGGAATAGTATGTTGGTAGTCCC
80	TAGTCCCTTGGAAAGGCTCTCTCGTCTTAATCGAGATTAGGCCTAGGCCAGCC
81	GCCAGCCGCGAATGGGTAACGTTAGGTAATCGGGGAGAGACGTAGAGCC
82	TAGAGCCTACCATCCAGGACTCCTTTCATATCTTCCATTCTATATTCGACG
83	TTCGACGTGCACCTTCCCTAAGTCTGGACTACTTCGAGTCGATGTGGCGAGC
84	GGCGAGCGAGTGGCCAAAGTTAGCCGTGCGGCTGCTAGTGGGAGGTGGGAA
85	GTGGGAACCGCTACTCTGATACACAGCCTTAACTGAGCCTTATTCTCTCCG
86	CTCTCGCAGTCAAGTTATGATTCCGAGCGGAAAGCCACTGATCCTTTGATT
87	TTTGATTGAATGGATTGCGAGTAGGAGGGTCTTAGCAGTTTGTGAAATGTT
88	AAATGTTTCTTCTTGGTAACCTTCTCTGTCGGGCTCAAATAAGAACGGC
89	GAACGGCGACTAATCTATCTCTGTAGGTCAACCTGCCGAGATCTTTGTAAG
90	TTGTAAGCGTAAGAAGGAGAATGTGCGTCAATTTAGTATCCTGAATAACACT
91	TAACTCATGAAGGATAAAAACTAATGTTAGCGCAACGTACTATCTTATGC
92	CTTATGCAGTTACTTTTACAGACACTCCTATTCCGGTGTGGAATTTCTCTTG
93	TCTCTTGGCCAGTCTTATTCCGTAGCTAACTGACTCACGCGCGGCAGCT
94	GGCAGCTAACTGGGTACAATAATCTTGAATGTCACTACGTCCAGTAGAT
95	AGTAGATCTTAACTGATATACGCGCGTTGCGTCTTTTGGGTTCTTACACTC
96	TACACTCACAAATCAGCTTAAAGTTGACAGGAACTTGAAGTATATTATGGACA
97	ATGGACAACGTGTGGAGTGCGAACCCCATACCATAGTGACCCCGACTGAGGT
98	CTGAGGTTGCGAGAATGACATGTCGACACAGTCCAAATGAACTAGCGAATCT
99	CGAATCTCCTTGATCGGATGCGGGCCTTCGATCATGTTGTAACCCACCG
100	CCCACCGTTCGATGGGGCTCCGTGCGCCACTTTATGAGCTGAGGAGCAGAT
101	AGCAGATAATCGATAAAAAGAAGGGTGGCATCAAATTTAGTTGGGACTT
102	GGACCTTAGGGTTAAAAGGGGTGCCCGTCTGTGACGCCCGCTCTCGTCCG
103	CGGTCCGATGGGCGAGGACCTGGTTGAGGTCAAGCCACCGCTAACCATATAAC
104	ATATAAGCTGCGTAACAAGTATTGTTTCAAGTGAAGCAGACCACCGGAAG
105	CCGGAAGGTCTGCCGTTAATAAAAAGGACGGAACGTAATCAGTTTACTTGGG
106	ACTTGGGACATACTGTATCATGTGATGTCCTCGGCACCAAGCTGTCTTGG
107	CCTTGAGTTAACATCCATACGAGCATCTCCTCTAGTAGTGGTGGCTTGACAA
108	TTGACAATTATATACAGCAACCTGTCCGTGCAATTTGCCACTCACGGCAG
109	ACGGCAGTGATCGTGATTCTGGTCTCCCTTTTCCATATGCTGCATATTATTA
110	ATTATTAGAAGTCTATCTGCGCATACAGCCGCGGCCCAAATCTACCTGGT
111	ACCTGGTAAGGCCCCCTCAGAATTCGGACGTTTCTCAGTTGGACCCAGGTTG
112	CAGGTTGTCTACGGCAATAGAATTCACACTACGAACATGTCGCGCGGCCG
113	GGCGCCGACGCTCGTGTACAGTAAGTCCCGCTCACCCCATGAATATACCG
114	TATACGGAAGTTTCGTGGATCTCGTCCCATGAGAACCACAGTTTCGCAACGAA
115	CAACGAAATACGATTGCCACCGATTGTAACCAAGGAACCATTATCCCCCC
116	TCCCCCGAACTATCGCACAATTTTCAAGCAAAACAATGACCGGATGACCCAC
117	GACCCACGCTTTTAGGTATTTTGTCTTTGATGAGTCATGACGTCAGTTGAGT
118	GTTGAGTTCCGGTACAAAGTGACAGTCAACGTCGCACCCGCGGCCGCCAC
119	CGCCCACTGAGATTCTCGGTATGTTAAGAGCACTAGGGTGGCGACGCGTGAC
120	GCGTGACGGAAGCTACATAGACACCGCATTGTTGGTGCCTACTTGAACAGT
121	GAACAGTACCTGCGAAGGTCAGTTTGTCTCAAGAGCCGACTCTACGAGACGTT
122	AGACGTTTTGTGGCCAAATGTAGAGTACACTAACGCACACCACGCCGCGAG
123	CCGCGAGTGCAAAACGGGATAAGAATAGACCCGTGACATCGTCCAACATCCC
124	ACATCCCATGTAACCCCTGGGGTAGTGATGAAAGTGCCCGCAAACGCTCAA
125	CGCTCAAGTTTTTTTACCCTTTCGGCTCTGCGTTTTCCCTCCCGTGGTATGCA
126	GTATGCATATGCAAGGAGTATATGAGACTACGAGCTCGAGCGAAATGAGTCC
127	TGAGTCCGGACGTCCTCAGGGGGAGGCCACGCCACCCAGGGCACTCGAAC
128	CTCGAACGAAAAAAATTCGGGCTACAGGATGTAATTAATTAAGTTACGGG
129	TTACGGGCGTTCCATGTTGCGGGTATCATCCGCACTATAGAGCTTCGATATG
130	CGATATGGGTTTGTTCGCTTATCCGAAGCTGTAACCTCATAGCACGGCACCC
131	GGCACCCATGATGGTCCGCTTGCCTCGTTAGGCTTTAAAGCTTTGTTTGTTF
132	TTTGTTCAGTATAGTTACAACAAGGGTCCCGGTCCTGTGAAGGTCGTGGCA
133	CGTGGCAGCTTGTGGTTGCAACGTTGTAAGGATTTTTTTCAGTTGCTATGCGC
134	TATGCGCGAAAGGAAACAGTCTGAACTTTTCTTAGTGAGATGAGTATGGT
135	GTATGGTTCGCTTTTCGCACAGGCTTATGGATAAACATCTGATAAAGGACGAT
136	GGACGATTACAAGTGGTCTCATTAAGAGGGACCCCTCCCGACCACTTGAT
137	ACTTGATTAAGGTTCAACTTCACTATAAGAAACGGAATTTTCTTGTGACTA

138	GTGACTACTACTTTTCAAAAAGAAATCCGCTACCTAGGTACCTAGTAGATGGA
139	AGATGGAACACGCACGGGAACAAGTTGATTTAATAGCAGTCCTTGGGAACGC
140	GGAACGCTCTTGTGAGTACTGCCCATATATACCCCTAGGTGACACTTGTG
141	ACTTGTGAAAAATCTGCTTGTGGTACCGAACAATGTGCACATGTGTGTGAA
142	GTGTGAACGCCACAATCCGGATCCGAGTAGCTTAAACTATAACTGCTGGTTC
143	CTGGTTCCCGGACCACGGACTACGCTGCCGGAAGTCCGCGCGCCAATAGC
144	CAATAGCCGGTCTTCGCTTTGGATCCTAAGTGAATCCCGATTCTAAAATGG
145	AAAATGGCTCGTGCCTGAAATACAAAGCGGGACTTCAGATCATCTAATTACT
146	AATTACTGGCCAGCGGACAGTTAAGCTTGGAAATCCGGGAGGTAGATATTCC
147	ATATTCCGAGCAGTATCGATCAGTAGACATATTGTGTTTTACGGAGTGCC
148	GAGTGCTATGCAGGCAATGCCACTTCTTAGGCCCCGAAAGCGCGCAAGGAG
149	GAAGGAGTCCCTGGGACAGGGCGTGGGTCGAGGATTTATGATCTCGAGCA
150	TCGAGCACCCCTACGCTGAGGTCTCTATTTCGATCTTTCCGCTCGTAAGGTG
151	TAAGGTGTGCTACATCCCGTACTCATTCTGCCGAAGGAAGGAATGACGTTCT
152	ACGTTCTTTTTTGGTATTTCTTCAAGATCCTGGGGACCGATAGCAAAATTG
153	AAAATTGCGTGTAGCTACATGAATGATGAGAAGTTCGCTCCCTTGACGA
154	TTGACGAGTGGCGTCTCGATGGAAATGGAGTCGTGCCGTACACGCCCCCGT
155	CCCCGGTCAAACACCGCGTTTGAGTCGCACTCCGCACAAAGTCCCTAACAC
156	CTAACACGCCTTCGCTCTATCACTCAAATCGTACTCTTCCAGGTGATGATGA
157	ATGATGACAATACATACATGCAGAGTTTTTCAGACGAGCAGTTCCTATTGCGG
158	ATTGGGGCTCGAGTCCATTGAGAATTATCGGCAACGGCCAATTAGAATAA
159	AGAATAAAGGGATCCGGAACACTGACATAGCGTACTAAGTGCTGTTATGGC
160	TCATGGCCCGGAGCCTTGTTGCTGGGATTTGGCCCGCTTAACAGAACTCCC
161	AACTCCCATATGTATCCAGCCTATTTCACCGGGGAACACCCCAAGAAATTAGG
162	AATTAGGTGATCATAGGCCTCTATAAAAATAACTAGCCTGGGCGCAAAATGGG
163	AAATGGGGGTCAAAGTCACTGAAGACAGACGCACGTTGGCTCGCTTCTTCA
164	TTCTTACGATACGTTTCCGTTCTCTCTCATAGATGCTGTAGGGCGATCTAA
165	GATCTAACTACGTGGTCTGGGATCTTCCGCCCCACGGAACCAAAATCTATA
166	ATCTATAGATATCTGTGGTACGCCTCATACGCGCCAGGGACATGTGGTGCA
167	TGGTGCAGGCTGTCTACCGCTCGACAGGTACCCAAACACGAGTAACACGACA
168	CACGACACCTCCATAAAGAGACAGGTCAAACCGGAACCGTATACCTGCAG
169	CCTGCAGACTTAGCCAAGGTTTGATTTGCCCGAGGTTAGGGCTATACACA
170	ATACACACTGGACTGGCATTTCAGACAGTGAAGATTGGAGTCTATACTCGGGG
171	CTCGGGGTTTGGTTATATTTACCATACACAAGGAGCGATCGGTGACAGGTG
172	ACAGGTGCTGTCTCAATCCACGTCAAGACCGACCTAGGACGGCGGTGTGCAC
173	TGTGCACTTAACCCATTTAACACGGTGAAAACCCCGTGATGTACGGTGACGC
174	GTGACGCATTTACTCCTGTCCAGAGGTATGAACAAAACCTACCAACTGTTAC
175	CTGTTACGGCATCAGGGCTTTGGTTTGGGCGGGCTGCTTGTTFACCAATTGT
176	CAATTGTCTTGAAAGAACATCACTTCAACCTTACTAAAAAATTGAGCAGCAAT
177	CACGAATACACATGCAACTGTCTCGTCAGAAAAGAAAGTCGTGATCCCACGCA
178	CCACGCACTGGGCAGCAATCGAGCCAGCTCACGGTGGTAAGACTTGACAGA
179	TGACAGATGCAGCTCGATCCTCCGCTTCTCTCAAACCTGGTCAATACAGCTGG
180	CAGCTGGGATAGTCCAATGAGCGCTTTCGGGTAAGTACTAGTTCGAATGT
181	CGAATGTACCTACCAGCATCCGTTCCAAGTCTGACCAATTCCTCGCCTGC
182	CGCCTGCGCAGCGCTGACCTAGAGGGCAATAAAAACATAAATAATTAACCTAC
183	AACTTACACCGAATAATACCGGAATAATCCATATCAGTACAGATGTCAAAA
184	GTCAAAGTGTGATATTCCCAAACCTGTATAAGAGGCTCTTCTGATTGGCGGT
185	TGGGCGTACATGCGATGTACAACCGGCTGGTCAAAATGCCTAACAGAGCAAA
186	GAGCAAATCTCAATAGCACCTAGTCGAAGTACCTCCAGGAGGAGGTAAGC
187	GCTAAGCATGCAAGCTAACGCCCTAAGATTCCGGTTGGTTTTTCTGGGTCCG
188	GGGTCGGTAGCCGCCCTTATCTTTGACTGACGCTATAGGGGCCGCTAAGAGC
189	TAAGAGCTGAACTGCCGACAATTTACGGTCTGAGATGGGAAGCCTACGTGC
190	TACGTGCGAGACCAACGTTTTAGATTAAATACACCCATCCTTGTCTTAGATG
191	TTAGATGTGAGAATCGCGTCCGAGGCTATGGTTGATGAACATTACTACAGC
192	CTACAGCTAGGAACTGCGCACGCATTGGTCTACTTGGCTGGTAGAATTGCT
193	AATTGCTAAAGCGAGGATGATGATGTGCGGCACCTCGCACCGAACAGACGG
194	CAGACGGGAGTAGATACTGTGCATCGTGCGCCCTACTCAAGTGGCTACTCA
195	CTACTCATGTGTTTACGTAAGCGACCTTATCGGTAATGGCCAATCGACGCA
196	CGACGCATCGTTGGAACCTGATCGTTTTGGTCCGTGATCCGGCTGAGGGCAG
197	AGGGCAGTGTAAATCTTGATCAGTAATGAAGGTGGGATTGACTCCACTAA
198	CCACTAAGGCCCCAGTCGATTAGACGAGGCTTAGGGCGGGTAACTGTGAGTT
199	GTGAGTTCACGCAAGCCGTGGTTCCTAGGCCCTCGGAGAACACAAAAATC
200	AAAAATCAGAGATTTTTCGGAAGAGCAGAGGCCCATTTTGGGCAGAGCTTGT
201	AGCTTGTGAGGGGTTTAGAAATCAAGAACAATCTACGGTGCCTGTATTTG
202	GTATTTGTTGGGGTCATGGCTACCTGTTTGGCAAGCTGACGTGCCCTCTGGC
203	CTCTGGCGGCAGGGATCAAGTAAGCGCTCCCGCGCCGCAGTGGAGCACTC
204	AGCACTTAAGCCATTGACACTGTGTTTTTACGGTATTCTTCTTTCGATGA
205	TCGATGACGCCGGGGCGCAATGCCTTTGGACGGATGTTCCGATACCTGAA
206	ACCTGAATCAAATCCGGCCGAGTACCTAAGGGTAGCGCTCAATGTATCTTA
207	TATCTTAGAGAGCGACTCCTACGTCGGTACGGTATTCCACGCGTATTATAA
208	ATTATAAATGTTATTAATCTTACGGTTCGCGACAGACGTGTCTAGTCTACA

209	TTCTACACCGCTTGGCTCTGATCTCAGGGCGAGCCGCTTCGCTCCCGTAAG
210	CCGTAAGCCGCCATTAGGGTAAATCCCGTGTAGACAATGCTACACTTGCACA
211	TTGCACATACTTGAACAGTATAATTAGCGTGTCTCGGACTGAAGGCCTGA
212	GGCCTGATATAGGGACAAGAATATCCGTCATCTCCCTGGTATGATTGTGACC
213	TGTGACCAGACCCTTCGGGCATCTGAGCGACAGGACAGTGTGCCACGTTTTTC
214	CGTTTTCCCGTATCCACCCCTCTGACATTAAGCTCTCATTGAGGTTGGGTT
215	TTGGGTTCAATATGCGCTCACAGCGAACCCGACACGGGTAAAGTTGCAA
216	GTTGCAAACCTAGCTACAGTAGCATGAAAAATGTATTAACACCTCAGGAATCC
217	CGAATCCCTAATCCACCTCTATCCTTGACTTACATAGTCTCGCTCATATTGC
218	ATATTGCCCTCCCTCAATGAGGATCTGTCTGCCAAAGCGACTGACACGTACC
219	ACGTACCAAACATAGAGGCAATTCATCCCCCACACAGCGCTAAACGCATG
220	ACGCATGTTCTAGAATTCTCCGGAAGCAACATAACTGAGGCCACAGCTGAG
221	CGCTGAGTAACGGACCGGACAAAGAGCGGCCTTCCTGGACTCACTGTTAAC
222	GTTTAACTTCTGGACGAGCGCAATTGTATCGCTCGGTGCAAGCATCACTGGA
223	CACTGGATGACGTATGGGTGGGACGAGACATTCTACTTGTATACCTGCA
224	ACCTGCATGCCCAGATATGCGTAGTAGGGATGATTCAACTATGACTGTGCC
225	CTGTGCCCTACTCTCTATCTTTAATCGGTAATAAATTATTGCGCGCGTAC
226	CGCGTACGTAGAACAAAGCCTTTACCCCGTCTCAAAATAGAGTTGTGCCAAC
227	TGCCAACCGCAGTCGGCAGCATTACTTTAAAATCGTGTCCGTGGTGAACGC
228	GGAACGCAATCGCCGACTATTGTTGATCCACAGTCTGGTACACCGGTTCCAG
229	GGTCCAGTTACGACAACCGAGTGTGAGAAGTTCGCTTCGGTGTAGTTGTGCGA
230	TTGTCGATAATACAACAGAACATGCTGGGTTCCGATAAACGCCACCGGA
231	CCACCGACGAGAAAGTGAATAACTGTTCCGTAAGTGAACAATCACTCTTCGTA
232	CTTCGTAGCTGGGATCTATTATTACAGCGGTAATCCCGCCCGAACCGCT
233	AACGCCTGGTGCATATGAATGGTAGCGTCCATTGCTATTCTACGTTCCGCT
234	CTTCGCTGGGTTAACACAACAGGCTAATTCTTCATGCGGACTGCAAACGGA
235	AAACGGATAGGCCATGGCCATGTAGCCTTCTGGCAGGTGAGAAGTATCCGTG
236	ATCCGTGGAGCCGCTTAAGGCCTTAGCGACTTCAATCCGACTAGAGAAACAA
237	GAAACAAGACAGCGTGAAGGATCCTAGGAATTTGCTGTTGACATTTACGAT
238	TTACGATACATCATGTCAAGGGCATTTCACCTTTCCGCCATGATTCCGCGGC
239	TCGCGGCTAGTCGACTAGTCGTATCGCATTAAGTAGGTGCGCGCTCCCTCGG
240	CCCTCGGCTCATTAGTCACTAGACTCGGCTAACCTGAGCTATGTGTAGAGA
241	GTAGAGAGGCGGGCGGCCCTGTAGAAGCACGCTTCTAACCTCGTCCGCGCT
242	CCGCGTCTATATACTGGTGTGAAACCATCTGGAATGTATGGCCGGCAATG
243	GGCAATGAAATCGCCAGAGCGTCTTAGGAGCTAGGGCCCGAGTTACATCTC
244	ACATCTCGAAGACTACTCGCTGATAGGGGGTTCTTGTTAGACGTGATCGGT
245	GATCGGTCTAACGGGACGACGAAAACGGTTCGGTTAGTTATATAGCATGCTAG
246	ATGCTAGCAGGTAGTGGGTTGATTCCCACCGGATCGTAGGGCCTTGCAGG
247	TTGCAGGACAATGGAAGTGTATAACAAGCTAGCCGTATCACTCCCTGTAT
248	CCTGTATCTAGCAATATCAGAACACGGGGAGTGCCTAGTGGTGCGAACG
249	GCGAACGACCGAAGATGATGCCAGCACTCGGCGGTGCGACAAACATTGTA
250	CATTGTATGTGCAATTGCGACCCGGACATAAAGTCCAGTGGTATTCCGGCC
251	TTCCGCTAGACCGTTCAAACAAGTCTAGAGTTAAGAATCTAAGGGAGGTGC
252	GAGGTCTTTGGGACGACCCAGAAACCAATCTGCAGCACATATAAAGCGTAT
253	AGCGTATAACTTCTTATGACATCAGTTAGCACTCACCATCACCAGTAGGG
254	GGTAGGGCTCAACCTCTCCGTTAGGGTGTGAGCCATATAGGCAAAAACCTCG
255	AACTCCGAGCAACAAGTGACGATTCTTTCCGATCGTTAGTAGAGGGCATATA
256	CGATATAGCGGATGTCTGGCCATCTTGACCGTTAAACGCGCACAGACCAGC
257	GACCAGGCACATAACGCGATGAACGTATGACACCGAGCTTCTATGGGGAA
258	TGGGGAAAATTAATTCGCCGTAGAATTACCTCTCGTAAAAGATTGTCCGG
259	TGTCGGGAGAACGTACACTTAAACCAGCCGTCGACGTGACCCACAATTGA
260	CAATTGAGTGCCAGTACCCTTCGGACTTTAGCGTATCTCCAGTAAAGATG
261	AAGAATGGTCTCTCAGCGGAAGGACATGAGATCGTGTAGTGAAGGCCGACGG
262	CCGACGGATTGCGCATTGCTGTGACAGCGTACAGGCTCCGCCAGCCAGTGCC
263	CAGTGCCAAATTTATGAGGATGTGCGATATGAGTTGATAAAGCATCCTTT
264	ATCCTTTTCGCATGCCAAAATTACAGCGATCACCACAATCTCATCCGTAGG
265	CCGTAGGTGAGACCCCACTGTAGATGCGAATTTAGTTAGTATGGGAACATCT
266	AACATCTACGTGACACACTTAGCTTTGGAATATTCCGCAATAGGGCTAGTAA
267	CTAGTAAATTCGCGTGTATCGGATACCGGTGCGATTAAACCGAGATTCCCGA
268	TTCCCGACTACCTCGAGAAGAGGACGGGGTTCGGGGCAGCCCTACTTTATAC
269	TTTATACCTCAAGAATCCGTTGACTGTTTTGTACCACGGTGCAGGCCATA
270	GCCCATAGTCACCTTCGAATAAGTGGCGCACGGATCTTAGTAGTTCTTAAGC
271	CTTAAGCGTCTAAGCTGCTGCTCTCGCCAGCGCCGCTCAACAGACGGCGACT
272	GGCGACTTTATTAGGTAGTTTACCTATGAGTGAGACACTGATGACGATCCTG
273	GATCCTGTGTGACGTGTTCAACCGGTGAAAGTGTTCGCGATTGTGTCTTTT
274	GTCTTTCTGCCCGTCAAAGGACAGGATCAATAATGACGGCATTAACTTATA
275	CCTTATATATATCCACGCTGATCAGCTGAGATACAAGGTGGTACGAAGGAC
276	GAAGGACGCTTGGACGTTAACAGGAGCTCCGGAGTGACCGGTTCCAGGCTT
277	CAGGCTTCTATTCAACCAAGCGCGACGCCCTTCAAGCGTTCGTGCGGGCAA
278	CGGGCAAGACACCCCTGCGAGTGAGTTACGTAATTTAATGTGGCCAGCAGCA
279	AGCAGCACCGCTCAGTGCAGAGACCTGGGCGCAGACTTCGTGGTGGGTGAC

280	GGGTGACCTTTTGGATTGCTCCTTAAATATAAAATCTTACTCGAGGGTCAC
281	GGGTACGGAAGACGGAATGAGCATTAGTGATCCAGACAACGCGGACGCTC
282	GACGCTCAGCTACTGGTTATTTGTCTCTCTTTCGTAGAAATGACGAGCCCTG
283	AGCCCTGCTGCCATTGTAATCTCGTTTAAAGTCCGTAAGGACCGCAACTCGT
284	AACTCGTCTCTAATAGAAAAGGTATATGTCCACTCTGGTTGTAGTTGGT
285	GTTTGGTGGGCTTCTGCAATGTAAGTACTCAGAGAGTGAATACTGGCTACTCTG
286	CACTCTGTTGGGTAACGGTGTCCATGGGTCCGTGTGAAACCTGTAGGATCAC
287	GGATCACATCCGGCAATTTGACGCCTACGGGTGTACATTTCTGTCTAGCAG
288	GTAGCAGCATGGGTTGTGGGGTCTTGTCTGTATTATAGCTAGCTAGAAGACG
289	GAAGACGATCATATATCGGAGATCCCCGCGGGAATGTGGGGATAGCCGCAT
290	GCCGCATGCTGACGATGATTTCCCTTCAACACGGAGGGTATCGGGGTGGGGCT
291	TGGGGCTGACCATGGAAGAAGTCAATTAACATGGGGTAATGCCTGTGTACACG
292	GTACACGTGTACAGTACATCTGCAAAATGGCATAAGCGCTAGACTGACTCGTC
293	ACTCGTCTCACACAAGTACGCACTTATCTCTAGGGGGCATCCAACGTAG
294	AACGTAGGTCTACCAATTCGGTCTGCGGTCAATCTCTCGGAACGAAGCAA
295	AAGCAAAGACTTTACGGGTAGTCTACTCGAATCTACGATGGACCGTATGGG
296	TCATGGGCATGCTCAATTTATTTCCCTTAAAGGGGTAACCGCACTGTACCG
297	TGTACGCAGGAGATTGGGATCCACTATGCATTCTCGCGGGCCTATTTTAGT
298	TTTTAGTCTCAGGCAGCTCTGGCATAAGTAGTACCCGATGTGCCTAGTACA
299	AGTATCACGAGCAATACACTGTTATCTACTCGTTTCGGTTACAGGAGAGCTG
300	AGAGCTGTACAATCCTCTACGCAGTCTTCAATCTTATCGTCCCTAGACGGTG
301	GACGGTGACCCATTAGCTGAATGGCGAACGGGAGACACGGATTTCCGGGAGC
302	GGGAGCCCAATACTAGTGGAGTACTTGACCACTGACTTGTGAATCGGAGT
303	TCGGAGTGGGGGAGACGACATCGGAAGTAGGCTGTAGCGGGCAGTAAT
304	TTGAATTATTCAATTATCTTTCATCGATCTCGCGTGAAGTAAAGTAAAAATG
305	GAAAATGAGGGTACCAGTGGCTAGCCGTTACGCCGTTGGACTCCGCTCGTA
306	GCTCGTATGATGAATGTAAACATTCTGGGGTAAACCGTCCGTCACAACTG
307	ACAACCTGCACTAACCCACCTCGATATCGATATTGTAGCAAACGTGAGGCAT
308	CAGGCATAACTCCACCTATGACGCAAAACCTTTCCGGTACTCGATAACCA
309	ATAACCACCAATAATAGCCTCTAAATACTATGGCTTTGTGGGACAAACGGTG
310	AACGGTGAGATTAAGCGATAAAGTTCCCGAGAGTGGCTCGGGACATTGCGTC
311	TTGGCTCAATCCCCTAAATATGGCGCAACTACTCCTAGCGGACTCCAGC
312	CTCCAGCGTATTTGGGTTAGGAGAAGGATGAAGTATACCGACAAATCTACA
313	ATCTACACTAGTCTCCTTTTGTCTTGGCAACTACTGTGGGCTGCCCTGAAAG
314	CTGAAAGAGCCCTGCTCTACTCTTCTTAAATAAGCAGTAACTACTTGGCCG
315	TTGCCGCGGATAGTGATAATATCGCGGAATGCTTGTCAATTTTAACTCGACC
316	CTCGACCTCATTATCGCAAATTTGTAAGATATTTAACTATTAAGCAGCAACGG
317	GCAACGGAGCTTATCTACAACATCAAGACTGAATCGTTTAGTCGCGAAAAAG
318	GAAAAAGCTCTTGCAAAAAGTTGTAATGCAAGACCTGGCTGCAGCAAGATTGC
319	AGATTGCTTGGCGTGTTAGTGACGTCGCCGCGTTCGCAAAATTTGAGAA
320	TCGAGAATACTACTACTACTGCAAGTGGAGACTCAGGTCTATAGGACGTA
321	GGACGTAACCCACGCTGCACAACATAGCATCTTGGAAATTCACATATGTAT
322	TATGTATACATAATCGTAACCGTGTGGTCTTCATTTTGAACCGGGCGCAGC
323	GCGCAGCTCCACATGCGGTTACTCTGTCTCAGATACAGCGTAACTAGGC
324	CCTAGGCTTACCTGCCTATTGGTTGGGCTTACTCATACCGGCTTGTAGAATG
325	TAGAATGCCGAGCGCGCAGCCTATTAATTGTAGTGTCTTCAAGATAAAT
326	ATAAACTGCAGTACTGTCTTACTACTGAACCGTACATTGTACATCTGTTAAC
327	CGTTAACCATCGATTTGCGTGTGAGCCGTTAGCCTCACATTAGCGCCGGCT
328	CCGGGCTCCTATCCCGTGGCCCTGCGGTGAGCCGCTAGTAGCTATCAGTCA
329	TCAGTCACTCTATGACAGCTCGCATGAGGGTGGATTCATTTGGCACCTTGGC
330	CCTTGGTACTGTGTAGGTAAGGTCATTAATCAAGTCGTTAAAGCCGACCTA
331	GCACCTAAAAAGTCCGGCCGATGAGGACAGAACGCGGACCCACAACACTTAC
332	CACTTACTCTAAAGAGTACTTTGCTAGTATAGGGTTGACGCTTTCACAACC
333	CACAACCTCCCATCGCAGTTCTCCCTAAAATTTAGGCCACGAGAAGCAATGA
334	GCAATGATTCTTAAACTCTTAAATGGAGCTTGACGGAGGCATTGAGAAAATC
335	GAAAATCAATGGATCGATCGATTCTGACTATTCCGGGTTATGAGGAGGATAG
336	AGGATAGTAATTTCTGACTTCCCCTGTCAGCGAAAACAACGGACATGGTAA
337	ATGGTAAGCACATGACAGAAAAGTGGAGGCTTCAAGGAAAGTCCAGTCTCG
338	CTGCTCGGCTATGATTTTGTCTTGTGCTGCTGTAATCCAGCTTCTAGTCTAA
339	GTCCTAACATGTGCCGCTTGTGGAGCTCGAGGAAACAACAAAAGCATACT
340	GCATACTACAACCTTGATGCGTATTCTGGAAGTTAGTTGCCAGGCCGACCCG
341	CGACCGCCGTCAGCTCGGCCCTCCGCAACCCAGTCCGAGACCCAGCGTGAGG
342	CGTGAGGCCTTTATGTGATTTTCTAATGTCTTGCCTGATGTACCTG
343	TACCCTGAGTGTACTTAGGCGGAGGTCAGGACATTTCCCGCATAGCGCGC
344	AGCGCGCTATATAGGATATCAGCCGCTTACGCTCGCTGTCCGAGTTGAGCA
345	TTGAGCAGCGCATGGTAGTAAAAGAGATATTGATGATCAGATTTCTGTGAGA
346	GTGCAGACCAAAGGTAGGGGAATACATCCGAGCCTCCGAGTCATCGGTATT
347	CGGTATTTGATTATGGGATCGCTATGACCTGACTCTGTGATTTATATGA
348	ATTATGAGATTTGCCCTTGTAAACCGGAAAAATACATGGTGACTCTTGGTTG
349	TTGGTTGTCTAAGAGGGCGTATATGCTTGGCGCGGATGGCTCATGCTGT
350	ATGCTGTGGCCCTATCTTCGCGAGGACGTCTTATGTGGGAAGGAGCTTC

351	GAGCTTCTTTATAGAAGAACAATTCATCCATCATTAAGGCATGACCGTGGGC
352	CGTGGCGGAGGAAGCATAGGACTTATGCGATCATTCCCAGAACTTCGATGTA
353	CGATGTAAGCAAGAGAGTTTCTTTAGTACCGCCGAGGTCTTATCATCGGCCA
354	TCGGCCAAGCCGCGCTCTCAGGCGAAACATCAGCGCAGTCCCTGTAGTAAT
355	TAGTAATAGGGTAGTAGACGTCCCCTCGCTCTGAAGGTATTGCATGGAATCG
356	GGAATCGAAGAGGGTCCATGACTTTAATCCGCGTACCCATAGACTAGGATGG
357	AGGATGGACAGTACCGCCTCAAGCGAATTCAGTACGTCGAGGGAGAGGCC
358	AGAGGCCTGCTTACCACACACCGTCCATAGGCTTGCAACGCCTCTGCCTG
359	TTGCCTGTTCCAATCGTTGTCAGAAGTACAGAGCCTGAGCCTTACGCACCC
360	CGCACCCCGTGAGAACAAGGTCCTACACTGCCATCAGGTGATCTAGACCT
361	TAGACCTTACCCCTCGGTCTCGAGTTCGTATCAGCTGCAGTGGTTGAGCG
362	TTGAGCGAATACGTGTGCTCCGGTGCCTTTTCATCCCGGAGGTTTACTCGG
363	TACTCGGTAGGAGATAGCCCATGCACGGGTGAAAGGGCGCACTCGTAGGCAG
364	TAGGCAGCACTGTTGGAAGTAATCAATGCAGGCGCCCGTCTAAAAGGTTCCG
365	AGGTTCCGGCTAGATTCACCTCAGGTGTCTTAGCATATCAAAAAGAGGTGAT
366	AGGTGATTTGGGGTATGACTTGAGACTTCAAAGTTATCCCTTCCGTTGGGGG
367	TTGGGGGGCCACGCCAGCTTTAGTTCTCTACGGGACAATTCGTGGGTGTGTT
368	GTGTGTTTATTTCGGAATGTGTAAGCCAGACCTGGCCAGCTGGCCCTTCGTC
369	CTTCGTCAATTAACACACAGAGACTCAACCGTTATCTTTTCGATTTC
370	ATTTCCAGTGTACGGTCCCCTGCAAGGGGAAGTGTTCGCTCCACTG
371	TCCACTGTTCCCGGTAGCAGGCAGTTAGGATGACAGGCCGTCTATGAAACG
372	TGAAACGATAAATTTGGCGTAAGTAAGGAAGCGACAAAGTTGAGGATGCCTAG
373	TGCCCTAGAGCACGCGCCCAAGTCCCACGTTGACGGCTATAAATCTAGAATCT
374	AGAATCTTTTGAGGAATGGCTGGATCTTTAGCCAGGGTAGGCGCAGATAAC
375	AGATAACGGAACAACTTGGATTAAGTACGCGTACCTACCTGATTTCCGCA
376	TTTCGCAGAAGCGCTCTAGCTCGAATCAATGGTGGTATATTTCTGTTGCAT
377	GTTGCATTTGCCGCCACCACATATCGTTACCAGCGTCTCAGTACTATAGCT
378	TATAGCTTAGTCTGCCCTTCTATGGTCCCCTGGTCTGAGCAATTAATTATC
379	TATTATCCAGGGATGTGGACGTGACCTAACAAATCAGGCTGCCAAAGTGTCT
380	AAGTGTGTTACGTTTGGGGACGGTAACCCCTGATTCTCGGGGCTTCCCTA
381	TGCCCTATGGCGATAACGACCATAACTATCCATTTAGTGGCCAAACCTGGA
382	ACCTGGACCTTCTTCTAGTCAATGCGCTATCAATCACCCCTAACCTATCAA
383	TTATCAATTCAAAGGGGGGCTATTTGTGCTCATGGTGCAGCTTTGCGAATGTT
384	GAATGTTACGTGCCATTTCGATTACGCGGGGGGACGCTGTTAGATAAATGTG
385	TAATGTGCTAATAGGTTTGAAGCTTGAAGAATCAGACTCCGGGGGGCGC
386	GGGGCGCGGCTCTAGTTAGGCCGGCTAGGACTCGGTGGTGCACGATAGTC
387	GATAGTCGTGAAAAAGATGGGCGTGAGATAGTTGCACGCCGAACAGCGGAG
388	AGCGGAGAACTAGGTCCCGAATGTCGACTGCATGATTGACATATCTCGTTT
389	CTCGTTTATACGGGGCCGGGCTGCGTATGTGTGACTATAAGTACTTT
390	ATACTTAGTCTCTACTGTCAATAACTCGTTAAGTAACAGCAGGCGGGTTAC
391	GGGTTACTGAGCTGCTTTGGCGCTCTGTGCGGACTAGCACCGTTGATGGC
392	TGATGGCTATCATCAGGATTGCCATCTCGCCTCGTAATCTTTGGCTGAAGAG
393	TGAAGAGGATCCGCTGTAATGAAAGCGTGTGCTTAAGATCTTGGGAGC
394	TGGGAGCACGGAATCTTGTTACCAAAAGTACGACGCAAGTGATATCATATG
395	TCATATGACAAATTTGCAAAGAAGTAGATTACTAGGAAGTATGGCGTCCACC
396	GGTCACTCCGAAACACTTCGTATGGTGTACTGCAATCATTGCTCGGCCAGT
397	GGCCAGTTCCGAAGGTTCCCTCTAGATAACCTACATACGGCCCTACCTCAGCG
398	CTCAGCGCTAGCATAGCCCTTTTTGTAAACAAATGGGGATCAGAATGCGT
399	AATGCGTAATTTGATGGAAGGCTGTTGTCCCCATCAATTAAGCTATGGG
400	CTATGGGACCGGAGATGGAGAAAGCGATGCTATTGGATCTGGATTGTTCC
401	TTGTTCCGACTTCTGCCTTATCCTGTTTGTAGTTGAGATGCAACGAGACCTTT
402	GACCTTATCCCATCATAGAACCACAGAGCCGATCTATACAGGAAGTTGCC
403	AGTTGCCGGCCAAAACACTCTAGTGTGTTAGAGTTCACACATACACCAATGG
404	CCAATGGGCTGGGCGGGAACGGCCGCTGGCCGCTTATACCGCCAAGATAT
405	AAGATATCCCTTACACGGTAATCATATCGGGCGAATATTGTGTTGTGCCA
406	GTTGCCATTAATGCTGTCTCATACTGAGGATAATCTGTGTGTTGATTAGT
407	GATTAGTACAACAGATCGCGCTGCAATAGTAAGCATAAATGGACCTGCTTGA
408	TGCTTGAGATATATCGCGCGATACTTTGTTGCGTTGAAGCGAAAGTGAAGTC
409	TGAAGTCAACTACAGGCCAGTGAACATACAGGTGATGCGCCCTCATCCAA
410	CATCCAAGAGCATGTAGTCGTTGGTTAGGTTTCTCGACTGATGGTACTTCCC
411	ACTTCCCTGATCTGGCTATGTATGCCAAGGCGAGCAAGAACTTTGGCGTCCG
412	GCGTCCGACACCATGACATTTGACGGTGGAGCGACAGTAAAGGCACACCCCC
413	CACCCCGCTCTATAACCGCCAGGCTCGCAAAATAAACCTTCCAGTCTATT
414	GTCTATTTCCGCAAGAAGGCAGTAAAAACATCTTATGTGGCTATTTTCTGA
415	TTTCTGAACGTTTGTAGTGAACCGGCGTGTCTAGAGGTTACCTGGGATGA
416	GGGATGACTTATAACCCAGACGACTAATACCCCGAGCCGATCAGTTGTAGGG
417	TGTAGGGTATTTATACAATAGATGGGGATGAGCCTCATGTTATGCTATCACC
418	TATCACCGTGTCTCAATTAGGAACCTGCGCCTCAGGAGGGAGTCTTGCTA
419	CTTGCTAGAAACAGGGCGGATAGGCGGCGCATGTGCAGATATTACTGTCGA
420	CTGTGCAACCTTATCAGCTCCTCGGTTCCAGTTGACGCTGATTATACCCA
421	ATACCAGCCTAGTTGGCAACGTGACGCGTAGACACTTATGTAACGTGGTAA



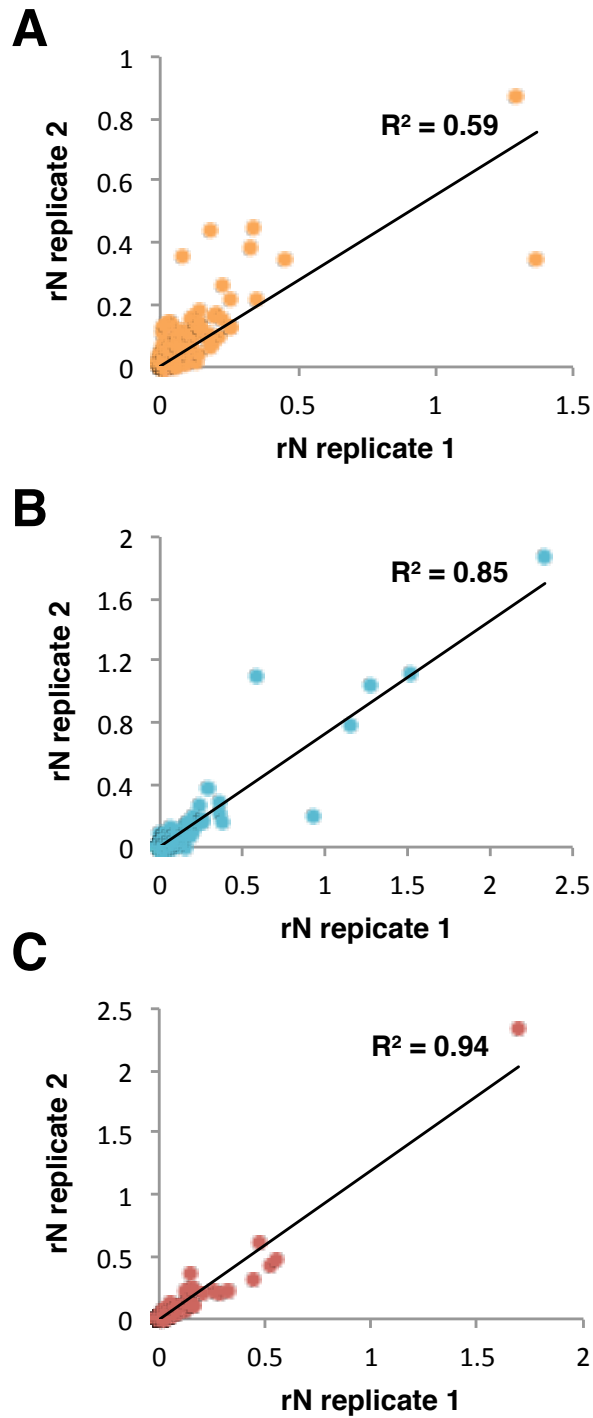
422	GTGGTAATAGAGGAACCTCGGCAAAAGCCGTAGAAGTAACGAGCATAAATGGC
423	AAATGGCGCGATTTGGATAGCGGTCAGGGAGGAAGATGCCTTACCGGCTAA
424	GGGCTAATAACCCGGATGCCAACATGCATTAATTAGGCTCCATGTCTGATAG
425	CTGATAGTTTGTGCATGTCCGATAGCTTGGCAATTACCGCATATTGGCCCT
426	TGGCCCTCCGATCCGTACCTCGTGACGACTGTTAATTCGTTATAGGGCCATT
427	GGCCATTTCAAGTGTGGCGGTGGACCCCGGACGCGCCGGCCCATCCGTGAG
428	CCGTGAGGGTGTGAAGTGCTTTCCTGGCAGCGAATGACCATTAACATACATAT
429	TACATATCCAGAGTTGGATCGGGTTAATTGCGGCATACGATGTTATCTTAAA
430	TCTTAAACGAATCAGCAACAGGACTTTTCTTAACCAATGTGGATGATCCAT
431	GATCCATAGTTTTCGAAGCGATTGATGTCGATGCCGCTGAATGTGTAATC
432	GTAATCTGTCCCGAAATAGAAGCTGCCTAGCTAATATATTATATAAACTA
433	TAAACTAACGTCACCAGTCCGTTTTTGGGAGAAGCCAGCGTTGATACCACAA
434	ACCACAAGTACAGTCTTTTTACGCGAGTTGGTCCCGCCAGTCTCAGGTAAT
435	AGGTAATACTAATAGTGAGCTCGCTAATCAGGCTCTCCCGTTCTACTGACT
436	ACTGACTGAGTTTTGAAAACACCTTGAATCCTGCATCACAGCACAGGTGGTG
437	GGTGGTGTAAAGTACCCCGGCGTGAACGAATAAACTCGGTGAGTCCGTGATC
438	CCTGATCCGTTAACTTTCAACTGTAAACGACTGCCTCGAGGCCTCGAAGCCG
439	GAAGCCGGGTGTCGCGAGCCCATATATTGACTATCACAAGGATTAGTT
440	ATTAGTTAACCGGGTATTCATAGTACATGGATCCGTCTCCCAAGACATCAT
441	ACATCATTTATTGAGTTTCGAGGGCATCGTACAGAGTATCTTTATCTATATG
442	CTATATGACCTATACCGTTCTCTGCCGGGGAGGAGCGTGAACATAGGCGC
443	TAGGCGCCTATAGGCTCGTTCCCGCGTCTGCTGGTCTGTACAGGCTGAA
444	GGCTGAAAGGCGGTTAGGAAGAGACTTAACTTGTCTGTGTTCAAGGTGAA
445	AGTGAATCTTTGTACGGGGGCTGCACTTGATGGTCTTCGCAAGGTAAG
446	GGAATAGCTTCAGGAGCGGTAGCGAAGAGAGGACTTGTACAATTGTTATGG
447	GTTATGGAACCCGGTCTGCTAGGGGCTTCGCATCGCCAAATAATGCGCGTAA
448	GCGTAACTAGTTAACTGAAGCGTCACTCGAGTACTCCAGATAAAACGCTGTG
449	CGCTGTGCGGGAATTAACCTTGTGCAAATATCCCGAACCCACTTCTCTCCT
450	CTCTCCTTAGAGCGTATGCGGGGCCATAAACCATTCGAATATGGTCTTTCAT
451	CTTTCATGACGAAACAATCGTCGCTCGTTGCTTGGTTCGTTAAACCGACGAAC
452	GACGAACTGGTTGTTAACATGTTGACTACACTCGTTTTCTTTGGGTCAGTA
453	GTCAGTAAGAGACCGGTCGACCGGCGGCCACAGAGAGCTATCGTATAGT
454	GTATAGTATATCGTGGTAGATGAGGCCCTTAAATTTCCCGTCAACATCGAT
455	CATCGATGCATCTAGATCTATGTCCTAAGGCTTAAAGCTTATTGGGTGGACT
456	TGGACTTAGCCGGGGTGCAGCTAGCAACGATCTACCGGATAGAATACAAGA
457	TACAAGAGACGAATCTGGACCGCCCTTAAAGGTTATCAGCAGTCATTGGCC
458	ATTTGCCTCATTCAAAGGGTAAGGAGTGGAAACGGTAGAGCGGATAACCGT
459	TAACCGTCCCTACACCCCATTGCCAGACGTCGTCGATCCGACGCTCGCAT
460	GTGCGATTCCACCACCCGAATCGGGTGCAGCTGCGTTGTTCTGCTTACGAAG
461	TACGAAGCCTTGGGTATGGTCGTAAGGAACTACCACATCAGATATAGAGTA
462	TAGAGTATAAGGCTAAAACCCAGTGCAAAGTTTAGGAATCGTGGGAGTCGA
463	GAGTCGAGCATCAGCCAGCACCCAAAGTCTGTGGACAGACTCATAAGTACGC
464	AGTACGCCATGCAGCCTTGCCATAGCGGGCTATACCGCAGCCGTGACTAGTT
465	ACTAGTTTACTTAGAGGTAGCACCCGGTACTCCTCGAACCAATTACGGGGT
466	ACGGGGTTTCAATGTGAAATCCCTCCATTCCCTACATCGGCTACACACAAT
467	ACACAATGAACGACGATGGGTTCTCTTTGCGCAATATTGCGTAGGGGCAAG
468	GGCAAAGTGCATCCTAATGCGTTAATCTGCCTAGGAGCTGGGCTAGGTTTC
469	TAGGTTCTAGGTAACAGGTTCCGCTCACCGGACCGGAGACACAGCTGTACA
470	CTGTACAGACCCGAGCATTTCGGCTCCGATGAAACTGAACCCTACGTATGC
471	CGTATGCTCTCTTAAATGATCTTTCTCATACAGTCGCGTTGACCCGTATTT
472	CGTATTTACAAATGTGCCAGGGGGAAACGCTATCAACATTTTCATAACAAGC
473	TACAAGCGACGCTAGATAGCGACATTTGGACCAGTGACCTGACCCAGTCCG
474	CTATCGTCAACACCTATTGACGTAAGGCGTTTACGTTGTTACCGGACTTGA
475	GACTTGAATACAGAGGTTTCGGCTCCCATGCCATAACGTTGATCTGCGGAGA
476	GCGGAGAGTAAAGACATAACACCATTTCGCCGAAGCCAAGAAGCAGTTGATCA
477	TTGATCAATTTGACAAGTCGAAACGTTTAAACTGTCATTGACCCAGTATGGAG
478	TATGGAGAGAGTCCGATTTACTGAATACGGCGGCACGCTATCTGTCTAAC
479	GTCTAACAGGCTATGGCACAGGAGCGACTCGCAATACTCCGACCCAGCTT
480	CCAGCTTACGCTACGCTCCAGTCAAGCAGGCTGGTGTGTACCTTTCGT
481	CTTTCGTGTCGATCCCGTCATAAACGGGCACGAAGGTAATGTAGGAAGGTAG
482	AAGGTAGCTCATTGATAGTAGGACCGGTTAGCAGGGTGGTAGGTGGCAGG
483	TGGCAGGAGTTAGGTGTGTCCAGTCGCCAATCTACCCGAGATGCCGGCGTTG
484	GGCGTTGAGAGGCACCTTCGACGCGGTCTTAAATAGCTAAGGTTTTATATCG
485	TATATCGACGGTTAGGGCAGGCTCTGCTTAGGAAAACCTCAAAGCCGGCTA
486	CCGGCTAGCCCCGCTGACGCCATTGCGATTGGGGGCCGAGGCCATTCTCTTC
487	TCTCTTCTAAGCACCCAGAGACTGTTGCGATCGTACCCGCTAATAGCTCTT
488	ACGTCTTGAGTGGGTGCACCCGTGCACTGGGGTCTGATCGCGGTGCAGTCAG
489	CAGTCAGGCGCAAGTCCAAGGAGTTAAAGGTAAGTGAGGAGAACCGGAGG
490	GCGAGGGTTGCATACTGCCAGTTAAAGATGCGCGAAACGAACCTAAGGTGAGT
491	GGTGAGTCGACGATACCTGTACTAGATGTCCCGGAAACCGGATTAGAAGCTT
492	GAACGTTGCTTTCAGTCATGCTTCTGGTTAAACTCACGATTAGGACCATTG

493	CCATTGTTTAGCAGCTATCTGTTACATTTAAATTTACCGGTACCGTACTAG
494	GTAAGGATCTAGTCTGTTATGAAATGAATGCTATTTTCAGCGGCTACACA
495	GTCTACATAAACACCACACTTGACGCGGCCCGCCGATCCTTGCCCAATGACG
496	AATGACGCGCAGTAAGGCAAACCCGACTCCGAACTCATTGCGCGGTGTGTAA
497	TGTGTAAGTACTACCGGGCTTACGTTGAGTAATGGTGTAAAGGACAAGCC
498	ACAAGCCAATATACACCTCTGCCTGGACGGGTCCCAGCAGAGAACAGATGAT
499	AGATGATATACGAGATGATTAATTCATGATTATCGTGTCTGTGTCGCTGGTT
500	GCTGGTTGCTTAGTATAACGGGTTACCTGGCTCCGCTAAAAACAAGGAAG
501	AAGGAAGTCTTGTAGTCTGGCGAAGCAGGGACCAGGAGATCTCGAGAGCCCA
502	GAGCCACGCGCGGTGATCTATCAGGAATATAATCGCGGCCAACACATTGC
503	ACATTGCCGCTCCTAGGTTTAAACAAAAGCGGAAACACCCGATAGTCTAGA
504	TGCTAGATCTGCTAGGCTAACGGTACACTCCCTTACCCTCGGAGGATTAT
505	GGATTATAGGCGAATTGAAAAAGCGGAATCGGTTATTGTCTCCAGCGGATT
506	GGCGATTACTCTGGCTTATGGGCGACTATACCTACTGGCGTCTGTTGGAT
507	TTTGGATGGGTCGAGCCTTATGTTTACCCTAGAAAGACAGGCTCACGCA
508	TCAGCCTACTTATCCACTTACGATCAAATCTGGTCTCTTTCCTCACCCG
509	TCACCCGCTTGATATATTCTAGTAAGTTACCCCAATAGTTGGTGTGTCTT
510	TTGTCTTCCCGGATCCCGCACCTTATTATTCAAGGCACCCGACCGCAGG
511	CGGCAGGCGTAGATCATGTATAATAGACGACGCCCCCGGGCCCTGAGGGA
512	TGAGGGAAGCTTAGCGTCTGATGAGCTTGGGCTAAATCCAGGTGGGGGGAG
513	GGGGGAGCTCATGATCATGATAGTGGCAGTCACACTGCTAGACGTACATAGG
514	ACATAGGGGTTGCCAGGTCGTTGTTGGCGGACCGTTGGCAGTTGCCACAG
515	CCTACAGTCAGCCTCAGATCTGCAGAATAATGTACGGAGGACTCTAATGACA
516	AATGACACAAACGATGGCTGCCAATCAATACCACGGCTCGCTACTATTTT
517	CTATTTTATGGGGCCCTCAAGGGCTCCAGCATTCTCAGCACTACCGCGTAT
518	CGCGTATCCCGGGACACGCATAGGTTACATAGTAAAGTGTAACTCTAGC
519	CTCTAGCGAGTCTTAGTTTAAATTAAGGGAAAGTAAAGCCCGAGACAACTG
520	ACAACCTGGTGTCTGCTCAGTTCGAGCGCTCGCGAGCACATCCTTAGGCTGAT
521	GGCTGATTGCTACCCGGCGCGCGGATACGTATACGCGATCCTATTACT
522	ATTTACTTCGGTACACAGTGGGGTACAGTTAGAGCATCGCAAGAGTCTTTG
523	TTCTTTGAACTCTGTGATACGTGCTAAAATCTTCTTCTGTAAGTGCCATCGG
524	CCATCGGTGCCGGGGGCAACACAGGTAATGGTTGGTACAACCTCCGCGACC
525	CGCGACCACGCACCTGGTCTGTCCCTTGGCGGGTACAGGGTTGGCTGCAA
526	GCTGCAAGATAGCAGAAGGTGTCAAGAGTGTTCAGATTGTGCACCATAGGTC
527	ATAGGTCGTCGAGACTTGAAGCTTGAAGTCAACCAAATCTTACCTTGGGGT
528	TTGGGGTTTTAGAGCCACGGGGGAAACCTTAGCTCACAACGTCGTTGATT
529	GTTGATTGCGGATTTGACAGCGCGCCCTCGTCTGGGCGCTTGTATCCGATG
530	TCCGATGCTTGGTTCAGGTTAATTTGGGATATGCACCTGTCGTTCTTGACAA
531	TGGACAAAAGTCTCCACTCGACAATGATAATCCCGTTGGTGTCTGAGCTT
532	TGAGCTTTACATACCCTTGAAGTGGTGGCTCAGACAAACCGTAGGGAGGCA
533	GGAGGCAGGTCAGTACGTACGAGAGAAGAAAAGTTTGGCCTGAGGAACATAT
534	AACATATCATGGATTGGCTGTGCTCGGGCACAAAAGAGCAAGTCCGGCCAC
535	GGGCCACTTTGATCATTGATGCTGCGATCTCCCGCATGTCACAAACCTCCAC
536	CCTCCACCAGGTTTCTGAGCGTTGACAGGTTCTTACAAGCTTTCAAATTT
537	TCAAATTTAACAATAGTTCCAGGCGCGACTCTCGCGCACCCGCTGTTGAA
538	ATGTGAATGGGACTGATTGAGCTAGTACCAATCCTATCATTGTGTGCTGA
539	GTGCTGACAAGACGGCACTACGCGCGGATGTGTTACGAGCGGTTGCCAGATT
540	CCAGATTACCATTAGTAGGTGCCATTGTTAATCCTGTCTGGTATTAAAC
541	ATTAACCTGAAAAGAGTGCAGGTCGAGATGCTTAAACGCCAAGGGGCTGGA
542	GGCCTGACCAGGGCTGATCTAATATCAATGTTGTTATTTACGCAAACCATAA
543	ACCATAATTCGGCTTGAGCCAACTTGCATCAACAGGTGTAACGAAACCGAC
544	AAACCAGCGCTCGGCTCCGAGACGGCTGGGAGGCTGCCTCCAACCGTACC
545	CCGTACCAGGCCCTGACAGTGGTGTCTCTGCGCTTACATGCCCAAGCA
546	CCAAGCAGTCGAGGCACTAGATACCGCAGCCCTCGAAGCGCTTATAAGAT
547	ATAAGATCGCTTGGTCCATACCTAGCTTCTGAGAACGAGAACTCAGGAACAA
548	GGAACAACCGTCCGTATCTGTAACGGGGTCCGCGGATATCCGAGGTGAGGG
549	GTGAGGGATATAGTTCAGGAGAAATAAGGATACCCCGCAGAAATGCGTT
550	TGTGGTTTCGAACTCGGGTTCTGACTGCGGAACCTACTCCGACAACGGTTT
551	ACGGTTCATTAGGCAATAACGAGAGTTGAAGATGTTAGTCCCTGGAGCATA
552	GAGCATATTCAGATAAGAGATGTGTGATAGGTTAACCGGTTCCGACCAGC
553	GACCAGCTTGTGCCAGGACGGTCTCAGCCACAGTGGCGATTCTAAAAAGTC
554	AAAAGTCAGATTACGCCGGATCTGACGGGTGCATGACAACCCCTAATTC
555	TAATTCGGGAATTCCTTCCGACAGAGCCGTTAAATCGTTCTGTCGCAAC
556	TGCAACTACCGGGGAAGGGTACCACGCGCACTTGTGTGGCACCCGTAACAT
557	TGAACATCGGCAACTTACGGCCAGTGGCCACCCCTCGATCACAGTGACA
558	AGTGACAACAATGCCCGGATAAATGCACACTAATTGCACAGTTTGTCTCGA
559	TGCTCGAAAACTGGCGGGAAGGCACGTGCACGAGGCCCGTCCGATCTGATT
560	TCTGATTAGGTTAGTCCGGTGCATAACCGGAGTAAGGGCCGTGCCAGTACAC
561	GCTACACCAGGTTTGAATAAGCGTGGCTAAGGGGGCTTTTTAAATAATCGGA
562	AATCGGATTACTGCATTATGCGCGCATATATGTACATGTAGATCCATCCGAA
563	ATCCGAATCAATCCTTAGACTGTCGCAACACAAGATGTATATCTGCGAT

564	CTGCGATATGCCGTTGAAACTCCTAAACCTAGACAGTCTCGTGGAGGCCAG
565	AGGCCAGAGATCGGGCTTACTACGTAAGAGGTACCCGACCACGTCCTGAA
566	CCCTGAAGATCAGTCTGCTAATCGGCATAATCTTGGTGTGGGAGCTATACCA
567	TATACCAGAGAGAGGTGCGATTGAAAATCGTCTCGGCCACAAGCCCTCTCA
568	CCTCTCATCGACGAAGATTCAAATGCTGTTTATAACTTAGACCATGCTCGC
569	TGCTCGCACACAGTTGTGATTGATAAGTCTAATACTTCGCGTCAGGTTGGC
570	GGTTGGCCCAATGCTCTGCAGTCTCGGTGCTCCTACAATCAAGGGGTTCCC
571	GGTTCCCAACTTCGGGGCAATCACACCGATATCTAAATTAGTCTTGGTCTGT
572	TGGTCTGCTGAAACGTGTTATCATGCTCCCTCTGCAACACCGACACTAAGT
573	ACTAAGTCGGTCCAAAGAGTTGCAGCTTTCCCTAGCCTTTGTGAGGCTTTT
574	GGCTTTTCGGCACGTTACCCGCGTAGGCCTTGAGCTGTTATTGGAGGAACC
575	AGGAACCGTCTTTGTTATCGGGAGCCTCGATTTTACACGTCGCTCATCTG
576	TCATCTGACCCGACAGTTCAATGAAGCCATCTAACGCTTCCCAGTAAGCTA
577	TAAGCTAAATAAGTCCGATCCAGCAATTGGATGTGTACGAGTCGGACCTGAA
578	ACCTGAAGTAGCCTGAAGGGGATTTTACCAGAACCAGTGTTCATAAAGC
579	ATAAAGCCGCTCGCGCCGGTCCAAAGTTTCTAAGTGGTTGAAAAGGCCAC
580	AGGCCACTGCCTAAGTTCGCTATTACATATTACGACTGGTCTATCGTGTTCG
581	GTGTTCCGGTCCGGCGAGTGTCCCTCCGGTAATCCTACGACCCGATGGCCG
582	ATTGCCGAACGGCACACATCACGAAGCGGTTTCATGTGGGCATACCCGACT
583	CCGACCTTCAATCGGCCTTTCGACGAGTACAGCCATGCGACGCAGAACTAT
584	GAACTATTCTCCGAAAATTTGCCACGATGTCTCATGCCAGGTATTAATA
585	ATTAATATTACCGGCAAAGGGTGTGGTTCATCGCTGGAAAATCGCTCACCTG
586	TCACCTGCCCGGACTGTTCTAAAGCCAACAGGGCATTCCTTCTCGGCGAGC
587	GGCGAGCTTCCCAATGCACCTGATACTGGATTCTGGTGAACGGGGTTCGCT
588	GGTCCGTGAACAGGGTGAATTAACAGCGCTCCTCTTGATAGGCATATAG
589	CATATAGTTCAGGAAGAAAGCAGTGAAGAATTGAGAGCCATCCATGCTTGA
590	TGCTTGATCTTGCCAGTGAGGCACCTACAAGGCCAAAGGGACAGCCCAAAGG
591	CCAAAGGAGGCAACAGCGACTACTGGGAAACATACGACCTCCTCGCATCAGA
592	CATCAGAGGCGCGCGGTTGTGCGTATCGGTTGCTACGATAACTGGGGCGTC
593	GGGCGTCTTGCGCAGCAGACCCGTCGCTGAGGCCGATACGCATGAGTAG
594	GGAGTAGCAAGCCATGTTCCACCCGCGCTTACATCAAAACACTCCAAACC
595	CCAAACCGCCTAAGCTCGACGGACCTCACTAGTTGCTCCGCGCATGACGCTC
596	GACGCTCCGCATCGGAAACTAAACCCGATGGGGAGCGCTGCGACACCGGTC
597	ACCGGTCCACGTAATGGGCGCTTGTGCTGGAATACTACCCTTGCTTCAGCT
598	TTCAGTGTCTGTACGCAGTGCAGTATGAGCCATGGGCCCAACTGTAA
599	ACTGTAACACCCGCAAGGCAACCGCTGGTGGTTCATACATCGTGACCGTACA
600	CCGTACAGCGCCCATGGCGACCTCGTATTACAATGAGATAAATCATTTCTCT
601	ATTTCTACGGACAGCTATATTGTTTATTGCACTTTTAAAATTTTCTCTGT
602	TTCTGTTCTGGAGAGCCGTACCCCGCTCCTCCCAATGTCTAAACAGGCGT
603	CAGGCGTCTACAATGCATTAATGTAATCTCCCTAGTCCCGTGCAGTGTGC
604	AGTGTGAGGCGCGCTCAGCCGAGACCCGGAGCTTGAACGCTATCGCGA
605	ATCGCGACCGACAGGGTAATTATGTGTCGGCCGATCGGGCCTATACGCAAC
606	ACGCAACCATCCCTCCAAGTACGTGATGGGCTTCTGCGAGCCTGTGGTG
607	TGTGGTGATATAACATAATGGCTAGGCAAGGTTATACCAAGGGTTGTTAAA
608	GTTTAAAGAATGCAGCGGGATACCTTGTACGAAACGGCAACCACCCACCA
609	CCCACAGTGTACAAATTAGACCCTAAAGTGGTGAGAAGCTTTATTGTTACT
610	TGTTACTAAGGGCATAGGGAGTTAATCAGGAGTGCTTGCACCTCTGTTTCG
611	TGTTTCGATCCGCTCGGGCCACTCTTATTACCAGATTGCACCGTCTGAGC
612	CGTGAGCGTAGGTTTGTGGCAAAGATGTGCATACAACGAATGGACCTCTCA
613	CCTCTCACGTTTGAAGAGCTCAGCCTTCTTGGCGACAAGGGGGTAAAGAAAT
614	AAGAAATGCAGGGCGCTACAAATATTAGGCGATGAGTGGGGTTCTGCCACA
615	TGCCACATGGAGCGAAGCTTGCCTTGACTAGGCGCACCAATTTCCGAACATT
616	GAACATTACCCAGAGGATGGCCTGGGGCACTGCAAGCGAGAGGGGATCGGT
617	GATCGGTAGTTATCAAAGCGCTGTTTACTAATTCGCACCATCGTGTGATCAC
618	TGATCACGTGATAAAAATGCATCAGTATCCCTATTGTGGTTGGATTTGCAATA
619	TGCAATATAAGCCCGTTGCACTACAGATCTCTATGCGAGTTAGAAAGCTGAA
620	AGCTGAAATCACAGGTCCTCCAGAGAATTCACGCCGTAACCCGGGGTG
621	CGGGGTGAGCAAGGCGGTAGCGATGGCCGTCACAACAACCTACCGTGGCCAT
622	TGGCCATAGATTCTTTGTAGAGCTGGTATTGGACACTGGGTGTCGAAGTCC
623	GAAGTCTTTCCCACTAAGATTTCTTTGGGGTAGGTCCACTTCATTCCAT
624	ATTCCATAGCAGCGAGATACCATTGGGTTTATCATATGTGTTCTCTACCT
625	TCTACCTTGAGAGTACTAACCCCAAGACACGCGGGTTGAGACCTAATGTATA
626	ATGTATAGTGGTCTAGTGTGTTGATGGGCTCGGTTTGTGCTGCATAGAT
627	CATAGATATGGTGGCTGAGTATCCAACACCGGTTGGTATGTGATCTACGT
628	TCTACGTTACCTCGCCAAGTCGCGGACAGTCAAGATGACCGATTTTGCAG
629	TTTGCAGGGGAGGTCTGACGACGGGAGCCACTCCAGGTACGTTTATTACAGA
630	TTACAGAGCTAATCTGAGTGGATGCACCAAAGACCGTGATGCCTCTGGTCCG
631	TGGTCCGGGAGTGTGGCCTACACGTAAGAGTCTAAGGCACTGATTACCG
632	ATTCACGCTACTGTGCTCTACCCAACGCCATCCGCGAAATAGGGGACA
633	GGGGACAGTATTGAGCCGTGTCCACGATGAGCAACCCGAACTGTGCGTGCC
634	GCGTGCTACCGCAAGTTACTATATATTTTGAAGCCACGATCCCTCGTGA

635	TCGTGTAACATATTTTTGGCTATTCGTAACTTCAGCATAACC
636	CATACCGTATTTGTAACATAAGCTAAAAGACCTAGCGTAGCGTAATAATG
637	AATAATGGATGCTACCAAGTAACTGGATCCATGGAGACGGAGCTACAAGT
638	TACAAGTTCGACCTTGGTCACCAATCACGTTCTGGCGTTTCGGTTCAGAGAC
639	CAGAGACGCAGTATACCGGAGATTGCGAAACCGCTCCAGTAACATCATAAG
640	TCATAAGGAACGACTCTTAGGTCGACAGCGGGTCCAATCCGAAATAACGCAG
641	AACGCAGGACGTTCAATCCGAGAAATCCATCTACCAAATTCAGAATATGT
642	AATATGTGACTCGACGCGCTATTTTCATCTATCCCTGTCTATAATTTGCATG
643	TTGCATGAGTTTGGAAAACATTAACGAACGGAGTTACTCCCCACGGCAAGG
644	CGCAAGGGCGAAAGCATGTGAGCACCTTTGCGGACATATGGCGTGTGTAC
645	GCTGTACCTATTCATCGTAATGTGTGGACTGAGCCGGGAGTCAGCAAAGGCC
646	AAAGGCCAATTTGGCCAATTTGGTTCGCGTGGCGCGGAACATTTGCTTATAG
647	CTTATAGTCCGTGGACATGCCACGGAGTACAAACGCCTAGGATTCACAAA
648	CTCACAATTAGCTTCGCGGTTTTCGGTAGAACGAACCGACTCAGAAAACGCGT
649	AACGCGTACAAAACATGATTTGAACATGAACCACTTAGGAGACAGAGGCTG
650	GAGGCTGGAGGGTTAGAGATTCATTGGTATCGAGAACATTGGATAACGCTCTG
651	ACGCTGCCTCAATTAACAACTCGCTTAGATAGAACCAACTCAGCAATAAGA
652	AATAAGATGACTAGCTTTTGTATCCATTGGCGGGAGGCTAGACCCATCA
653	CCCATCAGAAAATTTGAAAAGCTACGCGCGCGAGAGATCGGTTTGACACAGC
654	CACCAGCTGACTAAATAGTCTGAGTTAGTGGGCTACGTCACGGTTTTTTCCC
655	TTTTCCCATGGTGGGACGGACTCGTGTACTACATTCTAAGGTAACCAAGTA
656	CCAAGTATCATTCTAGGCTCAAAGGTTGGTTCGCACATGCGCACTGACGGTAT
657	ACGGTATATCTAGGTGTACGCTTGTCTATCCACACTATTCAGCAAGTAACCA
658	GTAACCATGTGTCTGACTGGTAATGACCTCGTGGCATCGGTCGGGAGTGGT
659	GAGTGGTGGCAACCTTCTAGGGCATGATCCCGGAGACTGAGAGGATACGAGT
660	TACGAGTTCAGTCGTGATTACGGTAAGAGTATGTACTTTCTATCATAATTT
661	ATAATTTTTCGCTGTGGTACATAACCTCACAGATCAGAAAATGTGATAAGC
662	GATAAGCATTAAACGTGCATTCGTACATCCAGAAAGGATATTAACGTCGGCT
663	TGCGGCTTCAGAAGGACCTCCAAGACGACCCCTTCGAGCCGTTCACTCAGCC
664	CTCAGCCAGGACCCTAGCATTATACGACGACTCGGACTCTTCACCGAGGC
665	CCGAGGCTGTTTGATACTCCAACGCATACTTATACGTGAGCATCCCCTCTC
666	CACTCTACAGGCGCTCAGGACCCTCGGAAACCTCAAACATGCTAAGCGT
667	TAAGCGTTGCTCTATTGATACATGTTCTGCTCGATGTCCATTACAGGTCGG
668	AGGTCGGCTGGACGCGTCGCGACGAGCTTGTGTACTGGTACTCTGCTTTG
669	TGCTTTGTCTAGAAGTTGTTTGGACAGCAGCCAGGATAGGTGCAACATTTCC
670	CATTCGCGTCCCGATCATCGAGTTTTATGCGTCTCAATCTGTATGGAAGCT
671	GGAAGCTGCGAGCGCCGGAATCCTAGAAATGATCACGGACGCTGTGACC
672	GCTGACCCTACTGAGTGCATTAGCCGACATTTACGAGTATGCCTTGTGGC
673	TTGTGGCATATTTCCCTGTGTCCCACTTGCCCCATTATAGTAGAATCAAGC
674	ATCAAGCCTCAACGATAGTCTTTATTTAAAAAATACTTACAAGATACGATCT
675	ACGATCTTGTGTCGAGGCGGACGAGATGTCGGGATCCTGAGAAGCGCCT
676	AGGCGCTGAATTGGAAGATTAGATTGATCTAGGGTACGTCGGATACACCGT
677	ACACCGTAACGCAAAGCATTGTCTCCACTTGGGGGAAGATAATAGGACAG
678	AGGACAGGCGACCGCTCTGCTACGTAGCTACCGTCGCGGGCCCCGACGTC
679	GACGTCATCAGCATCGTCTGCGAGGCATGCACCTTCTGTTTGTGGAAGGATT
680	AAGGATTAGCCATAAGGGCACCTGCTGCTCCGTAATTGCAACAATTTCCG
681	AATTTGCTATTTGGTCCCGTATATGAGGTGACAGACGATGACATAAGCGGTA
682	AGGCGTATTGAATATCGTCTAGTATGAGCCAGACTCGTTGACACGGGCCAG
683	GGGCCAGGAGTCACTTCTCGCTATAAGTTCTAGCCGGGATGCCAATCCTT
684	AATCCTTGGCCTGTTATAATGTTGGGAGTGAGGGCGGTAAGTCTTCTGAAC
685	CTTGAACCATGGCTCCTTATAAAGGACAAACGGGTGGCGTAGTTTCAGAAC
686	TCAGAACGGTATCGCCGGGATTGTCTGCGTACGATATCCAAGGACGGCCAT
687	GCGCCATCTGTGCAAGTGAATTTAGCATACACTCTACTTTTTGCTCCCGCCTA
688	CCGCCTACAAAAGGTGGCCGGAACCGCTCTGGAAGGCAATCTATTCATTT
689	TCCATTTTTGACTCGCGATCGCGATGCCTGCCATGTCAGCTATGCAATTT
690	CAATTTACAGAGTCCACGGCATGGTGCTAAAGCACGGGCAAACGCAAGATGA
691	AAGATGAGACGTGGGACCTATTTGGTCAAATAGTTTCGCTAACGTGTATAC
692	TGTATACGTGCTACGACAGGACGGCAGGGCGGACAACTTATGGTACCTTG
693	TACCTTGCTATAGAACGGCCGTACGTGACGGGAATCCATTAATCCAGTT
694	CCCAGTTCTTCCAATTAAGCCCTAAACGGTTTCTATGCCTCAGCAGGTTTT
695	AGGTTTTGAGCTCATGAGGTAACAATTAAGTCCATCGACCGTATGAATT
696	ATGAATTCCTTAGTCCATGCCCGGGGACTCAGCTTCACTATAAACCGTCA
697	CCGTCCTGCGCCAAACGCTAACTATGCTCAGGGATTGCATTCATAGTTGC
698	AGGTTGACGTTGCGTCTCCTAATAGCTCGACTATGATATTTTACTCAAAG
699	CTCAAAGTACTGGGCTTGGGATGGCATCTAAAGTTACCGACTGTGAAAATCC
700	AAAATCCACAACGGGTGGGAAACTTCCAGCGCGGGTTGGGAAATCCTGG
701	ATCCTGGAATCTGTTTGGAGGGCTCTATGGAGCAAGCGTACTCAGTGAGA
702	AGTGAGAGTACTTAGCAAATAAATCACGGCGTTTTTCCGGTCTGTAACG
703	TACAACGGCTGAACGGACGATCGACTTCCACGTTTCTCGTACATGTAAAT
704	GTAATAATAGCAAGTTGTGGTGGTGAAGCACTAAAAGCTTACAGGGTCTAC
705	GGTCTACGGAGAAGATCGGCACTCTGAGGAAGGGACCTCGACTACGTTCCG

706	CGTTCGCGGGGCACACGACCCTTTAGGATTTGTTCCAGCCCCGTGCCGGTGA
707	CCGGTGACGAATATGACGTCTTGTACCTGACTAGACCACCTCAATATTAAG
708	TATTAAGGACTACAGTTCTAACGACGGGGCAGCGCGAAGGGACCTGTTCTG
709	TGTTCTGCAATCGATTAATCGTTATGACTAGCTGGGGGAGTCCCATCCCGAG
710	TCCCGAGCTCTCGTACCTTAAACATGGTCCAGCGGGCATTTTGAGTTGCTT
711	GTTGCTTGGCGCCATATGGTAGACTGCTCATAGTGCACGCTATTGTCCAGGG
712	TCCAGGGTGATACCTATATTACACAGAGGAGGGGGACGCCAGTCTACCTA
713	CTACCTACGAATTGCAATTTGCGCAACTCCTCAAGTAGCAGTACCCACAGGT
714	CACAGGTTATTGCGATTGTGCCTGGGTAGTTGGAAACTCGTGAGACTCGATC
715	CTCGATCTGGTAGTGTAAACTCCAAGCACAGAAGGGGCTGTGGGATACAACC
716	TACAACCCGGCCCAAGAGGGGCAGATACCCATGTGCGAGATAGGCTACCCCGC
717	ACCCCGCCCCAAAACCATGCATATAATATTCACGAACACCTGGGGGCAACTC
718	GCAACTCCATACAATGGTTATGGTCAACTCCCTTTCTACCCTCAGTGTA
719	AGTGTAAATGTCTTATATGAAGGCTTTGCGAGCAGTGAGCGGGTCTTTAAGTT
720	TTAAGTTCCATTAGATACGTCGCGCTTAAAGTGAGAAATGGTGAATTAGA
721	AATTAGAGACGGTACTGAAGTTCCTTTGCATCGCGCTTCGGGAAGCACCG
722	AGCACCGGGCCGTTTGACTAAGTTTAGTAGCCGGAGCGGTTAATAGATAGG
723	AGATAGGAATAAAAGGAAGACCCAATCCGCTTACCGAGGTATTCCGATTGGT
724	GATTGGTCGATACGGTGGCACTAAGCGCAGCGATAGAAGTGTGCGCCGGATA
725	CCGGATATAAGTGATCAGCACGGTCTTATCGAACAGAGACATCTTACCTCCT
726	ACCTCCTAGTTCCAGAGCAGTGGGTAAGCTCAACGTATTGCCACGCATCTGG
727	CATCTGGTGGCGCCATGCTAGTTGATGTAGAAACGAGGTGGGCCCTGGAAC
728	CTGGAACGATCCAGGCCTACGCCCGTTTCATAGGAGAGAAAAACAGCATAAGG
729	CATAAGGTATGTCACCCGAGTATGGCGGGGACGAATTTGTGGTCCCGGACA
730	CCCGACACGTGAGGATTCTCCAACCTTCTGTATACTCAATGGGTGATGAGGT
731	ATGAGGTTCAAGTGGCGCTTTTTTATACTTGCAGACGGACACGTTCCGGATCT
732	CGGATCTCTTCATAAGCTGTTTCCATGAGTCGTAGACGGAGTCATATCCGCA
733	ATCCGCATTCAACGAAGGCTGGCTACAACGCTCCATCCACCAACCCACACGA
734	CACACGAACCTAAGCGGTCCCTACCAATATTTATCAGTGATTACAGTAAACG
735	GTAAACGGAGGTATCTTCTGGGCCCGTATGCAGAAGACCTTGACCCTTATA
736	CCTTATAGATGTAGTGGCGGATTTCAGAGGAATTACATCACCATAAATTACCA
737	ATTACCAAAGTAAGATCACGCCACGAAAAGTATTCGAGAGACTCTCAGAGGT
738	CAGAGTCCGGACACAACGCCCTAGGGCCCTTGGGCACTATTGAAGTTCATAT
739	TTCATATTAATGAAAAGTGATTTAGGACGAAAGCCTCTCGGCCACAGAATGG
740	AGAATGGGCAATTG



**Appendix Figure 21 MITOMI replicate agreement.**

Plotted are values of intensity of the replicates for (A) *CaCph2p*, (B) *CaHms1p* and (C) *CaTye7p* in MITOMI experiments.

**Appendix Table 3 Primers used in this study**

Name	Description	Purpose	Sequence (5' to 3')
JCP_1425	<i>CaHms1</i> into pLIC-H3 FW	Plasmid construction	AGCAGCCCCGGGTCAACAATTATTAATGAACCAAC
JCP_1429	<i>CaHms1</i> into pLIC-H3 RV	Plasmid construction	GTGGTGCTCGAGCTACATATCTTCATCAATAACTAAAC
JCP_1426	<i>CaCph2</i> into pLIC-H3 FW	Plasmid construction	AGCAGCCCCGGGGCCAAAGTCACAAAGCCAAAATC
JCP_1430	<i>CaCph2</i> into pLIC-H3 RV	Plasmid construction	GTGGTGCTCGAGTTATAATGACTTTGGATTCATGTTGGC
JCP_1513	<i>CaTye7</i> into pbRZ75 FW	Plasmid construction	CATATGGCTAGCACTAAACCAAAGAGAAGAGCACC
JCP_1660	<i>CaTye7</i> into pbRZ75 RV	Plasmid construction	GTGGTGCTCGAGCTATATTTACCACCCAATTTAATAAC
JCP_2044	<i>CpHms1</i> into pbRZ75 FW 1A	Plasmid construction	CATATGGCTAGCAGTACAACAATTACAGCTTCCAGTG
JCP_2006	<i>CpHms1</i> into pbRZ75 RV 1B	Plasmid construction	GATGGAATGGATTTGTCGATTCTTGATGATAAG
JCP_2007	<i>CpHms1</i> into pbRZ75 FW 2A	Plasmid construction	CTTATCATCAAGAATCGACAAAATCCATTCCATC
JCP_2008	<i>CpHms1</i> into pbRZ75 RV 2B	Plasmid construction	GTGGTGCTCGAGTTATTCTTCTTCATCAAATCAACACCC
JCP_2030	<i>AfSrbA</i> into pLIC-H3 FW	Plasmid construction	AGCAGCCCCGGGACTGGAAGCGATGACGATGGATCTA
JCP_2031	<i>AfSrbA</i> into pLIC-H3 RV	Plasmid construction	GTGGTGCTCGAGCTAGGTAACAATCTGATCAGCGGCCTTG
JCP_1783	Helix1 chimera into pLIC-H3 1A	Plasmid construction	ATTAAATCAGCTCATAATGTAATTGAACAACGATATCGAAATAAAATTAATGAT AAATTTAATGCTTTACAAAATTCT
JCP_1784	Helix1 chimera into pLIC-H3 1B	Plasmid construction	AGAATTTTGTAAGCATTAAATTTATCATTAATTTTATTTTCGATATCGTTGTTCAA TTACATTATGAGCTGATTTAAT
JCP_1357	Helix1 chimera into pLIC-H3 2A	Plasmid construction	GCACCATCATCATCACCATC
JCP_1785	Helix1 chimera into pLIC-H3 2B	Plasmid construction	CAATTACATTATGAGCTGATTTAATTTTGGATTTTGGC
JCP_1786	Helix1 chimera into pLIC-H3 3A	Plasmid construction	TAAATTTAATGCTTTACAAAATTCTGTGCCTGCTC
JCP_1358	Helix1 chimera into pLIC-H3 3B	Plasmid construction	CTTTCGGGCTTTGTTAGCAG
JCP_1839	Helix 2 chimera into pLIC-H3 1B	Plasmid construction	TTGTTCTTTATTCAATTTTCTTGCTGGTGTTAACC
JCP_1840	Helix 2 chimera into pLIC-H3 2A	Plasmid construction	GCAAGAAAATTGAATAAAGGAACAATATTAGCTAAATCTATTGAATATATTTAA TTTTTAGAAATGAAAAATGAAAGA
JCP_1841	Helix 2 chimera into pLIC-H3 2B	Plasmid construction	TCTTTCATTTTTCATTTCTAAAAATTTAATATATTCAATAGATTTAGCTAATATTG TTCCTTTATTCAATTTTCTTGC
JCP_1862	Helix 2 chimera into pLIC-H3 3A	Plasmid construction	ATTTTTAGAAATGAAAAATGAAAGATTGAAACAACA
JCP_1843	Helix1+Loop chimera 1B	Plasmid construction	GCTTTTTTTGGGCTAATATTCTTAAAGCAGGCACA
JCP_1844	Helix1+Loop chimera 2A	Plasmid construction	TTAAGAATATTAGCCCAAAA
JCP_1845	Helix1+Loop chimera 2B	Plasmid construction	TGGTTCTAATCCTTCTAAAT
JCP_1863	Helix1+Loop chimera 3A	Plasmid construction	TATTGATTTAGAAGGATTAGAACCAGCATCTAAGTTAAAC

JCP_2071	Loop chimera 1B	Plasmid construction	GCTTTTTTTGGGCTAATATTCTTAAAGCAGGCACAG
JCP_1844	Loop chimera 2A	Plasmid construction	TTAAGAATATTAGCCCAAAA
JCP_1949	Loop chimera 2B	Plasmid construction	ATTCAATTTTCTTGCTGGTTC
JCP_2000	Loop chimera 3A	Plasmid construction	ATTAGAACCAGCAAGAAAATTGAATAAAGCTAGTGTTTTAACC
JPC_1358	Loop chimera 3B	Plasmid construction	CTTTCGGGCTTTGTTAGCAG
JCP_1432	Cph2-myc tag fw	Myc-tagging	TGCCTTGCCATTTTCGTATTTGTTTCCTAATGCTATTCTTAACCCATCGCCATTGACTATTCAATTACGGATCCCCGGGTTAATTAACGG
JCP_1693	Cph2-myc tag rv	Myc-tagging	AACCCACATTCCTTATGAAACAAAATAAAATTAACAATCTATACCTGAAAAA AAGAAACACTGGCGGCCGCTCTAGAACTAGTGGATC
JCP_2261	qPCR – orf19.3242fw	qPCR	GCGGAGACAAAGATCGACAC
JCP_2262	qPCR – orf19.3242 rv	qPCR	TATTACTGACGCTGGTGGGG
JCP_1829	qPCR – orf19.921 fw	qPCR	GACACCAATCTCATCAATTAC
JCP_1830	qPCR – orf19.921 rv	qPCR	AAATCTACTGGGAATCCTCC
JCP_2462	qPCR – orf19.4941 fw	qPCR	CAAACCACAAGCAACTCCGA
JCP_2463	qPCR – orf19.4941 rv	qPCR	TGCGCAATACTGTCTGATGTG
JCP_1627	E-box fw	EMSA	TCTGGCCAACATCAGGTGAACGCAGGAAAA
JCP_1628	E-box rv	EMSA	TTTTCTGCGTTCACCTGATGTTGGCCAGA
JCP_1629	E-box fw	EMSA	TCTGGCCAACATTCTTTAAACGCAGGAAAA
JCP_1630	E-box rv	EMSA	TTTTCTGCGTTTAAAGAATGTTGGCCAGA
JCP_1442	Non-E-box fw	EMSA	ACAAAAAAAAAATCAGGCCATTTGTAATACT
JCP_1443	Non-E-box rv	EMSA	AGTATTACAAATGGCCTGATTTTTTTTTTGT
JCP_1498	Non-E-box mutated fw	EMSA	ACAAAAAAAAAATTCTAAAATTTGTAATACT
JCP_1499	Non-E-box mutated rv	EMSA	AGTATTACAAATTTTAGAATTTTTTTTTTGT
JCP_2314	Cph2 ChIP motif fw	EMSA	GGAAATTGGATGCAAAATCATAACGATTTTT
JCP_2315	Cph2 ChIP motif rv	EMSA	AAAAATCGTATGATTTTGCATCCAATTTCC
JCP_2316	Cph2 ChIP motif mutated fw	EMSA	GGAAATTGGACACAAACGCGTACGATTTTT
JCP_2317	Cph2 ChIP motif mutated rv	EMSA	AAAAATCGTACGCGTTTGTGTCCAATTTCC



**Appendix Table 4 Protein expression plasmids generated in this study**

Called	Plasmid name	Description	Reference
<b>His-<i>CaCph2</i></b>	JCP_567	pLIC-H3 with 6xHis <i>CaCph2</i> 197-302	This study
<b>HisMBP-<i>CaTye7</i></b>	JCP_726	pBRZ75 with 6xHis-MBP <i>CaTye7</i> 121-269	This study
<b>His-<i>CaHms1</i></b>	JCP_565	pLIC-H3 with 6xHis <i>CaHms1</i> 463-686	This study
<b>HisMBP-<i>CpHms1</i></b>	JCP_833	pBRZ75 with 6xHis-MBP <i>CpHms1</i> 486-659	This study
<b>His-<i>A/SrbA</i></b>	JCP_831	pLIC-H3 with 6xHis <i>A/SrbA</i> 463-686	This study
<b>His-Chimera1</b>	JCP_822	pLIC-H3 with 6xHis Helix1 chimera (1)	This study
<b>His-Chimera2</b>	JCP_823	pLIC-H3 with 6xHis Helix2 chimera (2)	This study
<b>His-Chimera3</b>	JCP_824	pLIC-H3 with 6xHis Helix1+Loop chimera (3)	This study
<b>His-Chimera4</b>	JCP_836	pLIC-H3 with 6xHis Loop chimera (4)	This study
<b>His-Anc4</b>	JCP_817	pLIC-H3 with 6xHis Ancestor 4	This study
<b>His-Anc5.3</b>	JCP_921	pLIC-H3 with 6xHis Ancestor 5.3	This study

(1) PCR was used to amplify DNA fragment of *Cph2* insert from plasmid JCP\_567 in two parts with primers 1357/1785 and 1786/1358. Helix1 of *Hms1* was generated by annealing the primer pair 1783/1784. The three fragments were joined by fusion PCR. The purified product was inserted in *SmaI/XhoI* sites of pLIC-H3.

(2) PCR was used to amplify DNA fragment of Helix1 chimera insert from plasmid JCP\_822 in two parts with primers 1357/1839 and 1862/1358. Helix2 of *Hms1* was generated by annealing the primer pair 1840/1841. The three fragments were joined by fusion PCR. The purified product was inserted in *SmaI/XhoI* sites of pLIC-H3.

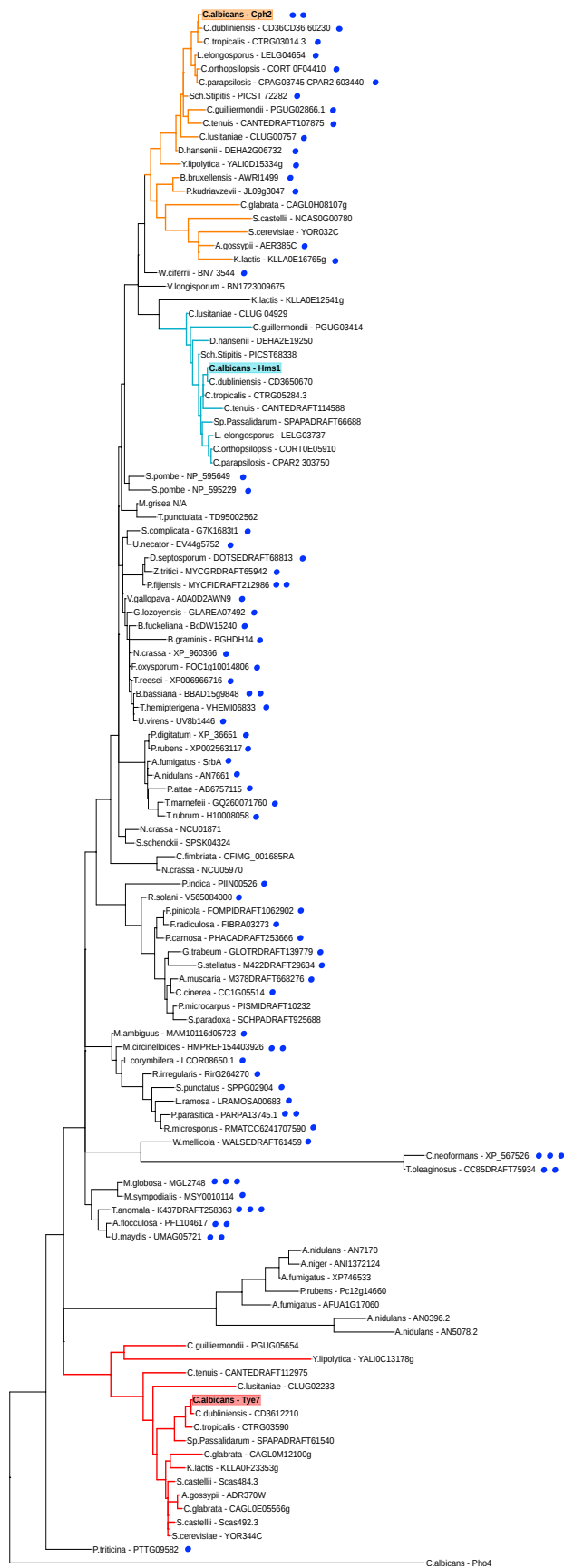
(3) PCR was used to amplify DNA fragment of *Cph2* insert from plasmid JCP\_822 in two parts with primers 1357/1843 and 1863/1358. The loop of *Hms1* was amplified from plasmid JCP\_565 with primers 1844/1845. The three fragments were joined by fusion PCR. The purified product was inserted in *SmaI/XhoI* sites of pLIC-H3.

(4) PCR was used to amplify DNA fragment of *Cph2* insert from plasmid JCP\_567 in two parts with primers 1357/2071 and 2000/1358. The loop of *Hms1* was amplified from plasmid JCP\_565 primers 1844/1949. The three fragments were joined by fusion PCR. The purified product was inserted in *SmaI/XhoI* sites of pLIC-H3.

**Appendix Table 5 Sequences of the DNA binding domain of fungal SREBPs**

Species	Systematic name	
<i>A.flocculosa</i>	PFL1 04617	KKVAHNAIERRYRNNINDRI AALRNSVPAATKLNKATILGKATDYIKYLKGRELR
<i>A.fumigatus</i>	AFUA 1G17060	.RAS..V..K..T.M.AKFTT.E.VITTPCSMK.CE..TS.IKC.QD.EE.NAA
<i>A.fumigatus</i>	Srba XP 749262	.R...V..K...A.L.EK..E..D...S.S...S..S...E..RH.EI.NK.
<i>A.fumigatus</i>	XP 746533	.RA...I..K...T.M.AKFV..EKAMSGPAS.K.SE..TN.IA.MQE.QDQNA
<i>A.gossypii</i>	ADR370W	Q.E...K..K...I...TKL.K.QQII.WTP...SM..E..V...LF.QNN.RL
<i>A.gossypii</i>	AER385C	ERMS..I..KK..T...K.LQ..EI...R...S..T.TIE...H.EEKCAF
<i>A.muscaria</i>	M378DRAFT 668276	P.TS.TT...T.L.A..QS..MA.L..R.CS..NV...VE...V..R..Q.
<i>A.nidulans</i>	AN0396.2	RRAS..VV.K...E.L.RKFHL.ETI.NKQYTSP...IDS.LS..ES.RSENHA
<i>A.nidulans</i>	AN5078.2	RRAS..IV.K...I.L.SKFRK.HEI.FCRSQPP..S.IDS.LN..ES.QREVHE
<i>A.nidulans</i>	AN7170	.RA...I..K...T.M.AKFV..EKAMSVSAS.K.SE..SN.IT.MQE.QEMNE.
<i>A.nidulans</i>	AN7661	.R...V..K...A.L.EK..E..D...S.N...S..S...RH.ET.NK.
<i>A.niger</i>	ANI 1 372124	.RA...I..K...T.M.AKFV..EKAMCGSAS.K.SE..TN.ITFMQE.QEENKV
<i>B.bassiana</i>	BBAD15 g9848	.T...M..K...T...K...D...S.H...V.S...E..RH.EK.NC.
<i>B.bruxellensis</i>	AWRI1499 3986	I.CS..L..KK..T...SK.VE..C..S.R...S...E..RH.EIKNEQ
<i>B.fuckeliana</i>	BcDW1 5240	.TT...M..K...L..K...D...S.H...V.S...E..RH.EK.NS.
<i>B.graminis</i>	BGHDH14 bgh01993	.T...M..K...T.L..K...D...S.H...VTTEYICHLEKRNK.NN.
<i>C.albicans</i>	orf19.1187 Cph2 XP 712449	D.NS..M..KK..T...TK.L...DA...S...SV.T...E...H.ESKNFI
<i>C.albicans</i>	orf19.4941 Tye7 XP 722152	Q.K...K..K...I...AK..GIQKII.WN.R...SM..E...E..LH.QKK.EE
<i>C.albicans</i>	orf19.921Hms1 XP 716760	I.S...V..Q...K...KFN..Q...T.R...G...A.SIE...F.EMKNE.
<i>C.cinerea</i>	CC1G 05514	P.TS.TT...T.L.A..QS..MA...R.CS..NV...VE..RV..K..M.
<i>C.dublinsiensis</i>	CD36 12210	Q.K...K..K...I...AK..GIQKII.WN.R...SM..E...E..LH.QKK.AE
<i>C.dublinsiensis</i>	CD36 50670	I.S...V..Q...K...KFN..Q...T.R...G...A.SIE...F.EMKNE.
<i>C.dublinsiensis</i>	CD36CD36 60230	D.NS..M..KK..T...TK.L...DA...S...SV.T...E...H.ESKNSV
<i>C.fimbriata</i>	CFIMG_001685RA	DRT...D...K..T.LK.K..E..DAI..PQ.VS.G.V.T...HT.ER.NKA
<i>C.glabrata</i>	CAGL0E05566g	Q.E...K..K...I...SKL.K.QQII.WTP...SM..E..V...L..QNN.KL
<i>C.glabrata</i>	CAGL0H08107g	.RAL.SIV.KK..T...E.L.E.K.T..T.R...S..H...E...F.ENENSD
<i>C.glabrata</i>	CAGL0M12100g	Q.L...K..KK..I...SK..Q.QKMI.WK..F..SI..Q..I..V..RNN.HL
<i>C.guilliermondii</i>	PGUG02866.1 PGUG 02866	D.IS..Q...K..T...TK.L...DA..S.S...SV.D...E..RH.ERKNET
<i>C.guilliermondii</i>	PGUG03414.1 PGUG 03414	GRS...I..Q...K...FDE.LAC...R...G...E.LVE..EF..LKNA.
<i>C.guilliermondii</i>	PGUG05654.1 PGUG 05654	Q.IG..K..KK...TKLEI.HRL..WIH.VS.SM.DG..E...Q.IE..K.
<i>C.lusitaniae</i>	CLUG 00757	T..S..M..KK..T...SK.ME..DA..T.S...SV.T...E...H.EHKNSL
<i>C.lusitaniae</i>	CLUG 02233	Q.E...KV.KK..V...QK.NS.QTII.WDL.V..SV..E..Y..T...AENEA
<i>C.lusitaniae</i>	CLUG 04929	T.S...V..Q...K...KFT..Q.A..T.R...G...T.SVE...F.ERKND.
<i>C.neoformans</i>	XP 567526 Sre1	TIGK..KT...QKVQAAQ.D..DAI..PNASA...IG.RV..EL.QK.SAK
<i>C.orthosilopsis</i>	CORT 0E05910	V.S...V..Q...K..NKFN..QE...T.R...G...A.SIE...F.ESKNDK
<i>C.orthosilopsis</i>	CORT 0F04410	D.SS..M..KK..T...SK.LI..DA...S...SV.T...E...H.EQKNA
<i>C.parapsilosis</i>	CPAG03309 CPAR2 303750	V.S...V..Q...K..NKFN..QE...T.R...G...A.SIE...F.ESKNDK
<i>C.parapsilosis</i>	CPAG03745 CPAR2 603440	D.SS..M..KK..T...SK.LI..DA...S...SV.T...E...H.EQKNAM
<i>C.tenuis</i>	CANTEDRAFT 107875	D.AS..V..KK..T...TK.LM..DA..S.A...SV.T...E..QH.EKKNEM
<i>C.tenuis</i>	CANTEDRAFT 112975	Q.A...V..KK..I...TK.ES.QKLI.SSK...SL..D..IE..MF.QQNQCCK
<i>C.tenuis</i>	CANTEDRAFT 114588	LRSS..V..Q...K...KFN..S...T.R...GV..S.SVE..RF.ELKND.
<i>C.tropicalis</i>	CTRG03014.3	D.NS..M..KK..T...TK.M..DA...S...SV.T...E...H.EAKNAI
<i>C.tropicalis</i>	CTRG03590.3 CTRG 03590	Q.K...K..K...I...AK..GIQKII.WN...SM..E...E..L..QK..KD
<i>C.tropicalis</i>	CTRG05284.3	I.S...V..Q...K...KFN..Q...T.R...G...A.SIE...F.ELKND.
<i>D.hansenii</i>	DEHA2E19250	T.S...V..Q...K...KFTV.Q.T..S.K...G...T.SIE...F.EMKNNS
<i>D.hansenii</i>	DEHA2G06732	D.SS..M..KK..T...SK.V...DA..S.S...SV.T...E...H.EHKNDM
<i>D.septosporum</i>	DOTSEDRAFT 68813	.TS..V..K...L..K.VE..A...H...VMA..E..RH.EK.NKT
<i>F.oxysporum</i>	FOC1 g10014806	.T...M..K...T.L..K...D...S.H...V.S...E..RH.EK.NN.
<i>F.pinicola</i>	FOMPIDRAFT 1062902	P.TS.TT...T.L.A..TG.KQA...R.MS..NV...AE...V..R..A.
<i>F.radiculosa</i>	FIBRA 03273	P.TS.TT...T.L.A..TG.KQA...R.MS..NV...AE...V..K..T.
<i>G.lozoyensis</i>	GLAREA 07492	.T...M..K...T.L..K...D...S.H...V.S...E..HH.EK.VK.
<i>G.trabeum</i>	GLOTRDRAFT 139779	G.TS.TT...T.L.A..QS..RA...R.TS..S...VE..RV..K..G.
<i>K.lactis</i>	KLLA0E12541g	TSSG..DS.KK...THFQF...TPS.V..VQ..S.SHE..H.ERKNSL
<i>K.lactis</i>	KLLA0E16765g	ERTS..V..KK..T...NK.VQ.KEII.S.K...S..V.TIE..QH.ENHVE.
<i>K.lactis</i>	KLLA0F23353g	QR.T..M..K...I...TK.GK.QKII.WKV...SM..E..V...L..QNN.RL





















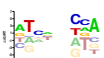






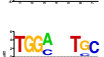

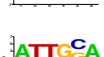
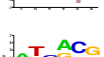
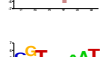

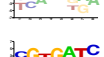
Appendix Figure 22 Extended phylogenetic tree of fungal SREBPs

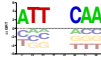
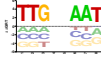
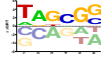


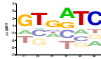
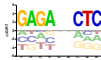
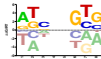

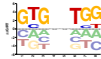
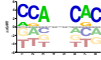

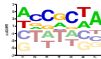


## Appendix Table 6 Comprehensive lists of DNA motifs derived from MITOMI.

The tables include all the DNA motifs resulting from the analysis of the MITOMI data with MatrixREDUCE ( $P < 1 \times 10^{-10}$ ). The derived motifs are ranked according to  $r^2$  and  $P$ -values. The model variants (topologies) X6, X7, X3N2X3 and X4N2X3 (in forward, reverse or both strands) were employed to query the data).

Cph2					
ID	Motif	Topology	Strand	P-value	$r^2$
27		X4N2X3	Both	4,83E-108	0,294332
9		X7	Forward	8,07E-92	0,257479
12		X7	Reverse	8,07E-92	0,257479
25		X4N2X3	Forward	1,86E-78	0,225693
15		X7	Both	5,09E-78	0,222197
16		X7	Both	4,19E-66	0,195116
19		X3N2X3	Forward	8,67E-62	0,184154
20		X3N2X3	Reverse	8,67E-62	0,184154
21		X3N2X3	Both	8,43E-58	0,173894
24		X4N2X3	Forward	4,08E-44	0,137717
7		X6	Both	4,01E-43	0,135035
28		X4N2X3	Both	1,29E-36	0,117239
2		X6	Forward	3,66E-31	0,102066
5		X6	Reverse	3,66E-31	0,102066
10		X7	Forward	7,96E-31	0,101118
13		X7	Reverse	7,96E-31	0,101118
1		X6	Forward	1,26E-29	0,0977422

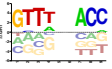
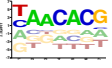

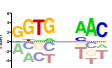
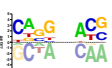




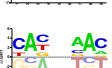

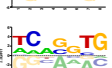
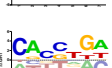
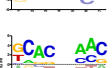

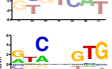
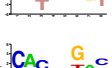
4		X6	Reverse	1,26E-29	0,0977422
26		X4N2X3	Forward	1,18E-27	0,0921713
30		X4N2X3	Both	1,28E-27	0,092074
17		X7	Both	1,27E-26	0,0892386
22		X3N2X3	Both	2,22E-26	0,0885481
11		X7	Forward	1,19E-24	0,0836163
14		X7	Reverse	1,19E-24	0,0836163
18		X7	Both	7,23E-24	0,0813749
29		X4N2X3	Both	2,48E-23	0,0798418
3		X6	Forward	6,92E-22	0,0756812
6		X6	Reverse	6,92E-22	0,0756812
8		X6	Both	2,71E-19	0,0681831
23		X3N2X3	Both	1,33E-16	0,0603417

Hms1					
ID	Motif	Topology	Strand	P-value	r <sup>2</sup>
11		X7	Forward	4,64E-101	0,277417
14		X7	Reverse	4,64E-101	0,277417
17		X7	Both	5,88E-88	0,247259
32		X4N2X3	Both	1,24E-76	0,220191
18		X7	Both	1,09E-66	0,195637
29		X4N2X3	Forward	4,07E-65	0,191682
12		X7	Forward	1,20E-62	0,185431
15		X7	Reverse	1,20E-62	0,185431
13		X7	Forward	1,62E-58	0,174884
16		X7	Reverse	1,62E-58	0,174884
28		X4N2X3	Forward	4,42E-44	0,136931
1		X6	Forward	1,15E-43	0,13581
5		X6	Reverse	1,15E-43	0,13581
9		X6	Both	1,55E-43	0,135468
22		X3N2X3	Forward	4,77E-38	0,120573
25		X3N2X3	Reverse	4,77E-38	0,120573
19		X7	Both	7,18E-38	1,20E-01

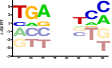
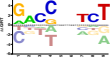
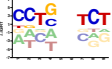
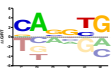



21		X3N2X3	Forward	2,60E-31	0,101953
24		X3N2X3	Reverse	2,60E-31	0,101953
3		X6	Forward	2,96E-31	0,099002
7		X6	Reverse	2,96E-31	0,099002
2		X6	Forward	3,51E-29	0,09599
6		X6	Reverse	3,51E-29	0,09599
31		X4N2X3	Forward	1,17E-27	0,0917065
27		X3N2X3	Both	2,29E-27	0,0908833
30		X4N2X3	Forward	1,35E-25	0,0858684
23		X3N2X3	Forward	2,16E-24	0,0824487
26		X3N2X3	Reverse	2,16E-24	0,0824487
10		X6	Both	7,02E-24	0,0781331
20		X7	Both	1,78E-23	0,0769749
4		X6	Forward	2,27E-21	0,0738074
8		X6	Reverse	2,27E-21	0,0738074



Tye7

ID	Motif	Topology	Strand	P-value	r <sup>2</sup>
39		X4N2X3	Reverse	9,29E-131	0,340835
13		X7	Forward	4,77E-95	0,263274
15		X7	Reverse	4,77E-95	0,263274
36		X4N2X3	Forward	7,55E-93	0,25821
42		X4N2X3	Both	1,38E-84	0,23888
18		X7	Both	1,52E-82	0,234024
1		X6	Forward	1,28E-70	0,205062
6		X6	Reverse	1,28E-70	0,205062
43		X4N2X3	Both	1,31E-63	0,187531
24		X3N2X3	Forward	1,39E-61	0,182394
29		X3N2X3	Reverse	1,39E-61	0,182394
14		X7	Forward	2,97E-61	0,181556
16		X7	Reverse	2,97E-61	0,181556
37		X4N2X3	Forward	9,00E-59	0,175214
19		X7	Both	2,35E-57	1,72E-01
23		X3N2X3	Forward	1,34E-55	0,167032
28		X3N2X3	Reverse	1,34E-55	0,167032

2		X6	Forward	5,95E-52	0,157535
7		X6	Reverse	5,95E-52	0,157535
11		X6	Both	1,02E-50	0,154293
22		X7	Both	9,17E-42	0,130435
35		X3N2X3	Both	1,55E-40	0,124399
20		X7	Both	5,92E-37	0,117349
33		X3N2X3	Both	1,67E-34	0,110609
5		X6	Forward	2,46E-34	0,110146
10		X6	Reverse	2,46E-34	0,110146
34		X3N2X3	Both	7,86E-33	0,105983
12		X6	Both	4,49E-30	0,0983051
17		X7	Reverse	1,17E-29	0,0971407
38		X4N2X3	Forward	1,06E-29	0,0944529
25		X3N2X3	Forward	5,66E-26	0,0867722
30		X3N2X3	Reverse	5,66E-26	0,0867722
26		X3N2X3	Forward	1,60E-23	0,0798098
31		X3N2X3	Reverse	1,60E-23	0,0798098
21		X7	Both	7,04E-23	0,0779758
27		X3N2X3	Forward	1,27E-22	0,0772401

32		X3N2X3	Reverse	1,27E-22	0,0772401
41		X4N2X3	Reverse	4,72E-22	0,0756153
40		X4N2X3	Reverse	2,67E-21	0,0734619
4		X6	Forward	6,34E-20	0,069509
9		X6	Reverse	6,34E-20	0,069509
3		X6	Forward	1,13E-19	0,0659003
8		X6	Reverse	1,13E-19	0,0659003

**Table 7 *C. albicans* strains used in this dissertation.**

Strain	Genotype	Source
SN250	$\frac{ura3\Delta::\lambda imm434::URA3-IRO1}{ura3\Delta::\lambda imm434}$ $\frac{arg4::hisG}{arg4::hisG}$ $\frac{his1::hisG}{his1::hisG}$ $\frac{leu2::hisG::CdHIS1}{leu2::hisG::CmLEU2}$	[156]
TF02	$\frac{ura3\Delta::\lambda imm434::URA3-IRO1}{ura3\Delta::\lambda imm434}$ $\frac{arg4::hisG}{arg4::hisG}$ $\frac{his1::hisG}{his1::hisG}$ $\frac{hms1\Delta::CdHIS1}{hms1\Delta::CmLEU2}$	[156]
TF138	$\frac{ura3\Delta::\lambda imm434::URA3-IRO1}{ura3\Delta::\lambda imm434}$ $\frac{arg4::hisG}{arg4::hisG}$ $\frac{his1::hisG}{his1::hisG}$ $\frac{cph2\Delta::CdHIS1}{cph2\Delta::CmLEU2}$	[156]
JCP880	$\frac{ura3\Delta::\lambda imm434::URA3-IRO1}{ura3\Delta::\lambda imm434}$ $\frac{arg4::hisG}{arg4::hisG}$ $\frac{his1::hisG}{his1::hisG}$ $\frac{leu2::hisG::CdHIS1}{leu2::hisG::CmLEU2}$ $\frac{CPH2(1-407)-myc::CPH2}{CPH2}$	In this study
JCP960	$\frac{ura3\Delta::\lambda imm434}{ura3\Delta::\lambda imm434}$ $\frac{arg4\Delta::hisG}{arg4\Delta::hisG}$ $\frac{his1\Delta::hisG}{his1\Delta::hisG}$ $\frac{ADE2}{ade2\Delta::CPH2p-myc-CPH2-URA3}$	[95]
JCP1018	$\frac{ura3\Delta::\lambda imm434::URA3-IRO1}{ura3\Delta::\lambda imm434}$ $\frac{arg4::hisG}{arg4::hisG}$ $\frac{his1::hisG}{his1::hisG}$ $\frac{leu2::hisG::CdHIS1}{leu2::hisG::CmLEU2}$ $\frac{SHE3-myc::SHE3}{SHE3}$	Lena Böhm

**Appendix Table 8 List including the most differentially expressed transcripts under anaerobic conditions.**

This table include the transcripts whose values of expression change were higher than  $\log_2$  fold-change  $|4|$ . Yellow colour represents up-regulation and blue indicates down-regulation. Description of transcripts was obtained from the Candida Genome Database ([www.candidagenome.org](http://www.candidagenome.org)).

ORF	Gene name	Log2 Fold-change	Description
orf19.6420	<i>PGA13</i>	12,56	GPI-anchored cell wall protein involved in cell wall synthesis
orf19.3499	<i>LDG8</i>	11,20	Secreted protein
orf19.6274	<i>PBR1</i>	10,28	Protein of unknown function; required for cohesion and adhesion
orf19.4690	<i>SMF11</i>	9,87	NRAMP metal ion transporter domain-containing protein
orf19.7455	orf19.7455	9,23	Ortholog of <i>C. dubliniensis</i> CD36 : Cd36_86630
orf19.2539	orf19.2539	7,33	Protein of unknown function
orf19.5674	<i>PGA10</i>	5,99	GPI anchored membrane protein
orf19.7550	<i>IFA14</i>	5,87	Putative LPF family protein
orf19.1321	<i>HWP1</i>	5,63	Hyphal cell wall protein
orf19.3908	orf19.3908	5,50	Protein of unknown function
orf19.7585	<i>INO1</i>	5,35	Inositol-1-phosphate synthase
orf19.5302	<i>PGA31</i>	5,30	Cell wall protein; putative GPI anchor
orf19.344	orf19.344	5,16	Protein of unknown function
orf19.4082	<i>DDR48</i>	5,12	Immunogenic stress-associated protein
orf19.2160	<i>NAG4</i>	5,05	Putative fungal-specific transporter
orf19.2481	orf19.2481	5,02	Protein of unknown function
orf19.7313	<i>SSU1</i>	5,02	Protein similar to <i>S. cerevisiae</i> Ssu1 sulfite transport protein
orf19.1539	orf19.1539	4,99	Protein of unknown function
orf19.1440.1	orf19.1440.1	4,91	Protein of unknown function
orf19.6661	orf19.6661	4,91	Predicted ORF
orf19.3374	<i>ECE1</i>	4,84	Candidalysin, cytolytic peptide toxin
orf19.535	<i>RBR1</i>	4,83	Glycosylphosphatidylinositol (GPI)-anchored cell wall protein
orf19.3655	orf19.3655	4,73	Ortholog of <i>C. dubliniensis</i> CD36
orf19.711	orf19.711	4,73	Protein of unknown function
orf19.1691	orf19.1691	4,70	Plasma-membrane-localised protein
orf19.4279	<i>MNN1</i>	4,66	Putative alpha-1,3-mannosyltransferase
orf19.6840	orf19.6840	4,65	Protein of unknown function
orf19.3740	<i>PGA23</i>	4,60	Putative GPI-anchored protein of unknown function
orf19.6852.1	orf19.6852.1	4,58	Protein of unknown function
orf19.2591	orf19.2591	4,55	Protein of unknown function
orf19.5814.1	orf19.5814.1	4,42	Protein of unknown function
orf19.4789	<i>ALD97</i>	4,39	Has domain(s) with predicted metal ion binding activity
orf19.265	orf19.265	4,30	Protein with a ribonuclease III domain
orf19.2451	<i>PGA45</i>	4,29	Putative GPI-anchored cell wall protein
orf19.489	<i>DAP1</i>	4,28	Similar to mammalian progesterone receptor
orf19.4664	<i>NAT4</i>	4,23	Putative histone acetyltransferase
orf19.6601	orf19.6601	4,20	Protein of unknown function
orf19.217	orf19.217	4,16	Ortholog(s) have sequence-specific DNA binding activity
orf19.849	<i>MNN4</i>	4,15	Ortholog(s) have activator activity
orf19.7221	<i>SET3</i>	4,14	NAD-dependent histone deacetylase
orf19.1369	<i>PYD3</i>	4,12	Protein with predicted peptidase domains

orf19.6484	<i>LDG4</i>	4,11	Protein of unknown function
orf19.3461	orf19.3461	4,10	Protein of unknown function
orf19.720	<i>GST3</i>	4,09	Glutathione S-transferase
orf19.6601.1	<i>YKE2</i>	4,06	Possible heterohexameric Gim/prefoldin protein complex subunit
orf19.5874	orf19.5874	4,06	Protein of unknown function
orf19.2833	<i>PGA34</i>	-4,02	Putative GPI-anchored protein
orf19.5992	<i>WOR2</i>	-4,04	Zn(II)2Cys6 transcription factor
orf19.7017	<i>YOX1</i>	-4,04	Putative homeodomain-containing transcription factor
orf19.3934	<i>CAR1</i>	-4,08	Arginase; arginine catabolism
orf19.3694	orf19.3694	-4,09	Unknown function
orf19.6010	<i>CDC5</i>	-4,12	Polo-like kinase
orf19.3869	orf19.3869	-4,12	Protein of unknown function
orf19.2475	<i>PGA26</i>	-4,25	GPI-anchored adhesin-like protein of the cell wall
orf19.2738	<i>SUL2</i>	-4,32	Putative sulfate transport
orf19.3810	<i>MTD1</i>	-4,34	Unknown function
orf19.6925	<i>HTB1</i>	-4,35	Histone H2B
orf19.6169	<i>ATO1</i>	-4,49	Putative fungal-specific transmembrane protein
orf19.7111.1	<i>SOD3</i>	-4,50	Cytosolic manganese-containing superoxide dismutase
orf19.7077	<i>FRE7</i>	-4,51	Putative ferric reductase
orf19.3895	<i>CHT2</i>	-4,52	GPI-linked chitinase; required for normal filamentous growth
orf19.4669	<i>AAT22</i>	-4,54	Aspartate aminotransferase
orf19.5188	<i>CHS1</i>	-4,56	Chitin synthase
orf19.1996	<i>CHA1</i>	-4,65	Similar to catabolic ser/thr dehydratases
orf19.385	<i>GCV2</i>	-4,79	Glycine decarboxylase P subunit
orf19.6644	orf19.6644	-4,82	Protein of unknown function
orf19.557	orf19.557	-4,82	Protein of unknown function
orf19.4527	<i>HGT1</i>	-4,93	High-affinity MFS glucose transporter
orf19.6139	<i>FRE7</i>	-4,99	Copper-regulated cupric reductase
orf19.6140	<i>FRE30</i>	-5,00	Protein with similarity to ferric reductases
orf19.5308	orf19.5308	-5,19	Protein of unknown function
orf19.1334	orf19.1334	-5,26	Protein of unknown function
orf19.1854	<i>HHF22</i>	-5,72	Putative histone H4
orf19.5307	<i>JEN2</i>	-5,74	Dicarboxylic acid transporter
orf19.1522	orf19.1522	-5,79	Unknown
orf19.1061	<i>HHT21</i>	-5,94	Putative histone H3
orf19.4933	<i>FAD3</i>	-6,21	Omega-3 fatty acid desaturase
orf19.3893	<i>SCW11</i>	-6,25	Cell wall protein; induced in high iron
orf19.3668	<i>HGT2</i>	-6,31	Putative MFS glucose transporter
orf19.670.2	orf19.670.2	-6,40	Protein of unknown function
orf19.4450.1	orf19.4450.1	-6,46	Unknown function
orf19.3981	<i>MAL31</i>	-6,65	Putative high-affinity maltose transporter
orf19.3664	<i>HSP31</i>	-6,73	Putative 30 kDa heat shock protein
orf19.7094	<i>HGT12</i>	-6,79	Glucose, fructose, mannose transporter
orf19.7586	<i>CHT3</i>	-6,84	Major chitinase; functional homolog of <i>S. cerevisiae</i> Cts1p
orf19.5305	<i>RHD3</i>	-7,12	GPI-anchored yeast-associated cell wall protein
orf19.1853	<i>HHT2</i>	-7,20	Putative histone H3
orf19.7218	<i>RBE1</i>	-7,44	Pry family cell wall protein
orf19.5267	orf19.5267	-7,51	Putative cell wall adhesin-like protein
orf19.4599	<i>PHO89</i>	-8,56	Putative phosphate permease
orf19.2758	<i>PGA38</i>	-9,78	Putative adhesin-like GPI-anchored protein
orf19.4910	<i>FGR41</i>	-10,05	Putative GPI-anchored adhesin-like protein

## REFERENCES

1. Carlin JL. Mutations Are the Raw Materials of Evolution. *Nat Educ.* 2011; 3:10.
2. Loewe L, Hill WG. The population genetics of mutations: Good, bad and indifferent. *Philos Trans R Soc Lond B Biol Sci.* 2010; 365: 1153 - 1167.
3. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science.* 2000; 290: 1151 - 1155.
4. Wagner A. Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends Genet.* 2001; 17: 237–239.
5. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics.* 1999; 151: 1531 - 1545.
6. Seoighe C, Wolfe KH. Extent of genomic rearrangement after genome duplication in yeast. *Proc Natl Acad Sci.* 2002; 95: 4447 - 4452.
7. Storz JF, Opazo JC, Hoffmann FG. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Mol Phylogenet Evol.* 2013; 66: 469–478.
8. Storz JF. Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. *Physiology.* 2016;31: 223–232.
9. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: Building the web of life. *Nat Rev Genet.* 2015;16: 472–482.
10. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol.* 2005;3: 679–687.
11. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet.* 2008;9: 605–618.
12. Britten RJ, Davidson EH. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol.* 1971;46: 111–138.

13. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science*. 1975;188: 107–116.
14. Lerman DN, Michalak P, Helin AB, Bettencourt BR, Feder ME. Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. *Mol Biol Evol*. 2003; 20: 135 - 144.
15. Lerman DN, Feder ME. Naturally occurring transposable elements disrupt hsp70 promoter function in *Drosophila melanogaster*. *Mol Biol Evol*. 2005; 22: 776 - 783.
16. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. *Nat Genet*. 2002; 30: 233 - 237.
17. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, *et al*. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*. 2007; 39: 31 - 40.
18. Wray GA. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet*. 2007;8: 206 - 216.
19. Sorrells TR, Johnson AD. Making sense of transcription networks. *Cell*. 2015; 161: 714 - 723.
20. Nosedal I, Johnson AD. How transcription networks evolve and produce biological novelty. *Cold Spring Harb Symp Quant Biol*. 2015; 80: 265 - 274.
21. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*. 2012; 69: 3613 - 3634.
22. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem*. 1992; 61: 1053 - 1095.
23. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, *et al*. Zinc-finger proteins in health and disease. *Cell Death Discov*. 2017; 3: 17071.
24. Landschulz WH, Johnson PF, McKnight SL. The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. *Science*. 1988; 240: 1759 - 1764.



25. Robichon C, Karimova G, Beckwith J, Ladant D. Role of leucine zipper motifs in association of the *Escherichia coli* cell division proteins FtsL and FtsB. *J Bacteriol.* 2011; 193: 4988 - 4992.
26. Jones S. An overview of the basic helix-loop-helix proteins. *Genome Biology.* 2004; 5; 226.
27. Vervoort M. The basic helix-loop-helix protein family : comparative genomics and phylogenetic analysis derivation of comprehensive sets. *Genome Biol.* 2001; 3: 754 - 770.
28. Prud'homme B, Gompel N, Carroll SB. Emerging principles of regulatory evolution. *Proc Natl Acad Sci.* 2007; 104: 8605 - 8612.
29. Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M. Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr Biol.* 2007;17: 1007 - 1013.
30. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. The evolution of combinatorial gene regulation in fungi. *PLoS Biol.* 2008; 6: e38.
31. Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol.* 2010; 8: e1000329.
32. Carroll SB. Evo-Devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell.* 2008; 134: 25 - 36.
33. Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, *et al.* Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature.* 2006;440: 1050 - 1053.
34. Gompel N, Prud'Homme B, Wittkopp PJ, Kassner VA, Carroll SB. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature.* 2005;433: 481 - 487.
35. Stern DL. Perspective: Evolutionary developmental biology and the problem of variation. *Evolution.* 2000; 54: 1079 - 1091.
36. Perez JC, Fordyce PM, Lohse MB, Hanson-Smith V, De Risi JL, Johnson AD. How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. *Genes Dev.* 2014;28: 1272 - 1277.

37. Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AFY. Ancestral Reconstruction. *PLOS Comput Biol*. 2016; 12: e1004763.
38. Merkl R, Sterner R. Ancestral protein reconstruction: Techniques and applications. *Biological Chemistry*. 2016; 397: 1 - 21.
39. Harms MJ, Thornton JW. Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology*. 2010; 20: 360–366.
40. Anderson DW, McKeown AN, Thornton JW. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife*. 2015;4: 1 - 26.
41. Bridgham JT, Ortlund EA, Thornton JW. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature*. 2009; 461: 515 - 519.
42. Choi J, Kim S-H. A genome Tree of Life for the Fungi kingdom. *Proc Natl Acad Sci*. 2017; 114: 9391 - 9396.
43. Bruns T. Evolutionary biology: A kingdom revised. *Nature*. 2006; 443: 758 - 761.
44. Suh S-O, Blackwell M, Kurtzman CP, Lachance M-A. Phylogenetics of Saccharomycetales, the ascomycete yeasts. *Mycologia*. 2006;98: 1006 - 1017.
45. Pérez JC. *Candida albicans* dwelling in the mammalian gut. *Curr Opin Microbiol*. 2019; 52: 41 - 46.
46. Gow NAR, Van De Veerdonk FL, Brown AJP, Netea MG. *Candida albicans* morphogenesis and host defence: Discriminating invasion from colonization. *Nat Rev Microbiol*. 2012; 10: 112 - 122.
47. Mayer FL, Wilson D, Hube B. *Candida albicans* pathogenicity mechanisms. *Virulence*. 2013; 4: 119 - 128.
48. Iliev ID, Leonardi I. Fungal dysbiosis: immunity and interactions at mucosal barriers. *Nat Rev Immunol*. 2017; 17: 635.
49. Wang X, Briggs MR, Hua X, Yokoyama C, Goldstein JL, Brown MS. Nuclear

- protein that binds sterol regulatory element of low density lipoprotein receptor promoter. *J Biol Chem.* 1993; 268: 14497 - 14504.
50. Robinson KA. *Saccharomyces cerevisiae* basic helix-loop-helix proteins regulate diverse biological processes. *Nucleic Acids Res.* 2000; 28: 1499 - 1505.
  51. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell.* 2014; 158: 1431 - 1443.
  52. Shimizu T, Toumoto A, Ihara K, Shimizu M, Kyogoku Y, Ogawa N, *et al.* Crystal structure of *PHO4* bHLH domain-DNA complex: Flanking base recognition. *EMBO J.* 1997;16: 4689 - 4697.
  53. Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, *et al.* Genomic regions flanking E-Box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* 2013;3: 1093 - 1104.
  54. Ellenberger T, Fass D, Arnaud M, Harrison SC. Crystal structure of transcription factor *E47*: E-box recognition by a basic region helix-loop-helix dimer. *Genes Dev.* 1994; 8: 970 - 980.
  55. Rosenfeld JM, Osborne TF. *HLH106*, a *Drosophila* sterol regulatory element-binding protein in a natural cholesterol auxotroph. *J Biol Chem.* 1998; 273: 16112 - 16121.
  56. Reed BD, Charos AE, Szekely AM, Weissman SM, Snyder M. Genome-wide occupancy of *SREBP1* and its partners *NFY* and *SP1* reveals novel functional roles and combinatorial regulation of distinct classes of genes. *PLoS Genet.* 2008; 4: e1000133.
  57. Longo A, Guanga GP, Rose RB. Crystal structure of E47-NeuroD1/Beta2 bHLH domain-DNA complex: Heterodimer selectivity and DNA recognition. *Biochemistry.* 2008; 47: 218 - 229.
  58. Kim JB, Spotts GD, Halvorsen YD, Shih HM, Ellenberger T, Towle HC, *et al.* Dual DNA binding specificity of *ADD1/SREBP1* controlled by a single amino acid in the basic helix-loop-helix domain. *Mol Cell Biol.* 1995; 15: 2582 - 2588.
  59. Párraga A, Bellolell L, Ferré-D'Amaré AR, Burley SK. Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 Å resolution.

- Structure. 1998;6: 661 - 72.
60. Geertz M, Maerkl SJ. Experimental strategies for studying transcription factor-DNA binding specificities. *Brief Funct Genomics*. 2010; 9: 362 - 373.
  61. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc*. 2007; 2: 1849 - 1861.
  62. Djordjevic M. SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering*. 2007; 24: 179 - 189.
  63. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*. 1990; 249: 505 - 510.
  64. Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. *PLoS Comput Biol*. 2009; e1000590.
  65. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, *et al*. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010; 20: 861 - 873.
  66. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*. 2006; 24: 1429 - 1435.
  67. Bulyk ML, Huang X, Choo Y, Church GM. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci*. 2001; 98: 7158 - 7163.
  68. Tanay A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*. 2006; 16: 962 - 972.
  69. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, *et al*. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*. 2009; 19: 556 - 566.
  70. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*. 2007; 315: 233 - 237.

71. Fordyce PM, Gerber D, Tran D, Zheng J, Li H, De Risi JL, *et al.* *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat Biotechnol.* 2010; 28: 970 - 975.
72. Xu YZ, Kanagaratham C, Jancik S, Radzioch D. Promoter deletion analysis using a dual-luciferase reporter system. *Methods Mol Biol.* 2013; 977: 79 - 93.
73. Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics.* 2009; 10: 669 - 680.
74. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2017; 45: D158 - D169.
75. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for Bigger Datasets. *Mol Biol Evol.* 2016; 33: 1870 - 1874.
76. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics.* 2005; 21: 2104 - 2105.
77. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics.* 2008; 24: 1662 - 1668.
78. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003; 13: 2498 - 2504.
79. Ward LD, Bussemaker HJ. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics.* 2008; 24: 165 - 171.
80. Hernday AD, Noble SM, Mitrovich QM, Johnson AD. Genetics and molecular biology in *Candida albicans*. *Methods Enzymol.* 2010; 470: 737 - 758.
81. Langmead B, Salzberg SL, Langmead. Bowtie2. *Nat Methods.* 2013; 9: 357 - 359.
82. Homann OR, Johnson AD. MochiView: Versatile software for genome browsing and DNA motif analysis. *BMC Biol.* 2010; 8: 49

83. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc.* 2012; 7: 1728 - 1740.
84. Zordan RE, Miller MG, Galgoczy DJ, Tuch BB, Johnson AD. Interlocking transcriptional feedback loops control white-opaque switching in *Candida albicans*. *PLoS Biol.* 2007; 5: e256.
85. Nobile CJ, Fox EP, Nett JE, Sorrells TR, Mitrovich QM, Hernday AD, *et al.* A recently evolved transcriptional network controls biofilm development in *Candida albicans*. *Cell.* 2012; 148: 126 - 138.
86. Pérez JC, Kumamoto CA, Johnson AD. *Candida albicans* commensalism and pathogenicity are intertwined traits directed by a tightly knit transcriptional regulatory circuit. *PLoS Biol.* 2013; 11: e1001510
87. Teste M-A, Duquenne M, François JM, Parrou J-L. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*. *BMC Mol Biol.* 2009; 10: 99.
88. Cain CW, Lohse MB, Homann OR, Sil A, Johnson AD. A conserved transcriptional regulator governs fungal morphology in widely diverged species. *Genetics.* 2012; 190: 511 - 521.
89. Lohse MB, Zordan RE, Cain CW, Johnson AD. Distinct class of DNA-binding domains is exemplified by a master regulator of phenotypic switching in *Candida albicans*. *Proc Natl Acad Sci.* 2010; 107: 14105 - 14110.
90. Hanson-Smith V, Johnson A. PhyloBot: A web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLoS Comput Biol.* 2016; 12: 1 - 10.
91. Hughes AL, Todd BL, Espenshade PJ. SREBP pathway responds to sterols and functions as an oxygen sensor in fission yeast. *Cell.* 2005; 120: 831 - 842.
92. Sato T, Lopez MC, Sugioka S, Jigami Y, Baker H V., Uemura H. The E-box DNA binding protein Sgc1p suppresses the *gcr2* mutation, which is involved in transcriptional activation of glycolytic genes in *Saccharomyces cerevisiae*. *FEBS Lett.* 1999; 463: 307 - 311.
93. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature.* 2007;449: 54.

94. Askew C, Sellam A, Epp E, Hogues H, Mullick A, Nantel A, *et al.* Transcriptional regulation of carbohydrate metabolism in the human pathogen *Candida albicans*. *PLoS Pathog.* 2009; 5: e1000612
95. Lane S, Di Lena P, Tormanen K, Baldi P, Liua H. Function and regulation of Cph2 in *Candida albicans*. *Eukaryot Cell.* 2015; 14: 1114 - 1126.
96. McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell.* 2014; 159: 58 - 68.
97. Yokoyama C, Wang XD, Briggs MR, Admon A, Wu J, Hua XX, *et al.* Srebp-1, a basic-helix-loop-helix-leucine zipper protein that controls transcription of the low-density-lipoprotein receptor gene. *Cell.* 1993; 75: 187 - 197.
98. Hua X, Sakai J, Ho YK, Goldstein JL, Brown MS. Hairpin orientation of sterol regulatory element-binding protein-2 in cell membranes as determined by protease protection. *J Biol Chem.* 1995; 270: 29422 - 29427.
99. Bien CM, Espenshade PJ. Sterol regulatory element binding proteins in fungi: Hypoxic transcription factors linked to pathogenesis. *Eukaryot Cell.* 2010; 9: 352 - 359.
100. Horton JD, Goldstein JL, Brown MS. SREBPs: transcriptional mediators of lipid homeostasis. *Cold Spring Harb Symp Quant Biol.* 2002; 67: 491 - 498.
101. Espenshade PJ, Hughes AL. Regulation of sterol synthesis in eukaryotes. *Annu Rev Genet.* 2007; 41: 401 - 427.
102. Chang YC, Bien CM, Lee H, Espenshade PJ, Kwon-Chung KJ. Sre1p, a regulator of oxygen sensing and sterol homeostasis, is required for virulence in *Cryptococcus neoformans*. *Mol Microbiol.* 2007; 64: 614 - 629.
103. Todd BL, Stewart E V., Burg JS, Hughes AL, Espenshade PJ. Sterol regulatory element binding protein is a principal regulator of anaerobic gene expression in fission yeast. *Mol Cell Biol.* 2006; 26: 2817 - 2831.
104. Nomura T, Horikawa M, Shimamura S, Hashimoto T, Sakamoto K. Fat accumulation in *Caenorhabditis elegans* is mediated by SREBP homolog *SBP-1*. *Genes Nutr.* 2010; 5: 17 - 27.
105. Maguire SL, Wang C, Holland LM, Brunel F, Neuvéglise C, Nicaud JM, *et al.* Zinc finger transcription factors displaced SREBP proteins as the major

- sterol regulators during Saccharomycotina evolution. PLoS Genet. 2014; 10: e1004076
106. Sudbery P, Gow N, Berman J. The distinct morphogenic states of *Candida albicans*. Trends in Microbiology. 2004; 12: 317 - 324.
  107. Noble SM, Gianetti BA, Witchley JN. *Candida albicans* cell-type switching and functional plasticity in the mammalian host. Nat Rev Microbiol. 2017; 15: 96 - 108.
  108. Böhm L, Torsin S, Tint SH, Eckstein MT, Ludwig T, Pérez JC. The yeast form of the fungus *Candida albicans* promotes persistence in the gut of gnotobiotic mice. PLoS Pathog. 2017;13: 1 - 26.
  109. Brand A, Gow NA. Mechanisms of hypha orientation of fungi. Current Opinion in Microbiology. 2009; 12: 350 - 357.
  110. Moyes DL, Runglall M, Murciano C, Shen C, Nayar D, Thavaraj S, *et al.* A biphasic innate immune MAPK response discriminates between the yeast and hyphal forms of candida albicans in epithelial cells. Cell Host Microbe. 2010; 16: 225 - 235.
  111. Phan QT, Myers CL, Fu Y, Sheppard DC, Yeaman MR, Welch WH, *et al.* Als3 is a *Candida albicans* invasin that binds to cadherins and induces endocytosis by host cells. PLoS Biol. 2007; 5: e64.
  112. Braun BR, Johnson AD. Control of filament formation in *Candida albicans* by the transcriptional repressor *TUP1*. Science. 1997; 4: 105 - 109.
  113. Lo HJ, Köhler JR, Didomenico B, Loebenberg D, Cacciapuoti A, Fink GR. Nonfilamentous *C. albicans* mutants are avirulent. Cell. 1997; 90: 939 - 949.
  114. Murad AMA, Leng P, Straffon M, Wishart J, Macaskill S, MacCallum D, *et al.* *NRG1* represses yeast-hypha morphogenesis and hypha-specific gene expression in *Candida albicans*. EMBO J. 2001; 20: 4742 - 4752.
  115. Klengel T, Liang WJ, Chaloupka J, Ruoff C, Schröppel K, Naglik JR, *et al.* Fungal adenylyl cyclase integrates CO<sub>2</sub> sensing with cAMP signaling and virulence. Curr Biol. 2005; 15: 2021 - 2026.
  116. Rocha CRC, Schroppel K, Harcus D, Marcil A, Dignard D, Taylor BN, *et al.* Signaling through adenylyl cyclase is essential for hyphal growth and



- virulence in the pathogenic fungus *Candida albicans*. *Mol Biol Cell*. 2013; 12: 3631 - 3643.
117. Roman E, Nombela C, Pla J. The Sho1 adaptor protein links oxidative stress to morphogenesis and cell wall biosynthesis in the fungal pathogen *Candida albicans*. *Mol Cell Biol*. 2005; 25: 10611 - 10627.
  118. Posas F, Saito H. Osmotic activation of the HOG MAPK pathway via Ste11p MAPKKK: Scaffold role of Pbs2p MAPKK. *Science*. 1997; 276: 1702 - 1705.
  119. Lu Y, Su C, Solis N V., Filler SG, Liu H. Synergistic regulation of hyphal elongation by hypoxia, CO<sub>2</sub>, and nutrient conditions controls the virulence of *Candida albicans*. *Cell Host Microbe*. 2013; 14: 499 - 509.
  120. Stoldt VR, Sonneborn A, Leuker CE, Ernst JF. Efg1p, an essential regulator of morphogenesis of the human pathogen *Candida albicans*, is a member of a conserved class of bHLH proteins regulating morphogenetic processes in fungi. *EMBO J*. 1997; 16: 1982 - 1991.
  121. Sonneborn A, Bockmühl DP, Ernst JF. Chlamydospore formation in *Candida albicans* requires the Efg1p morphogenetic regulator. *Infect Immun*. 1999; 67: 5514 - 5517.
  122. Synnott JM, Guida A, Mulhern-Haughey S, Higgins DG, Butler G. Regulation of the hypoxic response in *Candida albicans*. *Eukaryot Cell*. 2010; 9: 1734 - 1746.
  123. Setiadi ER, Doedt T, Cottier F, Noffz C, Ernst JF. Transcriptional response of *Candida albicans* to hypoxia: linkage of oxygen sensing and Efg1p-regulatory networks. *J Mol Biol*. 2006; 361: 399 - 411.
  124. Mulhern SM, Logue ME, Butler G. *Candida albicans* transcription factor Ace2 regulates metabolism and is required for filamentation in hypoxic conditions. *Eukaryot Cell*. 2006; 5: 2001 - 2013.
  125. Doedt T. APSES proteins regulate morphogenesis and metabolism in *Candida albicans*. *Mol Biol Cell*. 2004; 15: 3167 - 3180.
  126. Giusani AD, Vences M, Kumamoto CA. Invasive filamentous growth of *Candida albicans* is promoted by Czf1p-dependent relief of Efg1p-mediated repression. *Genetics*. 2002; 160: 1749 - 1753.
  127. Albenberg L, Esipova T V., Judge CP, Bittinger K, Chen J, Laughlin A, *et al.*

- Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology*. 2014; 147: 1055 - 1063.
128. Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol*. 2015; 14: 20 - 32.
  129. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30: 2114 - 2120.
  130. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29: 15 - 21.
  131. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015; 31: 166 - 169.
  132. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014; 15: 550.
  133. Chung D, Barker BM, Carey CC, Merriman B, Werner ER, Lechner BE, *et al*. CHIP-seq and *in vivo* transcriptome analyses of the *Aspergillus fumigatus* SREBP SrbA reveals a new regulator of the fungal hypoxia response and virulence. *PLoS Pathog*. 2014; 10: e1004487.
  134. Brown Jr DH, Giusani AD, Chen X, Kumamoto CA. Filamentous growth of *Candida albicans* in response to physical environmental cues and its regulation by the unique *CZF1* gene. *Mol Microbiol*. 1999; 34: 651 - 662.
  135. Ghosh AK, Wangsanut T, Fonzi WA, Rolfes RJ. The *GRF10* homeobox gene regulates filamentous growth in the human fungal pathogen *Candida albicans*. *FEMS Yeast Res*. 2015; 15: fov093.
  136. Shapiro RS, Sellam A, Tebbji F, Whiteway M, Nantel A, Cowen LE. Pho85, Pcl1, and Hms1 signaling governs *Candida albicans* morphogenesis induced by high temperature or Hsp90 compromise. *Curr Biol*. 2012; 22: 461 - 470.
  137. Lane S, Zhou S, Pan T, Dai Q, Liu H. The basic helix-loop-helix transcription factor Cph2 regulates hyphal development in *Candida albicans* partly via *TEC1*. *Mol Cell Biol*. 2001; 21: 6418 - 28.
  138. Sudbery PE. Growth of *Candida albicans* hyphae. *Nat Rev Microbiol*. 2011; 9: 737.

139. Rosenbach A, Dignard D, Pierce J V., Whiteway M, Kumamoto CA. Adaptations of *Candida albicans* for growth in the mammalian intestinal tract. *Eukaryot Cell*. 2010; 9: 1075 - 1086.
140. Cote P, Hogues H, Whiteway M. Transcriptional analysis of the *Candida albicans* cell cycle. *Mol Biol Cell*. 2009; 20: 3363 - 3373.
141. Willger SD, Puttikamonkul S, Kim K-H, Burritt JB, Grahl N, Metzler LJ, *et al.* A sterol-regulatory element binding protein is required for cell polarity, hypoxia adaptation, azole drug resistance, and virulence in *Aspergillus fumigatus*. *PLoS Pathog*. 2008; 4: e1000200.
142. Desai PR, van Wijlick L, Kurtz D, Juchimiuk M, Ernst JF. Hypoxia and temperature regulated morphogenesis in *Candida albicans*. *PLoS Genet*. 2015; 11: e1005447.
143. Falvo J V., Lin CH, Tsytsykova A V., Hwang PK, Thanos D, Goldfeld AE, *et al.* A dimer-specific function of the transcription factor NFATp. *Proc Natl Acad Sci*. 2008; 105: 19637 - 19642.
144. Osborne TF, Espenshade PJ. Evolutionary conservation and adaptation in the mechanism that regulates SREBP action: What a long, strange tRIP it's been. *Genes Dev*. 2009; 23: 2578 - 2591.
145. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science*. 2007; 317: 1544 - 1548.
146. Weinreich DM, Delaney NF, DePristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*. 2006; 312: 111 - 114.
147. Mayer FL, Wilson D, Hube B. *Candida albicans* pathogenicity mechanisms. *Virulence*. 2013; 4: 119 - 128.
148. Sudbery PE. Growth of *Candida albicans* hyphae. *Nat Rev Microbiol*. 2011; 9: 737.
149. Frankel N, Wang S, Stern DL. Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. *Proc Natl Acad Sci*. 2012; 109: 20975 - 20979.
150. Pires ND, Yi K, Breuninger H, Catarino B, Menand B, Dolan L. Recruitment

- and remodeling of an ancient gene regulatory network during land plant evolution. *Proc Natl Acad Sci.* 2013; 110: 9571 - 9576.
151. Klis FM, Mol P, Hellingwerf K, Brul S. Dynamics of cell wall structure in *Saccharomyces cerevisiae*. *FEMS Microbiology Reviews.* 2002; 26: 239 - 256.
  152. Ruiz-Herrera J, Victoria Elorza M, Valentín E, Sentandreu R. Molecular organization of the cell wall of *Candida albicans* and its relation to pathogenicity. *FEMS Yeast Research.* 2006; 6: 14 - 29.
  153. Znaidi S, van Wijlick L, Hernández-Cervantes A, Sertour N, Desseyn JL, Vincent F, *et al.* Systematic gene overexpression in *Candida albicans* identifies a regulator of early adaptation to the mammalian gut. *Cell Microbiol.* 2018; 20: e12890.
  154. Prieto D, Román E, Correia I, Pla J. The HOG pathway is critical for the colonization of the mouse gastrointestinal tract by *Candida albicans*. *PLoS One.* 2014; 9: e0087128
  155. Plaine A, Walker L, Da Costa G, Mora-Montes HM, McKinnon A, Gow NAR, *et al.* Functional analysis of *Candida albicans* GPI-anchored proteins: Roles in cell wall integrity and caspofungin sensitivity. *Fungal Genet Biol.* 2008; 45: 1404 - 1414.
  156. Noble SM, Johnson AD. Strains and strategies for large-scale gene deletion studies of the diploid human fungal pathogen *Candida albicans*. *Eukaryot Cell.* 2005; 4: 298 - 309.

# Curriculum Vitae





## List of publications

### Publications included in this thesis

**Valentina del Olmo** Toledo, Robert Puccinelli, Polly M. Fordyce, J. Christian Pérez (2018) *Diversification of DNA binding specificities enabled SREBP transcription regulators to expand the repertoire of cellular functions that they govern in fungi*. **PLoS Genet**, 2018; 14: e1007884.

### Other publications

Sergio Moreno-Velásquez, Su Tint, **Valentina del Olmo Toledo**, Sanda Torsin, Sonakshi De, and J. Christian Pérez. *The conserved yeast regulators Rtg1/3 govern sphingolipid homeostasis in the human fungal pathogen C. albicans*. Submitted (under review).

Fabien Cottier, Sarah Sherrington, Sarah Cockeril, **Valentina del Olmo Toledo**, Stephen Kissane, Hélène Tournu, Luisa Orsini, Glen Palmer, J. Christian Pérez, and Rebecca Hall. *Remasking of Candida albicans b-glucan in response to environmental pH is regulated by quorum sensing*. Submitted (under review).



## Conferences and courses attended

### Conferences

Eureka 10th International GSLS Symposium 2015, Würzburg, Germany. Poster: *“Understanding how a family of transcriptional regulators was repurposed to enable Candida albicans to thrive in mammalian hosts “*

EMBO, Experimental approaches to evolution and ecology using yeast and other model systems 2016, Heidelberg, Germany. Poster: *“Evolution of DNA binding preferences in fungal members of the SREBP family of transcription regulators”*

Eureka 13th International GSLS Symposium 2018, Würzburg, Germany. Poster: *“Diversification of DNA binding preferences enabled SREBP transcription regulators to expand the repertoire of cellular functions that they govern in fungi”*

### Transferable skills training

#### GSLS workshops

- Poster design, Barry Drees (2015, Würzburg)
- Scientific writing, Dzifa Vode (2016, Würzburg)
- Good scientific practice & copyright in science, Dr. Stephan Schröder-Köhne & Christian Schmauch (2016, Würzburg)
- Oral presentation skills, Avril Arthur-Goettig (2016, Würzburg)
- Marketing your skills – Job interview training, Robert Zaal (2016, Würzburg)
- Intercultural communication, Ernestine Schneider & Martina Wernz-Hornberger (2017, Würzburg)
- Introduction to biotech industries, Dr. Christian Grote-Westrick (2018, Würzburg)

#### Other courses

- Laboratory animal science – Basic course main focus: Mice/Rats as per FELASA B guidelines, Berliner fortbildungen (2015, Berlin)
- Statistical data analysis with SPSS, Daniel Keller (2016, Würzburg)
- Infection and Immunity – Short course, Institute for RNA-based Infection research - HIRI (2019, Würzburg)