Universidade de
Aveiro
2022

FILIPA AZEVEDO SAMPAIO

Caracterização genética de isolados de *Candida albicans* provenientes de amostras clínicas distintas e do ambiente clínico

Genetic characterisation of *Candida albicans* isolates from distinct clinical human samples and the clinical environment

Universidade de Aveiro
Ano 2022

FILIPA AZEVEDO SAMPAIO

Caracterização genética de isolados de *Candida albicans* provenientes de amostras clínicas distintas e do ambiente clínico

Genetic characterisation of *Candida albicans* isolates from distinct clinical human samples and the clinical environment

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Microbiologia, realizada sob a orientação científica da Professora Doutora Gabriela Moura, Professora Auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro e da Doutora Maria João Carvalho, investigadora do Instituto de Biomedicina da Universidade de Aveiro.

This study was funded through CANCYL (CANCYL-POCI-01-0145-FEDER-031849), COVID2123 (CENTRO-01-01D2-FEDER-000001), GenomePT (POCI-01-0145-FEDER-022184) and iBiMED (UIDB/04501/2020 and UIDP/04501/2020)

o júri

Presidente	Prof. Doutor Carlos Miguel Miguez Barroso Professor auxiliar c/ agregação, Departamento de Biologia, Universidade de Aveiro
Vogal - Arguente	Doutora Ana Rita Macedo Bezerra Investigador Doutorada (nível 1), Instituto de Biomedicina, Departamento de Ciências Médicas, Universidade de Aveiro
Vogal - Coorientadora	Doutora Maria João Mendes de Carvalho Investigador Doutorada (nível 1), Instituto de Biomedicina, Departamento de Ciências Médicas, Universidade de Aveiro

agradecimentos

À minha orientadora Professora Doutora Gabriela Moura por me ter dado a oportunidade de fazer parte do seu grupo de trabalho.

Um obrigado especial à minha coorientadora, Doutora Maria João Carvalho, por todo o empenho, simpatia, carinho, disponibilidade, apoio, ajuda e papel fundamental na realização deste trabalho.

À Carla Oliveira, Rita Guimarães e Inês Sousa pelo suporte e ajuda que forneceram ao longo deste trabalho.

À Ana Poim, um enorme obrigado por toda a amizade, disponibilidade, conselhos, dedicação e formação que me proporcionou.

Por último e igualmente importante, agradeço aos meus pais, melhores amigos e namorado pelo apoio e incentivo incansável para a realização desta dissertação.

palavras-chave

Candida albicans, Sequenciação total do genoma, Polimorfismos, Resistência antifúngica, Multilocus Sequence Typing, Análise da ontologia genética, Perda de Heterozigosidade

resumo

Candida albicans é um microrganismo comensal que pertence à microbiota normal humana, capaz de se tornar patogénico, e é uma das principais causas de infeções fúngicas em humanos, com taxas de mortalidade até 50%. Com base neste problema, é imperativo caracterizar este patógeno rotineiramente, não só fenotipicamente, mas também genotipicamente, identificar características genómicas responsáveis pela doença, resistência a fármacos, adaptabilidade a nichos ecológicos e desvendar a diversidade genética dos isolados causadores de infeções. Assim, neste estudo, uma coleção de 76 estirpes de *C. albicans* recolhidas de amostras de sangue, vaginais, orais e de dispositivos médicos foi caracterizada através da sequenciação do genoma total e análises de bioinformática. Explorou-se as sequências genómicas dos isolados para discriminar e contextualizar epidemiologicamente as estirpes, identificar genes com polimorfismos de nucleotídeo único (SNPs) e contribuir para o conhecimento do varioma de *C. albicans*. A análise do genoma dos isolados mostrou que um dos isolados foi anteriormente mal classificado como *C. albicans*, sendo *C. glabrata*. Foi identificado um número de SNPs maior do que o habitual, possivelmente devido ao número elevado de SNPs homozigóticos. Foram identificados genes com SNPs comuns a todos os isolados, com SNPs exclusivos de isolados de cada tipo de amostra, e com SNPs comuns a todos os isolados de cada origem. A frequência de genes com SNPs missense envolvidos em funções moleculares e processos biológicos foi muitas vezes significativamente superior à frequência na estirpe de referência para os mesmos conjuntos de ontologia genética.

keywords

Candida albicans, Whole genome sequencing, Polimorphisms, Antifungal resistance, Multilocus Sequence Typing, Gene ontology analysis , Loss of Heterozygosity

abstract

Candida albicans is a commensal microorganism of the human normal microbiota, capable of turning pathogenic, and one of the leading causes of human fungal infections with mortality rates as high as 50%. Based on this problem, it is imperative to routinely characterize this pathogen, not only phenotypically but also genotypically, to identify the genomic traits responsible for disease, drug resistance, adaptability to ecological niches, and unravel the genetic diversity among distinct isolates. As so, in this study, a collection of 76 *C. albicans* strains collected from blood, vaginal and oral samples, and samples from medical devices was characterized by whole genome sequencing and bioinformatics analysis. Their genome sequence was explored to discriminate and epidemiologically contextualize strains globally, identify genes with single nucleotide polymorphisms (SNPs) and contribute to the knowledge on the *C. albicans* variome. The isolates genome analysis showed that one of the isolates had been misidentified as *C. albicans*, being *C. glabrata*. A higher number of SNPs than usual was identified, possibly due to the high number of homozygous SNPs. Genes with SNPs common to all isolates, with SNPs exclusive of isolates from each sample type, and with SNPs common to all isolates of each origin were identified. Often, the frequency of genes with missense SNPs involved in molecular functions and biological processes was significatively higher when compared to the reference strain frequency in those gene ontology sets.

Index

Index of Figures

Index of Tables

Abbreviations

WGD – Whole Genome Duplication

GUT – Gastrointestinal Induced Transition

ORFs – Open Reding Frames

CGD – *Candida* Genome Database

MTL – Mating-Type Locus

RPS – Repeat sequence arrays

MRS – Major Repeated Sequence

GI – Gastrointestinal

SAP – Secreted Aspartyl Protease

MIC – Minimal Inhibitory Concentration

ABC – ATP-binding Cassete

MFS – Major Facilitator Superfamily

DREs – Drug Response Elements

Hsp90 – Heat shock protein 90

PKC1 – Protein kinase C

mRNA – Messenger RNA

TAC1 – Transcriptional Activator of *CDR* genes

MRR – Multidrug Resistance Regulator

CDR – *Candida* Drug Resistance

INDELs – Small Insertions and Deletions

CNVs – Copy Number Variations

LOH – Loss of Heterozygosity

WGS – Whole Genome Sequencing

MLST – Multilocus Sequencing Type

ST – Sequence Type

GO – Gene Ontology

SNPs – Single Nucleotide Polymorphisms

YPD – Yeast Peptone Dextrose

OD – Optical Density

TE – Tris-EDTA

TNE – Tris-NaCl-EDTA

BLT – Bead-Linked Transposomes

TB1 – Tagmentation Buffer 1

TSB – Tagmentation Stop Buffer

TWB – Tagment Wash Buffer

EPM – Enhanced PCR Mix

PCR – Polymerase Chain Reaction

SPB – Sample Purification Beads

BLAST – Basic Local Alignment Tool

CC – Clonal Complex

ATG – Autophagy Related Genes

TORC1 – Target of Rapamycin Complex 1

I. INTRODUCTION

1. *Candida albicans*

Candida albicans is a commensal microorganism that belongs to the human normal microbiota as a heterozygous diploid yeast of mucosal surfaces and the digestive system, capable of polymorphic switching and also of turning into an pathogenic organism[1–3].

C. albicans is one of the leading causes of human fungal infections, mainly superficial vaginal or mucosal oral and may also, under propitious conditions, enter the bloodstream leading to deep-tissue infections[2–4].

In healthy individuals, *C. albicans* is often inoffensive, and cohabits with the other members of the local microbiota. However, changes in the host microbiota, alterations in the host immune response, or variations in the local environment can lead to *C. albicans* abnormal growth leading to infection[5].

In a substantial number of immunocompromised individuals, *C. albicans* induces systemic infection and can switch from local opportunistic or commensal infections of the mouth, throat, and reproductive tract to a systemic invasive candidiasis affecting the circulatory system, bones, and brain[2]. These infections present an effective worldwide mortality rate, which varies across geographical regions and lies between 10% and 47% despite the availability of antifungal therapies[4].

C. albicans transmission occurs vertically from mother to child, and infections arise predominantly from the endogenous microbiota[4]. This is a distinct characteristic since other major pathogens are fundamentally environmental fungi that have developed traits that promote pathogenicity in humans[4].

Bloodstream infections caused by member of the *Candida* genus are responsible for mortality rates as high as 50% among the infected patients and despite fifteen *Candida* species can be pathogenic for humans, more than 90% of reported invasive candidiasis are associated with the five most common species, *C. albicans*, *C. glabrata*, *C. parapsilosis*, *C. krusei*, and *C. tropicalis*[2,6]. Of these, *C. albicans* remains the most common cause of lethal systemic candidiasis. Based on these reasons, *C. albicans* has gained importance as a human pathogen, which promoted the study of this organism to understand its biology. In the last two decades, substantial advances have been made in understanding pathogenicity, genome structure and dynamics, pattern of gene expression in different conditions, drug resistance, biofilm formation, and host-parasite interactions in *C. albicans*[6]. Thus, *C. albicans* gained importance as a member of an elite group of model organisms at least for fungal pathogens[6].

1.2. Phylogeny and Taxonomy

The phylum Ascomycota contains 3 subphyla. One of these is the sub-phylum Saccharomycotina which includes the class Saccharomycetes and the order Saccharomycetales. This order has approximately 16 families. *C. albicans* is a species within a clade of one of these families, *the Saccharomycetales incertae sedis* family. This family contains two sub-groups, the CTG clade, the members of which unusually translate the CTG codon as serine instead of leucine, and the whole genome duplication (WGD) clade, which consists predominantly of the *Saccharomyces* and of other species for which the genomes have undergone complete duplication. *C. albicans* belongs to the CTG clade, which also contains the majority of the medically relevant *Candida* species[7] (Figure 1).

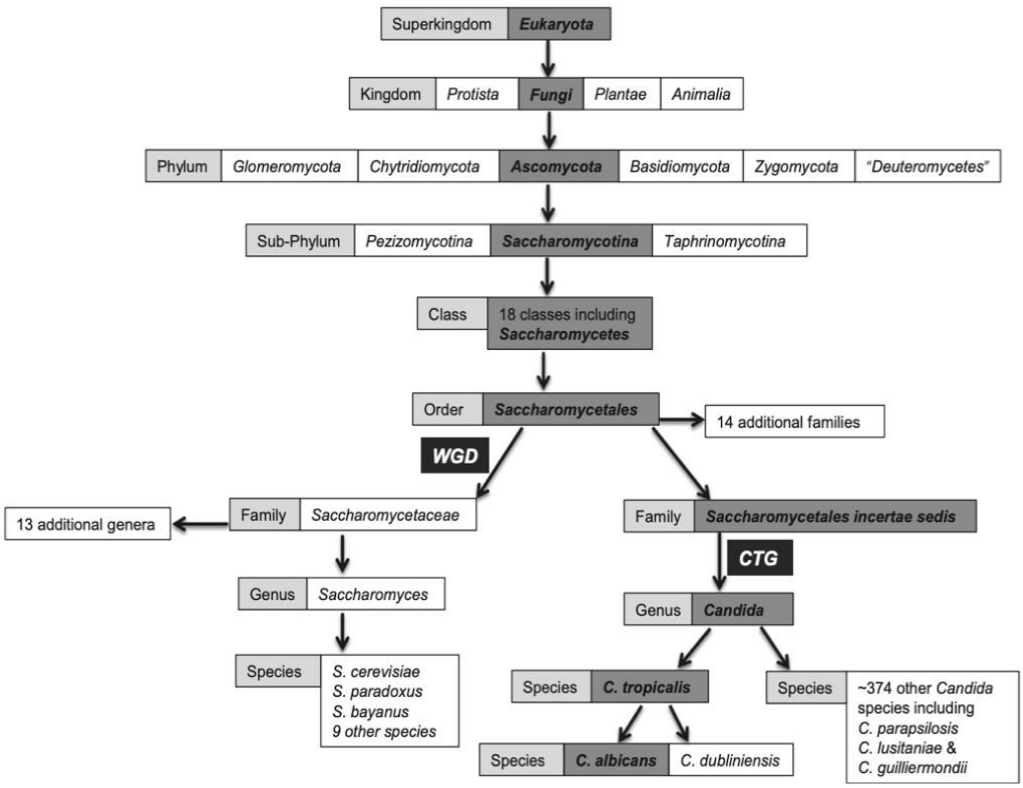


Figure 1: Ancestry and phylogeny of *C. albicans* [7].

1.3. Phenotypic features

C. albicans appears in several morphologic forms, such as yeast or blastospores, hyphae, pseudohyphae, chlamydospore, opaque, grey, and Gastrointestinal Induced Transition[8] (GUT) as pictured in Figure 2.

Yeast, hyphae, pseudohyphae and chlamydospore are the four cell types first being described, while opaque, grey, and GUT are an alternative to the standard yeast morphology[8–12].

Standard yeasts or white cells, are single cells with a round to oval morphology, measuring around 5-6 μm who reproduce by budding. In the budding process, the nuclear division occurs at the junction between the mother and daughter cells. As the descending cells separate themselves from the mother cells after cytokinesis they are considered unicellular[9].

Hyphal cells develop from an unbudded yeast cell and are characterized by branched chains of tubular cells, with no narrowing at the sites of septation[8,10]. The nuclear division process occurs inside the daughter cells. Afterwards, one progeny nucleus migrates back to the mother cells. Hyphal cells remain attached after cytokinesis, so that iterative cycles of cell division produce branched multicellular filamentary structures called mycelia[11]. This morphologic form is naturally invasive and penetrates the host's epithelial barriers through the expression of specific virulence factors such as degradative enzymes, cell surface adhesins and Candidalysin (a pore-forming toxin)[12].

Pseudohyphae are ellipsoid shaped cells with phenotypic features of both yeasts and hyphae. Such as hyphae, they remain attached after cytokinesis and generate mycelia after several cell divisions, but sites of septation are narrowed comparatively to the rest of the cell; similarly to yeasts, nuclear division occurs at mother-daughter junctions, and their shape is ellipsoid[8,11].

Clamydospores are large spherical cells with thick walls and are produced by cells at the distal ends of mycelial filaments, the suspensor cells. These cells' nuclear division occurs inside the parental suspension cell, followed by the migration of a progeny nucleus to the nascent chlamydospore, which remains linked to the mother cell[11].

Opaque cells is the first morphotype derived from white cells[9]. These present an ellipsoidal shape with protuberances on the surface and are slightly darker, matte, and flattened compared to the white cells. This cell type is related to the *C. albicans* parasexual cycle[9].

Gray cells also have an ellipsoidal shape and a similar length to the opaque cells but are smaller and do not present protuberances on their surface. These cells have a much

higher efficiency at mating compared to white cells and much lower compared to opaque cells[9].

The GUT phenotype is associated with the gastrointestinal tract of mammals. GUT cells are very similar to the opaque and gray cells, however, they don't exhibit protuberances on their surface and tend to be wider and less efficient at mating[9,10].

C. albicans commensal and pathogenic states are determined by the morphological transition from the yeast to the hyphae form, which may lead to tissue invasion, macrophage evasion, host-cell adhesion, and development of clinically relevant biofilm communities. Therefore, the post-transcriptional mechanisms underlying this transition control the virulence processes by influencing mRNA stability, alternative transcript localization and translation[2].

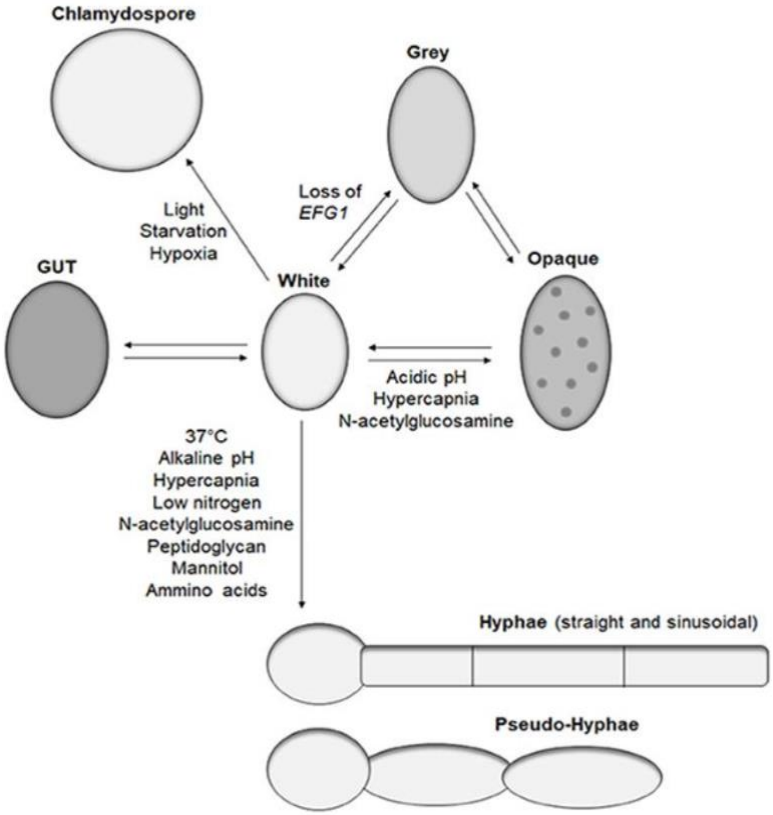


Figure 2: *C. albicans* morphological switches during the infection process[9].

1.4. Genome and life cycle

C. albicans has rarely been isolated in nature other than in an animal host and has probably coevolved along with humans for millions of years. It is presumed, therefore, that *C. albicans* genome contains the information that enables this fungus to thrive in its human host in competition with the immune system and with other microbiota[13].

This organism has a 15.4 megabase (Mb) genome arranged in eight chromosomes exhibiting 33.5% GC content and it contains 6107 protein-coding genes[6,14]. About 774 of the 6107 genes/ORFs (Open Reading Frames) are specific to *C. albicans* and homologues for these genes/ORFs are not available in *Saccharomyces cerevisiae.* Based on the gene annotation data available in CGD (*Candida* Genome Database), the functions of only 22.97% (1403 genes) of the genes have been experimentally verified, whereas 77.03% (4705 genes) of them remain uncharacterized in *C. albicans* and their functions have been assigned based on sequence analysis. Moreover, 152 genes/ORFs are still in the "dubious" category for which no experimental evidence is available and seems to be indistinguishable from noncoding sequences[6]. Generally, this genome exhibits some dynamics and a more frequently occurrence of truncations, translocations, and other mutational events compared to other microbes[6].

The complete genome of a *C. albicans* standard laboratory isolate, SC5314, has been sequenced due to its widespread use for molecular and genetic analysis worldwide. This strain has eight distinct chromosomes in duplicate ranging from 1030 to 3200kb[6].

In fact, C. albicans genome is highly plastic due to DNA small insertions and deletions (INDELs), copy number variations (CNVs), loss of heterozygosity (LOH) and karyotypic variation. These chromosomal rearrangements are tolerated in *C. albicans*, and often occur in response to stresses such as host-pathogen interactions, heat shock and the presence of antifungal drugs[3,7,15].

This microorganism has evolved unusual mechanisms in order to maintain genetic diversity in the absence of a complete sexual cycle[7,13]. These include the parasexual cycle, aneuploidy, gain and loss of heterozygosity and repetitive DNA sequences.

1.4.1. The parasexual cycle and aneuploidy

Unlike most yeasts, *C. albicans* is a diploid organism with no known haploid phase, and so it was considered to be asexual for a long time. However, early assemblies of the *C. albicans* genome sequence revealed a set of genes homolog to the *Saccharomyces cerevisiae* mating type locus, the mating type-like locus (MTL)[13]. Further work led to the

discovery that cells of the opposite homozygous mating type are able to mate by conjugation to form tetraploid zygotes[7].

Despite the efficient mechanisms for cell-cell conjugation, the attempts to demonstrate meiosis, and thereby complete a sexual cycle, have been unsuccessful so far. Instead, a parasexual cycle has been observed, which promotes the chromosome loss on the tetraploid zygotes until they reach a near diploid state with high levels of homozygosity and high frequencies of aneuploidy[7,13]. However, the parasexual cycle occurs rarely, possibly only under stressful conditions[7] and has only been demonstrated in mutants lacking the *BAR1* gene, a gene that encodes an aspartic-type endopeptidase involved in degradation of alpha pheromone[16,17].

The main function of the parasexual cycle is to enable diversity during times of stress, revealing new combinations of recessive traits by loss of heterozygosity (LOH), or resulting in aneuploidy and copy number variation promoting the adaptation to adverse environmental conditions[7].

Aneuploidy is the presence of an abnormal number of chromosomes and is an integral part of the *C. albicans* parasexual cycle[7]. Despite aneuploidy being commonly associated with fitness defects in eukaryotes, aneuploid forms of *C. albicans* can be advantageous under specific conditions. In particular, *C. albicans* cells harboring certain supernumerary chromosomes, such as an isochromosome of Chr5, can provide resistance against azole drugs[14]. This phenomenon can result from defects in DNA replication or division machinery and can involve complete or partial chromosomes[7].

1.4.2. Heterozygosity

C. albicans diploid genome sequence is highly heterozygous, even when compared to other *Candida* species[7]. It is estimated that approximately 4% of the *C. albicans* genome exhibits heterozygosity and that its frequency is of 4.21 polymorphisms per kb, or 1 polymorphism per 237 bases[7,13].

These heterozygosities are distributed unevenly across the genome, with a highest prevalence on chromosomes 5 and 6[13]. Over half of the approximately 6400 *C. albicans* genes contain allelic differences, and it is thought that two-thirds of these polymorphisms alter the protein sequence[13]. Also, this heterozygosity hides any recessive deleterious mutations that may be present in the genome, and may contribute significantly to strain fitness[7].

The considerable allelic variation in the *C. albicans* genome results from tandem repeat sequences, with many trinucleotides tandem repeats located in coding regions of the genome. This suggests that the frequency with which seemingly equivalent heterozygous mutants display phenotypic differences might be higher than expected[13]. Nevertheless, heterozygosity may be lost by several processes such as mitotic recombination, gene conversion between homologous chromosomes, DNA crossovers or by chromosome loss and duplication[7].

LOH is the loss of heterozygous positions between two homologous chromosomes, which transforms these regions into homozygous[14,18]. This event promotes genetic plasticity and can reveal recessive alleles with deleterious results. LOH can occur in chromosome segments of various lengths or even in the entire chromosome, however, such event is rare in entire chromosomes. Large-tract LOH are associated to mitotic crossovers or break-induced replication, and normally occur from the site of the DNA break to the end of the respective chromosome arm[14,18]. Short-tract LOH occur via gene conversion or double crossovers and are mostly located at telomeres, regions with repeats and in genes with repetitive elements[14,18].

1.4.3. DNA repetitive sequences

The *C. albicans* genome has several different types of repetitive DNA elements frequently associated with chromosomal rearrangements that contribute to genomic and phenotypic plasticity. Much of the karyotypic variability of *C. albicans* is due to the major repeat sequence (MRS) regions, the largest nontelomeric homologous sequences that have been identified in *C. albicans* and which account for approximately 3% of the total genomic content[7,15].

Complete MRS regions were found in 7 of the 8 *C. albicans* chromosomes, whereas only a partial MRS region is present on chromosome 3. These MRS hotspots are formed by several repeat sequence arrays (RPS), which are between two non-repeating sequences, the RB2 and HOK. Due to its size and presence on all chromosomes, RPS serves as a breakpoint for chromosomal rearrangements yielding chromosomal length polymorphisms, reciprocal translocations, chromosomal deletions and other aneuploidies. In the absence of a meiotic cycle, the MRS works as a source of homology across the majority of the *C. albicans* chromosomes, enabling reciprocal recombination events to occur between non-homologous chromosomes. These events promote a significantly increase of genomic diversity and can affect the phenotypes of the resultant cells[7].

1.5. Habitat

Unlike most fungal pathogens, *C. albicans* is generally considered to be obligately associated with warm-blooded animals[4]. Mammals are most exposed to microbes on the skin and mucosal of the gastrointestinal (GI), respiratory and reproductive trats[2,4,19]. Therebefore, epithelial surfaces on the mucosal tissues represent the main sites of *C. albicans* residence, where it establishes a commensal relationship with the host[2,4,19,20].

Colonization occurs during or shortly after birth, and transmission may be hereditary or through intimate physical contact. At this point, *C. albicans* becomes a commensal and its growth and location within the host is controlled by the normal microbiota, host physical barriers and immune system with which the yeast is continuously or transiently interacting[20,21]. At this stage, *C. albicans* competes with other microbes in different body sites for nutrition and adapts to different host conditions such as pH, temperature and nutrients[21]. Due to its high genomic plasticity, its adaptation occurs with ease and within short time periods[21].

In order to survive, *C. albicans* has to rapidly adapt to changing micro-environments, which implicates their lifestyle change to an opportunistic pathogen[21]. The mechanisms most implicated in this rapid adaptation is the yeast-to-hypha transition and the high frequency of phenotypic switching[22]. However, there are other virulence factors implicated in this process.

Although rarely, *C. albicans* can also be isolated from plants, soil or other environmental substrates such as trees, shrubs and grass[23].

1.6. Pathogenicity

C. albicans is a commensal microorganism often harmless in healthy individuals, However, in immunocompromised or immunologically deficient patients *C. albicans* can promote disease[1,5].

The several diseases caused by *C. albicans* are divided in three types: superficial and mucosal infections, and invasive candidiasis[24,25].

The superficial and mucosal infections comprises dermatitis and mucosal infections[24]. Dermatitis appears in both oncology and transplant patients due to the fragility of their immunity system.

Mucosal infections or candidiasis involve infections of oropharynx, esophagus, or vulvovaginal mucosa. These are usually due to changes in the normal microbiota after exposure to broad-spectrum antibiotics and/or after administration of chemotherapeutic agents that can also cause dysbiosis or mucosa breakdown. Candidiasis represents the most frequent fungal disease affecting populations worldwide, affecting both immunocompromised and healthy individuals[25].

Invasive candidiasis, a bloodstream infection that can be disseminated to multiple organs, includes both candidemia and deep-seated tissue candidiasis[26,27]. Invasive candidiasis can also be restricted to a single organ system or body compartment which results from local inoculation of *Candida.* This disease presents a mortality range of 40%, even when patients receive antifungal therapy[27].

Candidemia is the most common form of invasive candidiasis and represents the most significant and prevalent hospital fungal infection associated with a high mortality rate (up to 49%) in immunocompromised patients[25]. Ten to forty percent of candidemia cases are associated with sepsis or septic shock while *Candida* species are responsible for no more than 5% of the total number of cases as the main agent of sepsis or septic shock[25]. Deep-seated candidiasis results from both hematogenous dissemination or *Candida* species direct inoculation into a sterile site, such as the peritoneal cavity[27].

Candida species, as being an important cause of morbidity and mortality worldwide, represent a serious threat to public health[25]. Among them, *C. albicans* has shown to be the most prevalent *Candida* species and thus represents the most frequently species isolated from fungal infections[25]. However, other *Candida* species such as *C. glabrata, C. tropicalis*, *C. parapsilosis*, *C. krusei*, *C. famata, C. guilliermondii*, and *C. lusitaniae* have been increasingly isolated over time, due to the selection of less sensitive *Candida* strains by the widespread use of the azole fluconazole as a therapeutic agent[25,28].

C. albicans presents a high degree of metabolic and physiologic flexibility, which supports its ability to grow in extremely different environments. This fact, associated with the species' high resistance capacity to antifungals, their virulent features, such as capability of forming biofilms, spot this species as the most threatening to human health among *Candida* species[25].

1.7. Virulence factors

C. albicans has the ability to infect a diversity of host niches due to a wide range of virulence factors. Those include morphological transition between yeast and hyphal forms, cell surface expression of adhesins and invasins, thigmotropism, biofilms formation, phenotypic switching, and the secretion of hydrolytic enzymes (Figure 3). Some groups of virulence factors are responsible for colonization or the initiation of an infection, and the other groups helps the infection spreading[2,9,29].

1.7.1. Polymorphism

C. albicans has a polymorphic nature transitioning from a commensal form to a pathogenic one, which depends on a range of environmental cues[8,29]. This process is characterized by the morphological transition of blastopores into hyphae, termed as dimorphism[8,29]. While the hyphal form is the most invasive form, the smaller yeast form is believed to be the form primarily involved in dissemination[29].

1.7.2. Expression of adhesins and invasins on the cell surface

Adhesins are a specialized set of proteins with an important mediator role in the adhesion process to other *C. albicans* cells, to other microorganisms, to abiotic surfaces and to host cells[8,29].

C. albicans can infect the host cells through two different mechanisms: induced endocytosis and active penetration. In the induced endocytosis mechanism, *C. albicans* expresses invasins on the cell surface to facilitate binding to host ligands and hence promote endocytosis by host immune cells, while active penetration is a fungal-driven process[29].

1.7.3. Biofilm formation

A further important feature of *C. albicans* pathogenesis is its capacity to form biofilms on abiotic or biotic surfaces. Biofilms are form through a sequential process, which starts with the adherence of yeast cells to the substrate, continues with the proliferation of these yeast cells, the formation of hyphal cells in the biofilm upper part, the accumulation of extracellular matrix material and, lastly, yeast cells dispersion from the biofilm complex[29].

Based in *C. albicans* transition ability, its biofilm is a complex structure of different morphological forms[8].

1.7.4. Contact sensing and thigmotropism

In *C. albicans*, contact sensing is an important environmental signal that promotes hypha and biofilm formation. The contact with a surface, triggers the yeast cells switch to hyphal growth[29].

Thigmotropism, a directional hyphal growth, occurs on particular surfaces, for example with the presence of ridges. In *C. albicans*, thigmotropism is regulated by extracellular calcium uptake through the calcium channels[8,29].

1.7.5. Secreted hydrolases

C. albicans has the ability to secrete hydrolases, which occurs after the adhesion process to host cell surfaces and the hyphal growth. These enzymes facilitate active penetration into these cells and improve the efficiency of extracellular nutrient acquisition[29]. Moreover, hydrolases enable the invasion into the surfaces of mucous membranes and blood vessels, and also participate in avoiding the host's immune response[8].

There are three classes of secreted hydrolases expressed by *C. albicans*: proteases, phospholipases, and lipases. Among these, the main enzymes produced by *C. albicans* are SAP (secreted aspartyl protease), phospholipase, and hemolysin[8,29].

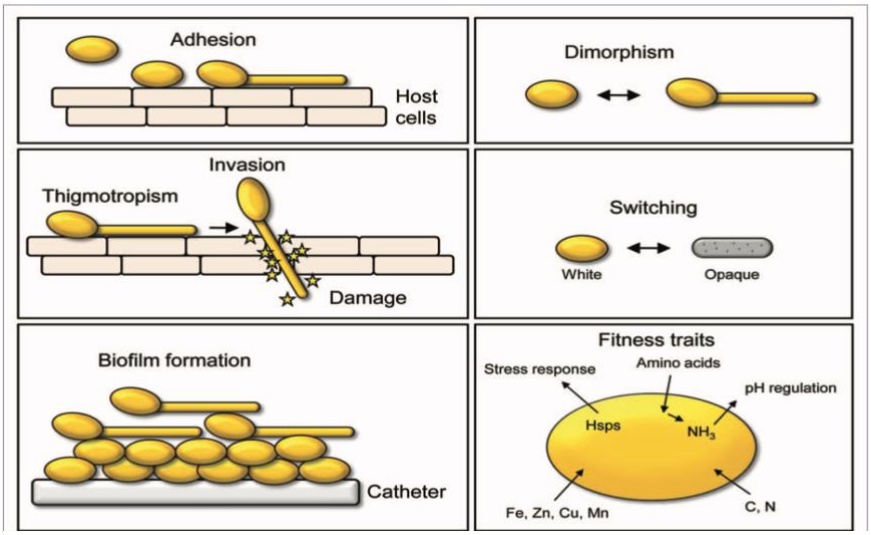


Figure 3: *C. albicans* virulence factors[29].

1.8. Treatment

Only a few classes of antifungal drugs are currently available to treat infections from *Candida* spp., these are: azoles, echinocandins, polyenes and nucleoside analogues[30]. Based on the European Society of Clinical Microbiology and Infectious Diseases, echinocandins is recommended as the first-line treatment for all patients with systemic candidiasis[28]. Figure 4 demonstrates the primary targets of each antifungal drug used in *C. albicans* treatment.

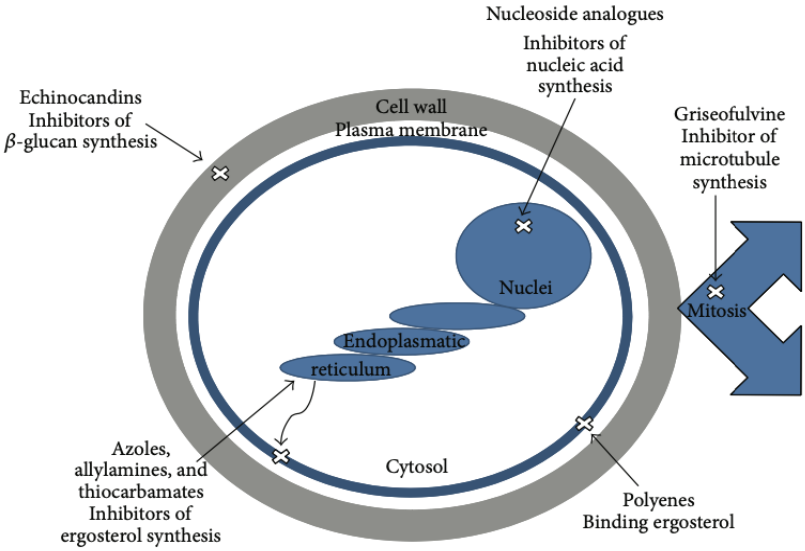


Figure 4: Primary targets of *C. albicans* antifungal drugs[30]**.**

1.8.1. Azoles

Azoles are represented by their lanosterol 14-α-demethylase activity inhibition mechanism, which leads to the disruption of the cell membrane[30]. Lanosterol 14-α-demethylase is an enzyme involved in the biosynthesis of ergosterol, the analogous to cholesterol in animal cells and the largest sterol component of the fungal membrane[30].

The azoles family is the largest antifungal drugs family, including imidazoles (miconazole, econazole, clotrimazole, and ketoconazole) and triazoles (fluconazole, itraconazole, voriconazole and posaconazole)[30].

1.8.2. Echinocandins

Echinocandins are a novel class of antifungal. They are lipopeptides modified from pneumocandins that non-competitively inhibit the 1,3-β-d-glucan synthase enzyme,

responsible for the biosynthesis of 1,3-β-d-glucan, a key component of the fungal cell wall. The cell wall disruption results in osmotic instability, lysis and death of the fungal cells[28,31,32].

This antifungal class includes caspofungin, micafungin, and anidulafungin[31].

1.8.3. Polyenes

The polyenes agents act trough binding ergosterol and disrupting the major lipidic component of the fungal cell membrane, causing the production of aqueous pores afterwards[30]. This process leads to an alteration into the cellular permeability which later results in the leakage of cytosolic components and, therefore, fungal death[30].

1.8.4. Nucleoside Analogues

The nucleoside analogues are represented as inhibitors of DNA/RNA synthesis. For example, flucytosine, a pyrimidine analogue, is transported into fungal cells and then deaminated to 5-fluorouracil and phosphorylated to 5-fluorodeoxyuridine monophosphate[30]. This product inhibits thymidylate synthase and thus interferes with DNA synthesis[30].

The 5-fluorodeoxyuridine monophosphate can also be further phosphorylated and combined into the RNA, affecting the RNA synthesis as well[30].

1.9. Antifungal resistance

The term resistance can be defined as a strain that has a minimum inhibitory concentration (MIC) for a particular antifungal above specific clinical breakpoints but, can also be used more broadly to indicate a strain with an increase in MIC to an antifungal drug relative to a control or reference strain[33].

Several adaptive mechanisms of antifungal drug resistance and tolerance have been identified, such as drug target alteration or overexpression, upregulation of multidrug transporters, and activation of cellular stress responses[33].

1.9.1. Azoles resistance

1.9.1.1. Overexpression of Efflux Pumps

The upregulation of plasma membrane efflux pumps is a major mechanism associated with azole resistance in many fungal pathogens, since it leads to drug concentrations decrease at the drug target in the fungal cell[28,33].

In *C. albicans*, efflux pumps are encoded by *Candida* drug resistance genes, *CDR1* and *CDR2*, of ATP-binding cassette (ABC) superfamily and *MDR1* gene of the major facilitator superfamily (MFS). The MDR-encoded efflux pumps are secondary transporter specific for fluconazole[28,30].

The transcription factor TAC1 (transcriptional activator of *CDR* genes) binds to cis-acting drug-response elements (DREs) in the promoters of *CDR1* and *CDR2* to regulate their expression, which links TAC1 to azole resistance. Moreover, the mediator complex was recently shown to be implicated in TAC1-mediated azole resistance, as the deletion of the mediator tail module in TAC1 gain-of-function mutants reduced *CDR1* transcription and increased fluconazole susceptibility[33].

95MF transporters are encoded in *C. albicans* genome and point mutations in *MRR1* increase azole resistance. Deleting *MRR1* promotes a higher fluconazole resistance reduction than the deletion of the MDR1 efflux pump itself, suggesting that *MRR1* may regulate additional determinants of azole resistance[33].

1.9.1.2. Alterations in drug targets

This common mechanism of azole resistance involves the alteration or up-regulation of the drug target lanosterol 14-α-demethylase, encoded by the *ERG11* gene[28,30,33].

The alteration mechanism is carried out through mutations in *ERG11* gene, which prevent the binding of azoles to the enzymatic site, leading to a decreased target affinity for the drug[28,30,33]. Over 140 amino acid substitutions in *ERG11* have been associated with azole resistance and the majority of these substitutions clustered into hot-spot regions ranging from 105−165, 266−287, and 405−488. It was revealed through the molecular mapping of *C. albicans ERG11* variant positions that mutations reside in the catalytic site of the enzyme, the fungus-specific external loop, and the proximal surface, as well as between the proximal surface and the heme[33].

The *ERG11* gene up-regulation mechanism promotes a intracellular increase of the target protein[30]. The overexpression of *ERG11* is very common and directly contributes to increase the target abundance, ultimately lowering the drug susceptibility[33]. However, a

minimal up-regulation of altered target enzymes has suggested so far that this mechanism plays a limited role in the development of resistance to the azoles[28].

1.9.1.3. Modulation of Stress Responses

This mechanism involves the modification of the ergosterol biosynthesis pathway. The exposure to azoles results in reduce ergosterol content in the fungal cell membrane and leads to the accumulation of the toxic product 14-α-methyl-3,6-diol[28].

Loss-of-function mutations in the *ERG3* gene, prevents the formation of 14-α-methyl-3,6-diol from 14-α-methylfecosterol and thus its cellular accumulation. Alternatively, 14-α-methyl fecosterol is incorporated into the fungal cell membrane, allowing them to continuously grow and replicate in the presence of azoles[28,33].

Heat shock protein 90 (Hsp90), the essential molecular chaperone, is considered as a global stress response regulator in diverse fungal pathogens and in *C. albicans*, Hsp90 potentiates the rapid evolution of azole resistance[33].

Another component that regulates cellular response to the azoles, is the *C. albicans* protein kinase C, PKC1, through a MAPK cascade consisting of Bck1, Mkk1/2, and Mkc1. Compromise of PKC-MAPK signaling phenocopies both Hsp90 and calcineurin inhibition, reducing azole-resistance phenotypes in distinct *C. albicans* clinical isolates. Moreover, genetic depletion of Hsp90 in *C. albicans* destabilizes Mkc1, Bck1, and Pkc1, consequently blocking PKC signaling[33].

1.9.1.4. Genomic modifications

Fungal genomes are highly plastic, which allows them to adapt to environmental perturbations and acquire antifungal resistance[33].

Karyotype variability conferring azole resistance has been extensively studied in *C. albicans* and the most prevalent phenomenon is a specific aneuploidy on the left arm of chromosome 5. This segmental aneuploidy consists of an isochromosome composed of two identical left arms of Chr5 flanking a centromere (i(5L)). Gain and loss of azole resistance is correlated to the gain and loss of i(5L). The resistance associated with i(5L) is mediated by increased copy numbers of ERG11 and TAC1, as well as gain-of-function mutations and LOH of these resistance determinants[33].

Aneuploidies of chromosomes 3,4 and 6 are implicated in *C. albicans* azole resistance. Experimental evolution in the presence of azoles and the calcineurin inhibitor

FK506 revealed several aneuploidies in strains that developed resistance to this drug combination. The most common aneuploidy in all resistant lineages was an increased copy number of chromosome 4, which suggests that this chromosome may have important resistance determinants. Furthermore, the experimental evolution in a strain sensitized to azoles, resulted in an amplification of both chromosomes 7 and a large segment of chromosome 3, which accompanied suppression of the azole-sensitivity phenotype. The overexpression of a transporter gene on the affected region of chromosome 3 was sufficient to confer the suppression phenotype[33].

LOH also has a significant impact on drug resistance and is most common in genomic regions containing determinants of azole susceptibility. In *C. albicans* LOH for the transcription factors TAC1 and MRR1 have been reported in azole resistant isolates[33]**.**

1.9.2. Echinocandin resistance

1.9.2.1. Alterations in Drug Targets

C. albicans acquires echinocandin resistance through point and intrinsic mutations in the essential gene, *FKS1*, the gene that encodes FKS subunits of (1,3)-β-D-glucan synthase[28,30,33]. These mutations have been found to cluster around two highly conserved regions, the hot-spot regions[28]. In *C. albicans*, the hot-spot regions correspond to amino acids 641−649 (hot-spot 1) and amino acids 1357−1364 (hot-spot 2). Mutations in these regions have been shown to substantially decrease the (1,3)-β-D-glucan synthase sensitivity, elevate MIC values, and result in cross-resistance between echinocandins. In *C. albicans* the serine 645 (S645) present in the hot-spot region 1 reveals the highest frequency of substitution and is associated with the most resistant phenotype[28,33].

Less is known about the role of *FK2* and *FK3* genes in the echinocandin resistance in *C. albicans*, however it has been shown that their deletion results in higher *FKS1* transcript levels and lower echinocandin susceptibility[33].

1.9.2.2. Modulation of Stress Responses

In *C. albicans*, the inhibition of (1,3)-β-D-glucan synthesis promotes an increase in chitin synthesis as a response. This response is mediated by protein kinase C, high-osmolarity glycerol response and Ca^{2+} calcineurin signaling pathways, which have been shown to contradict the lethal effects of the echinocandins[28].

Similar to the Azoles, Hsp90 also has a critical role regulating echinocandins resistance in *C. albicans*. Besides Hsp90, the homozygous deletion of the transcription factor CAS5 reduces *FKS1*-mediated echinocandin resistance[33].

1.9.2.3. Genomic modifications

The emergence of echinocandin resistance from chromosomal abnormalities is uncommon compared to azole resistance in *C. albicans*. Nevertheless, genomic alterations have been reported in echinocandin resistant isolates. An identified mechanism of echinocandin resistance in *C. albicans* is the homozygous hot-spot mutations in *FKS1*, followed by LOH[33].

1.9.3. Polyenes resistance

1.9.3.1. Alterations in Drug Targets

Polyenes resistance is extremely uncommon given the fitness cost associated with the process, however, when it does occur, is due to changes in enzymes that reduce drug-binding affinity or deplete ergosterol from the membrane[33]. Membranes of polyene-resistant *Candida* isolates show a relatively low ergosterol content, due to mutations in the *ERG3* or *ERG6* genes which encode some of the enzymes involved in ergosterol biosynthesis[30].

Candida albicans mutations in ergosterol biosynthesis enzymes, such as ERG2, ERG3, ERG5 and ERG6 promote the amphotericin B susceptibility reduction[33]**.**

1.9.3.2. Modulation of Stress Responses

Clinical isolates of *C. albicans* with loss-of-function mutations in *ERG3* have demonstrated cross-resistance to the polyenes and azoles[33].

The survival of amphotericin B-resistant isolates is dependent upon Hsp90 expression and function. Therefore, the inhibition of Hsp90 eliminated amphotericin B resistance[33].

2. Genomics and epidemiologic surveillance

Whole-genome sequencing (WGS) allows for the determination of the entire nucleotide sequence of a genome. Based on these terms, genomic studies using WGS, allow an analysis of not just coding genes and, the readily identification of a full range of common and rare structural variants, including deletions, amplifications, chromosomal translocations, and LOH involved in either pathogenicity or drug resistance[34,35]. Microbial WGS data further allows to elucidate phylogenetic relationships among isolates belonging to disease-causing lineages, which enhances their traceability and monitoring over time[36]. Besides that, WGS data can be used in multiple secondary analyses, such as virulence gene detection, antibiotic resistance gene profiling, synteny comparisons, mobile genetic element identification and geographic attribution[36].

In this study, short-read Illumina WGS, which yields paired-end reads of approximately 150 bp with low error rates in the range of about 0,1% to 0,5%[35] was used. The Illumina DNA Prep Kit integrates DNA extraction, fragmentation, library preparation and library normalization steps, which enables the fastest library preparation workflow. Furthermore, this kit offers On-Bead Tagmentation, a strategy that uses bead-bound transposomes to mediate a more uniform tagmentation reaction. The bead-bound transposomes get saturated with DNA, making impossible additional tagmentation, enabling a highly uniform saturation-based normalization process[37].

Multilocus sequence typing (MLST) is a DNA sequence-based typing method that involves the determination of the alleles of up to seven well-conserved housekeeping genes using nucleotide sequences of internal fragments of these genes. The different sequences of each housekeeping gene present within a microbial species are assigned different allele numbers and, for each isolate, the combination of the alleles (allelic profile) at each of the seven loci defines their sequence type (ST)[38,39]. Two isolates showing different STs are necessarily different strains[40]. Two isolates with the same ST, belong to strains not distinguishable by MLST, however they are not necessarily identical strains, since this methodology gives no information about the global extent of SNP differences between the two sequences[40]. MLST uses housekeeping genes because they code for proteins under stabilizing selection for the conservation of metabolic function, and they are sufficiently diverse to identify multiple variants within the isolate collection[38]. The accumulation of nucleotide changes in housekeeping genes is a relatively slow process and they are sufficiently stable over time, promoting an universal nomenclature MLST scheme for storing and analyzing the nucleotide sequence data to infer evolutionary relationships[38,39]. The MLST scheme database for *C. albicans* hosted in PubMLST (Public databases for

molecular typing and microbial genome diversity) is based on fragments of seven *C. albicans* genes: *AAT1A*, *ACC1*, *ADP1*, *MPIB*, *SYA1*, *VPS13*, and *ZWF1B*[41].

MLST's main goal is to compare the strains under study to others of other parts of the world that are present in the PubMLST database, which means that MLST provides important information for population genetics and epidemiological studies[40]. From the different allele's identification, information such as GC content, codon usage, and polymorphism frequencies that show the different nucleotide changes present within an isolates collection, can be later obtained. This enables a retrospective and perspective analysis of data within a specific strain collection, country, or even globally. Comparisons and distinctions can be made among the data or against other data that are available[38]. In this study, MLST was used to understand if strains from a specific ecological niche shared the same ST with other strains from the same niche or another, in this collection and with the PubMLST strain collection.

Gene Ontology (GO) analysis identifies GO terms implicated in a list of genes differentially expressed (emerge of single nucleotide polymorphisms (SNPs)), for example in case of a disease. There are three types of terms in the gene ontology, such as biological processes, molecular functions, and cellular components[42,43]. GO analysis includes GO Slim Mapper and GO Term Finder. The GO Slim Mapper maps annotations, associations between gene products and GO terms, of a group of genes to more general terms and/or bins them into broad categories, while the GO Term Finder explores significant shared GO terms, used to describe the genes of our list to help to discover what the genes may have in common[42,43]. In this study, GO analysis was used in order to study the genomic variability of the isolates in study, using as reference the *C. albicans* SC5314 genome.

3. Scope, aims and objectives

Genomic studies are essential in the investigation of the biology, ecology, phylogeny, and epidemiology of pathogens, allowing the surveillance of microorganisms worldwide. While bacterial genomics has been widely used, only recently, fungal genomics is becoming part of clinical and public health microbiological laboratory practices.

C. albicans is a commensal yeast that belongs to the human normal microbiota. However, it has the ability to become an opportunistic organism and represents one of the leading causes of human fungal infections, mainly superficial vaginal or mucosal oral but also systemic, almost always deadly infections. Based on this problem, it is imperative to routinely characterize this pathogen, not only phenotypically but also genotypically, to

identify the genomic traits responsible for virulence, drug resistance, adaptability to ecological niches, and unravel the genetic diversity among distinct isolates. It is also important to implement pipelines for the rapid analysis of genomic data derived from these organisms.

In this study, a collection of *C. albicans* strains was characterized by WGS and bioinformatics analysis with the purpose of describing their genome sequence, discriminate and epidemiologically contextualize strains globally, identify genes with single nucleotide polymorphisms (SNPs) and contribute to the knowledge on the *C. albicans* variome by adding new data obtained to international databases.

This work will add knowledge on this pathogen, which can help the understanding of *C. albicans* pathogenicity and lead to higher effectiveness of therapeutics.

3.1. Objectives

The general aim of this dissertation was to study the genetic diversity of 76 clinical strains previously identified as *C. albicans* isolates. For this, this dissertation specific objectives were, to:

1. Identify Single Nucleotide Polymorphisms (SNPs) using *C. albicans* SC5314 reference genome.
2. Discriminate and epidemiologically contextualize *C. albicans* strains globally, through the identification of multilocus sequence types.
3. Identify genes with SNPs (i) common to all isolates; (ii) exclusive of all isolates collected from each sample type; and (iii) common to all isolates collected from each sample type.
4. Contribute to the knowledge on the *C. albicans* variome by adding new data obtained to international databases.

II. MATERIALS AND METHODS

1. Strains and growth conditions

C. albicans clinical isolates were provided by Prof. Teresa Gonçalves, from the Instituto de Microbiologia da Faculdade de Medicina da Universidade de Coimbra (FMUC). The 76 clinical strains were collected from vaginal (n=29), oral (n=13), and blood samples (n=20) and samples collected from medical devices (n=13; Table 1).

Candida sp. isolates were grown at 30ºC in YPD (Yeast Extract-Peptone-Dextrose) with 2% agar, for 96 hours. Strains were pre-inoculated in YPD liquid medium at 30ºC, overnight. Subsequently, 400 μL of the pre-inoculum were used to inoculated with 20 mL of YPD liquid medium, and the optical density (OD) was measured with the spectrophotometer Microplate Manager v6.3 until the OD values reached 1, indicating the log growth phase.

Table 1: *C. albicans* strains used in the study.

Isolates	Identification	Origin
YP0045	*Candida albicans*	Blood
YP0047	*Candida albicans*	Blood
YP0569	*Candida albicans*	Blood
YP0399	*Candida albicans*	Blood
YP0577	*Candida albicans*	Blood
YP0639	*Candida albicans*	Blood
YP1130	*Candida albicans*	Blood
YP0760	*Candida albicans*	Blood
YP0581	*Candida albicans*	Blood
YP0801	*Candida albicans*	Blood
YP0631	*Candida albicans*	Blood
YP0037	*Candida albicans*	Blood
YP0057	*Candida albicans*	Blood
YP0364	*Candida albicans*	Blood
YP0070	*Candida albicans*	Blood
YP0537	*Candida albicans*	Blood
YP0363	*Candida albicans*	Blood
YP0493	*Candida albicans*	Blood
YP0474	*Candida albicans*	Blood
YP0126	*Candida albicans*	Blood
YP0392	*Candida albicans*	Vaginal
YP0083	*Candida albicans*	Vaginal
YP0050	*Candida albicans*	Vaginal

YP0061	*Candida albicans*	Vaginal
YP0087	*Candida albicans*	Vaginal
YP0058	*Candida albicans*	Vaginal
YP0108	*Candida albicans*	Vaginal
YP0098	*Candida albicans*	Vaginal
YP0159	*Candida albicans*	Vaginal
YP0131	*Candida albicans*	Vaginal
YP0129	*Candida albicans*	Vaginal
YP0326	*Candida albicans*	Vaginal
YP0144	*Candida albicans*	Vaginal
YP0051	*Candida albicans*	Vaginal
YP0132	*Candida albicans*	Vaginal
YP0093	*Candida albicans*	Vaginal
YP0081	*Candida albicans*	Vaginal
YP0355	*Candida albicans*	Vaginal
YP0167	*Candida albicans*	Vaginal
YP0232	*Candida albicans*	Vaginal
YP0233	*Candida albicans*	Vaginal
YP0344	*Candida albicans*	Vaginal
YP0316	*Candida albicans*	Vaginal
YP0362	*Candida albicans*	Vaginal
YP0094	*Candida albicans*	Vaginal
YP0095	*Candida albicans*	Vaginal
YP0200	*Candida albicans*	Vaginal
YP0097	*Candida albicans*	Vaginal
YP0114	*Candida albicans*	Vaginal
YP0115	*Candida albicans*	Vaginal
YP0001	*Candida albicans*	Medical Devices
YP0048	*Candida albicans*	Medical Devices
YP0067	*Candida albicans*	Medical Devices
YP0162	*Candida albicans*	Medical Devices
YP0176	*Candida albicans*	Medical Devices
YP0188	*Candida albicans*	Medical Devices
YP0196	*Candida albicans*	Medical Devices
YP0211	*Candida albicans*	Medical Devices
YP0306	*Candida albicans*	Medical Devices
YP0382	*Candida albicans*	Medical Devices
YP0384	*Candida albicans*	Medical Devices
YP0386	*Candida albicans*	Medical Devices
YP0391	*Candida albicans*	Medical Devices
YP0016	*Candida albicans*	Oral

YP0019	*Candida albicans*	Oral
YP0024	*Candida albicans*	Oral
YP0026	*Candida albicans*	Oral
YP0028	*Candida albicans*	Oral
YP0031	*Candida albicans*	Oral
YP0034	*Candida albicans*	Oral
YP0035	*Candida albicans*	Oral
YP0036	*Candida albicans*	Oral
YP0054	*Candida albicans*	Oral
YP0100	*Candida albicans*	Oral
YP0101	*Candida albicans*	Oral
YP0103	*Candida albicans*	Oral

2. Whole Genome Sequencing

2.1. DNA extraction

All isolates DNA was extracted using the Lucigen's MasterPure Yeast DNA Purification Kit with an adapted protocol used at iBiMED. A 20 mL cell culture at mid-log phase (OD: 1-1.5), was centrifuged at 4000 rpm, 4ºC, for 10 minutes and the obtained pellet washed with 4ml of a TE buffer (10 mM Tris-HCL, pH 8.0; 1mM EDTA, pH 8.0). With the aim of cell lysis, 300 μL of a Yeast Cell Lysis Solution (previously mixed) and 10 μL of RNase A (Qiagen, 100 μg/μL) were used and then incubated for 2-3 hours at 50ºC. The suspension was incubated in ice for 5 minutes and 150 μL of MPC Protein Precipitation Reagent was added and the mixture vortexed for 10 seconds. The suspension was then centrifuged for 10 minutes, 13,000 rpm, at 4ºC. The supernatant was transferred to a clean tube and the process repeated with a reduced centrifugation time of 3 minutes. 500 μL of isopropanol at room temperature were added and gently mixed by inversion for 30-40 times. The suspension was centrifuged for 10 minutes, at 13, 000 rpm and 4ºC, and the isopropanol removed without dislodging the pellet. The pellet was then washed with 500 μL ethanol 70% (freshly prepared) and centrifuged for 1 minute at 13,000 rpm. The ethanol was removed, the last two steps were repeated, and the pellet was allowed to dry for no more than 5 minutes at room temperature. The pellet was then suspended in 23 μL of 1X TNE (Tris-NaCl-EDTA) Buffer and left overnight at 4ºC. After this, for RNA clean-up, 1 μL of RNase A and RNase I (4 U/μL) were added to the suspension and incubated overnight at 37ºC. After incubation, 25 μL of Yeast Cell Lysis Solution (1:1 proportion) were added and incubated on ice for 5 minutes. 50 μL of MPC Protein Precipitation Reagent were then added, mixed until becoming homogeneous and centrifuged for 10 minutes, at 13,000 rpm

and 4ºC. The supernatant was transferred to a clean tube, 100 μL of isopropanol at room temperature were added and gently mixed by inversion for 30-40 times. The suspension was centrifuged for 10 minutes at 13,000 and 4ºC. Without dislodging the pellet, the isopropanol was removed, and the pellet washed with 100 μL of ethanol twice with 1-minute centrifugations. The ethanol was then removed, the pellet was allowed to dry for 1-2 minutes and the DNA was resuspended with 35 μL of ultrapure H_2O and stored at 4ºC overnight.

2.2. DNA quantification and quality assessment

DNA quality was measured by assessing A260:280 and A260:230 ratios using the Denovix DS-11 Spectrophotometer and by running the DNA samples in a 1% agarose gel. The DNA quantification was performed using the Qubit 2.0 Fluorometer Assay.

2.3. Genomic DNA libraries preparation

Genomic libraries were prepared using the Illumina DNA Prep Kit protocol (Figure 6).

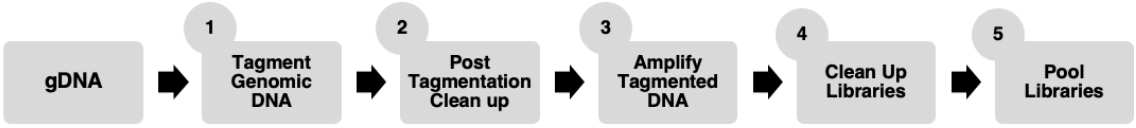


Figure 5: Illumina DNA Prep Workflow.

After an initial DNA dilution with ultrapure H_20 to reach a total input for sequencing of 100-500 ng, the protocol started with the Tagment Genomic DNA step, which enzymatically fragments and tags the DNA with adapter sequences. To do so, the previously prepared tagmentation master mix with BLT (Bead-Linked Transposomes) and TB1 (Tagmentation Buffer 1), were added to the samples and incubated in the thermal cycler for 15 minutes at 55ºC. Afterwards, during the Post Tagmentation Cleanup step, the adapter-tagged DNA was washed on the BLT before PCR amplification. Firstly, TSB (Tagmentation Stop Buffer) was added to the tagmentation reaction and samples incubated for another 15 minutes at 37ºC. After that, the samples were transferred to a 96-well-plate, placed on a magnetic stand and the supernatant discarded. The pellet was washed two times with TWB (Tagment Wash Buffer), and the beads were resuspended in TWB. The next step, Amplification of the Tagmented DNA, started with the supernatant removal,

followed by the addition of a PCR master mix with EPM (Enhanced PCR Mix) and nuclease-free water, previously prepared. The EPM includes a master mix and index adapters. The samples were then incubated in the thermal cycler with the BLT PCR program. After the PCR cycles, the Clean Up Libraries step, that purifies the amplified libraries started. Initially, the supernatant was transferred to a new plate, using the magnetic stand. For standard DNA input, nuclease-free water and SPB (Sample Purification Beads) were added to each well containing supernatant, and this solution was then transferred to a new plate containing undiluted SPB. Without disturbing the beads, the supernatant was removed, and samples were washed twice with 80% ethanol. After the 80% ethanol removal, beads were resuspended in RSB (Resuspension Buffer), and the supernatant was transferred to a new plate. Finally, the Pool Libraries step, where the DNA libraries quantity and quality are verified, was performed. Libraries were quantified using the Qubit fluorometric assay and samples were ran on the Agilent 2100 Bioanalyser to obtain the library size profile. Each library was prepared with an insert size of ~ 600 bp (Figure 5).

2.4. Sequencing

The sequencing run was performed by using the NextSeq 550 System by Illumina following the manufacturer's guide protocol. Before loading, samples were diluted to the starting concentration of 2nM. After dilution, libraries were denatured and diluted to the final loading concentration. Libraries were then loaded in the reagent cartridge and after checking the run and system status, the run was started.

3. Bioinformatic analysis

Quality control of paired-end reads was performed before downstream analysis. Trimmomatic v0.36[44] was used to (i) exclude reads with length below 33 bp, (ii) trim the first 10 bp and the last 3 bp of each read if Phred score <3, and (iii) to scan reads with a 5 bp wide sliding window, cutting when the average quality per base dropped below 20 (Phred score). Quality control reports were generated using fastQC v0.11.7[45] before and after trimming, and aggregated using MultiQC v1.12[46].

Genome coverage was calculated with an in-house script after reads were mapped against the genome of *Candida albicans* SC5314 (Candida Genome Database Assembly 22, haplotype A).

3.1. Multilocus Sequence Typing (MLST)

Using *C. albicans* MLST scheme published in PubMLST, the assembled genomes of each sample were compared to the reference database to identify allelic matches and a sequence type. Results were ordered by best match. With the lack of an exact match for all genes defined in the MLST gene set, the locus/scheme was selected using the closest match for the gene(s) lacking.

3.2. Genomic variant analysis

Filtered reads were mapped against the genome of *Candida albicans* SC5314 (Candida Genome Database Assembly 22, haplotype A) using bwa v0.7.17[47]. BAM files containing all mapped reads were sorted and PCR duplicates filtered using SAMtools v1.16.1[48] and Picard tools v2.21.3[49]. Base quality score recalibration was performed with GATK v4.2.6.1[50]. GATK tools haplotype caller, select variants and variant filtration, were then used for single nucleotide polymorphism (SNP) calling against the *Candida albicans* reference genome mentioned above. Low-quality variant calls were filtered out with parameters "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0", according to the GATK recommendations for hard-filtering germline short variants as SNPs[51]. SnpEff v5.0 was subsequently used to annotate SNPs based on their genomic location (intronic, untranslated region, upstream, downstream, splice site, or intergenic regions) and predict their effects (synonymous or non-synonymous amino acid replacement, start codon gains or losses, or frame shifts)[52]. Additionally, due to the large number of SNPs and genome heterozygosity in these samples (an expected feature in *Candida albicans* genomes), SNPs with >2 reads supporting an alternative allele for coverage 10× to 20× and >4 reads supporting an alternative allele for coverage more than 20× were excluded, as in Selmecki *et al.* 2015[53].

3.3. Gene Ontology (GO) analysis

To understand how the genes with SNPs present in the clinical isolates' collection in study distributed according to the different sample origins, a Venn diagram was built using the package UpSetR v1.4.0R v4.1.3 in RStudio v2022.02.1 Build 461. Gene sets obtained for further analyses were (i) genes with SNPs common to all isolates (clinical/clinical environment vs laboratory strain), (ii) genes with SNPs exclusive of isolates of each sample type (blood vs vaginal vs oral vs medical devices), and (iii) genes with SNPs common to all

isolates of each sample type (blood or vaginal or oral or medical devices vs laboratory strain). To identify the roles of these genes in biological processes, molecular functions, and cellular components, a gene ontology (GO) analysis was performed using the GO Slim mapper and the GO Term finder tools available through the Candida Genome Database[42]. The first was used to annotate the genes according to each of the GO sets in broader groups, and the second tool was used to find significantly shared GO terms in each GO set to describe the genes where SNPs were found (p-value ≤ 0.05).

III. RESULTS

1. Whole Genome Sequencing

1.1. Genomic DNA concentration

Genomic DNA quantification of each isolate revealed DNA concentrations ranging from 17.2 ng/μl to 30.2 ng/μl. Final library concentrations which were loaded for sequencing ranged from 11 ng/μl to 20.7 ng/μl. Annex A shows the DNA concentrations per isolate obtained in both quantification steps.

1.2. Sequencing run metrics

A total of 751154062 reads were obtained, ranging from 8667680 to 11332668 reads in samples YP0035 and YP0097, respectively. After quality checks to improve data quality, the number of total reads decreased to a range from 8401490 reads (sample YP0035) to 10985684 reads (sample YP0097).

Reads mapping to the reference genome ranged between 98.4% (sample YP0233) and 99.6% (sample YP0051), and weighted mean whole genome coverage ranged between 80.9% (sample YP0057) and 102.8% (sample YP0097), both very good results. GC content of the isolates was of 33%, 33.5% and 34% as expected.

However, the results obtained for isolate YP0129 were discrepant comparing to all others, as only 5.2% of reads mapped against the SC5314 reference genome, the weighted mean whole genome coverage was of 37% and the GC content of this isolate was of 37%, suggesting this isolate could have been previously misidentified as *C. albicans*.

Table 2 and annex B depict the sequencing run metrics per sample.

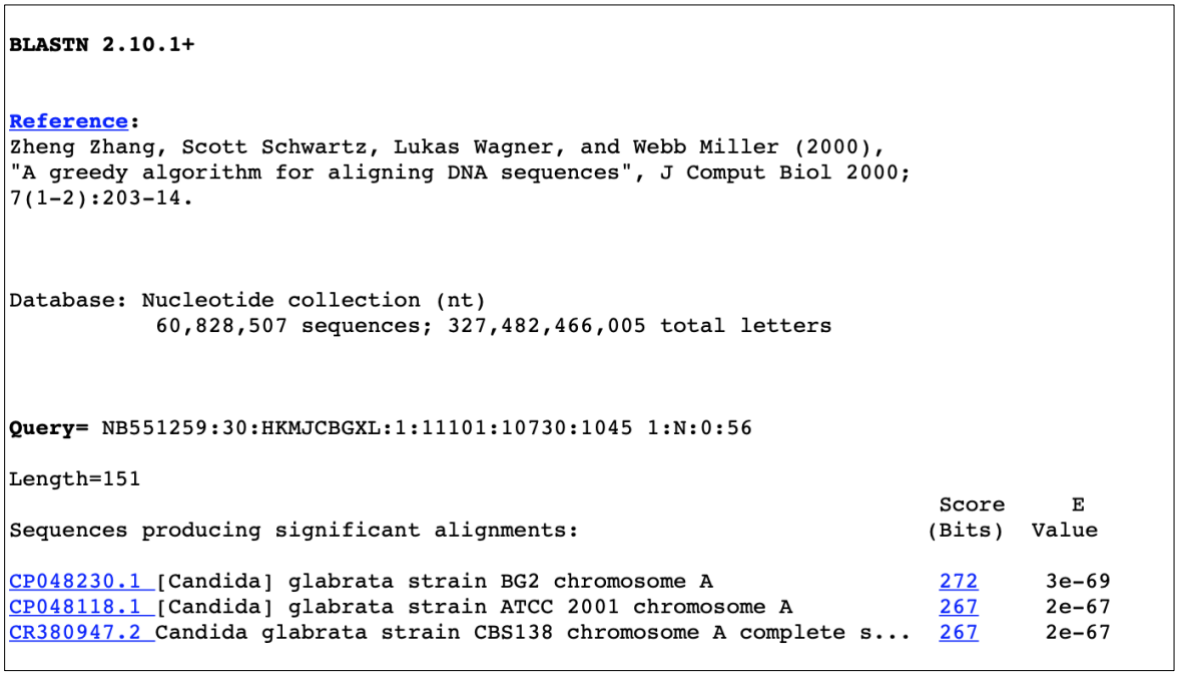
Table 2: GC content, weighted mean of whole genome coverage, and frequency of reads mapped obtained for all isolates.

Sample	Origin	%GC	Whole genome coverage, weighted mean (%)	Reads mapped (%)
YP0001	Medical Devices	33,5%	89,4	98,8
YP0016	Oral	34,0%	82,7	98,6
YP0019	Oral	33,5%	83,6	98,6
YP0024	Oral	34,0%	82,0	99,1
YP0026	Oral	34,0%	89,9	98,9
YP0028	Oral	33,5%	91,2	99,0
YP0031	Oral	33,5%	82,9	99,6
YP0034	Oral	34,0%	90,4	99,1
YP0035	Oral	34,0%	78,3	99,0
YP0036	Oral	33,5%	94,0	99,1
YP0037	Blood	33,5%	83,9	98,9
YP0045	Blood	34,0%	87,8	99,0
YP0047	Blood	34,0%	91,2	99,0
YP0048	Medical Devices	33,5%	88,5	98,9
YP0050	Vaginal	33,0%	88,8	98,8
YP0051	Vaginal	33,5%	88,3	99,6
YP0054	Oral	34,0%	97,5	99,0
YP0057	Blood	34,0%	80,9	98,9
YP0058	Vaginal	33,5%	89,0	99,6
YP0061	Vaginal	33,5%	93,0	99,0
YP0067	Medical Devices	34,0%	91,7	99,5
YP0070	Blood	34,0%	84,1	98,9
YP0081	Vaginal	34,0%	90,8	99,0
YP0083	Vaginal	33,0%	92,7	98,9
YP0087	Vaginal	33,0%	81,9	99,1
YP0093	Vaginal	34,0%	84,4	98,4
YP0094	Vaginal	33,0%	94,4	99,3
YP0095	Vaginal	33,5%	89,4	99,3
YP0097	Vaginal	33,5%	102,8	99,3
YP0098	Vaginal	34,0%	95,9	98,8
YP0100	Oral	34,0%	84,6	99,5
YP0101	Oral	33,5%	93,8	99,1
YP0103	Oral	34,0%	89,5	98,9
YP0108	Vaginal	34,0%	82,8	98,9
YP0114	Vaginal	33,5%	90,1	99,1
YP0115	Vaginal	33,5%	88,6	99,1
YP0126	Blood	33,5%	96,9	99,5

YP0129	Vaginal	37,0%	1,8	5,2
YP0131	Vaginal	33,5%	92,2	98,8
YP0132	Vaginal	34,0%	94,6	99,0
YP0144	Vaginal	33,0%	91,2	99,5
YP0159	Vaginal	34,0%	96,0	98,9
YP0162	Medical Devices	34,0%	93,5	98,9
YP0167	Vaginal	33,0%	85,3	99,1
YP0176	Medical Devices	33,5%	92,9	98,8
YP0188	Medical Devices	34,0%	95,9	98,7
YP0196	Medical Devices	34,0%	81,1	98,7
YP0200	Vaginal	33,0%	94,4	99,4
YP0211	Medical Devices	34,0%	96,3	98,9
YP0232	Vaginal	33,0%	90,9	99,5
YP0233	Vaginal	34,0%	91,4	98,4
YP0306	Medical Devices	34,0%	98,0	99,5
YP0316	Vaginal	33,5%	88,4	98,6
YP0326	Vaginal	33,0%	83,1	98,8
YP0344	Vaginal	33,5%	97,7	98,9
YP0355	Vaginal	33,5%	82,0	99,0
YP0362	Vaginal	33,0%	91,5	99,5
YP0363	Blood	34,0%	91,6	98,7
YP0364	Blood	33,0%	87,0	99,0
YP0382	Medical Devices	34,0%	82,6	99,0
YP0384	Medical Devices	34,0%	81,7	99,1
YP0386	Medical Devices	34,0%	84,2	99,1
YP0391	Medical Devices	34,0%	83,5	98,9
YP0392	Vaginal	33,0%	83,2	99,1
YP0399	Blood	34,0%	89,5	98,9
YP0474	Blood	33,5%	94,7	98,7
YP0493	Blood	33,5%	97,9	98,8
YP0537	Blood	33,5%	90,3	99,0
YP0569	Blood	34,0%	88,9	99,0
YP0577	Blood	34,0%	89,0	99,1
YP0581	Blood	33,0%	85,8	99,0
YP0631	Blood	33,0%	83,3	98,9
YP0639	Blood	33,5%	87,4	98,9
YP0760	Blood	34,0%	87,7	98,8
YP0801	Blood	33,5%	89,3	98,9
YP1130	Blood	34,0%	85,4	98,5

1.3. Species identification of YP0129

To ascertain YP0129 species identification, we performed a blastn search against the NCBI nucleotide database of this isolate genome (2002.05.16), which resulted in a nucleotide sequence similarity with the *Candida glabrata* genome.



```
BLASTN 2.10.1+


Reference:
Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000),
"A greedy algorithm for aligning DNA sequences", J Comput Biol 2000;
7(1-2):203-14.



Database: Nucleotide collection (nt)
           60,828,507 sequences; 327,482,466,005 total letters



Query= NB551259:30:HKMJCBGXL:1:11101:10730:1045 1:N:0:56

Length=151
                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

CP048230.1 [Candida] glabrata strain BG2 chromosome A             272     3e-69
CP048118.1 [Candida] glabrata strain ATCC 2001 chromosome A       267     2e-67
CR380947.2 Candida glabrata strain CBS138 chromosome A complete s...  267     2e-67
```

Figure 6: Blastn search results against the NCBI nucleotide database of YP0129 genome.

Sequencing reads obtained for this isolate were then mapped against the *C. glabrata* CBS138 genome (chromosome A) available through the Candida Genome Database[54], resulting in 98.7% of mapped reads and a whole genome weighted mean coverage of 104%, , which confirmed its previous misidentification as *C. albicans* and caused its elimination from further analyses.

2. Multilocus Sequence Typing (MLST)

The results of the MLST for *C. albicans* revealed that, for some loci, these isolates' alleles were not available in the PubMLST database and an assignment to a sequence type (ST) was not possible for any sample. Allelic profiles of isolates YP0031, YP0050, YP0051, YP0058, YP0067, YP0100, YP0126, YP0144, YP0200, YP0232, YP0306, YP0362 and YP0363 matched several ST profiles previously described. Table 2 shows the exact and partial matches to the alleles previously known, the clonal complex (CC) and STs nearest matches for each sample. Due to the lack of a definitive allocation to a ST for all samples, it was not possible to determine to which clade each isolate belonged to.

Nucleotide sequences of the seven loci present in the isolates in study were submitted to the *Candida albicans* PubMLST database for the definitive assignment of STs and await curation.

Table 3: MLST results.

Isolates	AAT1a	ACC1	ADP1	MPIb	SYA1	VPS13	ZWF1b	ST	CC	Nearest match ST
YP0001	4	7	43	5	108*	165*	11*	ND	ND	
YP0016	4	7	43	5	108*	165	11*	ND	ND	
YP0019	148*	3	10*	14	76*	32	107	ND	ND	
YP0024	14	7	30	4	34	3	159*	ND	ND	
YP0026	148*	3	10	14	76*	32	107	ND	ND	
YP0028	33	7	30	4	2	3	159*	ND	ND	
YP0031	3	2	6	4	2	20	20	ND	ND	96; 201; 940; 941
YP0034	118*	58	10	5	155*	79*	268*	ND	ND	
YP0035	148*	58	10	5	155*	79*	268*	ND	ND	
YP0036	8	3	39	19	155*	32	22	ND	ND	
YP0037	4	7	43	5	108*	165*	11*	ND	ND	
YP0045	13	2	39	19	155*	32	22	ND	ND	
YP0047	13	2	39	19	155*	32	22	ND	ND	
YP0048	13	3	39	19	155*	70	22	ND	ND	
YP0050	13	13	10	19	34	55	20	ND	ND	1483
YP0051	3	3	6	4	2	20	20	ND	ND	1527; 2055
YP0054	3	58	10	5	155*	4	268*	ND	ND	
YP0057	148*	58	10	5	155*	20	268*	ND	ND	
YP0058	3	2	6	9	2	20	20	ND	ND	2085
YP0061	14	7	30	4	34	13	159*	ND	ND	

YP0067	3	3	6	9	2	20	20	ND	ND	127; 1089; 1411; 1450; 1522; 1527
YP0070	4	7	43	5	108*	165*	11*	ND	ND	
YP0081	14	7	30	4	34	13	159*	ND	ND	
YP0083	106*	40	6	9	24	167*	161	ND	ND	
YP0087	148*	58	10	4	155*	79*	268*	ND	ND	
YP0093		3	10	152*	67	205*	22	ND	ND	
YP0094	148*	58	10	5	155*	79*	20	ND	ND	
YP0095	148*	58	10	5	155*	79*	20	ND	ND	
YP0097	36	83*	14	63*	108*	4	268*	ND	ND	
YP0098	4	8	6	4	34	13	130*	ND	ND	
YP0100	3	2	6	4	2	20	20	ND	ND	96; 201; 381; 940; 941; 1524
YP0101	13	2	39	19	155*	32	22	ND	ND	
YP0103	13	43*	10	19	34	55	22	ND	ND	
YP0108	3	3	10	14	76*	32	107	ND	ND	
YP0114	14	7	30	4	34	13	159*	ND	ND	
YP0115	148*	58	10	5	155*	79*	268*	ND	ND	
YP0126	3	2	6	4	2	20	20	ND	ND	96; 201; 381; 940; 941; 1524;
YP0131	4	7	43	5	108*	51*	11*	ND	ND	
YP0132	3	3	10	14	76*	32	107	ND	ND	
YP0144	3	2	6	4	2	20	20	ND	ND	96; 201; 381; 940; 941; 1524;
YP0159	148*	58	10	5	155*	79*	268*	ND	ND	
YP0162	107	7	14	14	2	93*	107*	ND	ND	
YP0167	14	7	30	4	2	3	159*	ND	ND	
YP0176	107	7	14	14	2	93*	107*	ND	ND	
YP0188	107	7	14	14	2	93*	107*	ND	ND	
YP0196	148*	3	10	14	76*	32	107	ND	ND	
YP0200	3	2	6	9	2	20	20	ND	ND	96; 201; 381; 940; 941; 1524
YP0211	148*	58	10	5	155*	79*	268*	ND	ND	

YP0232	3	2	6	4	2	20	20	ND	ND	96; 201; 381; 940; 941; 1524
YP0233	113	13	6	4	155*	32	264*	ND	ND	
YP0306	3	3	6	4	2	20	12	ND	ND	1450; 1522; 1527; 3238
YP0316	4	7	43	5	97	165*	11*	ND	ND	
YP0326	14	7	30	4	34	3	159*	ND	ND	
YP0344	107	3	10	14	76*	32	107	ND	ND	
YP0355	107	58	10	5	155*	79*	268*	ND	ND	
YP0362	3	2	6	4	2	20	20	ND	ND	96; 201; 381; 940; 941; 1524
YP0363	13	13	10	19	34	13	20	ND	ND	1470; 1483; 3611
YP0364	4	7	43	5	108*	165*	11*	ND	ND	
YP0382	107	58	10	5	155*	79*	268*	ND	ND	
YP0384	118*	3	10	5	34	79*	268*	ND	ND	
YP0386	8	3	39	19	155*	32	22	ND	ND	
YP0391	148*	3	10	5	155*	79*	268*	ND	ND	
YP0392	14	7	30	4	34	3	159*	ND	ND	
YP0399	14	7	30	4	34	3	159*	ND	ND	
YP0474	4	7	43	5	97	165*	11*	ND	ND	
YP0493	148*	3	43	9	76*	32	107	ND	ND	
YP0537	14	7	30	4	34	13	159*	ND	ND	
YP0569	4	2	10	19	53	201*	22	ND	ND	
YP0577	13	43*	10	19	34	32	22	ND	ND	
YP0581	4	43*	10	19*	108*	51*	268*	ND	ND	
YP0631	4	16*	43*	5	108*	165*	11*	ND	ND	
YP0639	14	7	30	4	34	3	159*	ND	ND	
YP0760	148*	3	10	14	76*	32	107	ND	ND	
YP0801	4	7	43	5	108*	165*	11*	ND	ND	
YP1130	4	7	43	5	97*	165*	11*	ND	ND	

***** alleles with a partial match

3. Genomic variants analyses

After variant calling using the GATK best practices workflow, a total of 9845324 SNPs were identified. Due to the large number of SNPs and heterozygosity of these genomes (an expected feature for *Candida albicans* isolates), an extra filter was applied

according to[53]: SNPs were filtered for alternate allele support and allelic frequency as described in methods, section 3.2. This resulted in a total of 9821009 SNPs, ranging from 176406 in sample YP0581 to 75537 SNPs in sample YP0306 (Annex B).

There were 6494819 homozygous SNPs and 3230695 heterozygous SNPs, ranging between 57440 (sample YP0001) and 148943 (sample YP0581) for homozygous SNPs and from 5420 (sample YP0232) to 64286 (sample YP0399) for heterozygous SNPs (Annex C). The number of homozygous SNPs is superior to the number of heterozygous SNPs, which allow us to associate the high number of SNPs found to a high LOH frequency, an event that could have occurred as a survivor mechanism in a pressured environment and a generator of a wider range of diversity.

From the 9821009 high quality SNPs found, 5650310 were in non-coding regions and 4170055 in coding regions. Among SNPs located in coding regions, 1258006 were missense, 2608803 were synonymous variants and 303246 belong to other variant types (non-synonymous amino acid replacement, start codon gains or losses, and frame shifts). The number of SNPs found in non-coding regions ranged from 40765 in sample YP0031 to 116999 in sample YP0493 (Figure 7). Among the SNPs found in coding regions, the number of missense SNPs ranged from 12 (approximately 0% of all SNPs) in sample YP0103 to 43959 (approximately 30% of all SNPs) in sample YP0019, and the number of synonymous SNPs ranged from 14355 (approximately 18% of all SNPs) in sample YP0031 to 53899 (approximately 31% of all SNPs) in sample YP058 (Figure 8).

Regarding the number of SNPs in non-coding regions (Figure 7), all *C. albicans* isolates seemed to follow the same pattern of about 51% of SNPs in non-coding regions, except for 17 isolates which exhibited close to 80% of SNPs in non-coding regions. These were YP0233, YP0093, YP0083, YP0050, YP0760, YP0639, YP0493, YP0399, YP0363, YP0047, YP0045 collected from vaginal samples, YP0103, YP0101, YP0024 collected from oral samples, and YP0386, YP0176, YP0048 collected from samples taken from medical devices.

Regarding the number of SNPs in coding regions (Figure 8), 58 *Candida* isolates exhibited a frequency between 30-31% of synonymous SNPs while 24 isolates exhibited a frequency between 17-18% of synonymous SNPs. Most *Candida albicans* isolates followed the same pattern regarding SNPs classified herein as “others” (non-synonymous amino acid replacement, start codon gains or losses, and frame shifts), of about 94%, except for isolates YP0392, YP0326, YP0114 and YP0061, collected from vaginal samples, that exhibited a frequency between 32-34% of these SNPs. Most *Candida* isolates seemed to follow the same pattern regarding missense SNPs with frequencies between 17-18%,

however two isolates exhibited a frequency between 30-31% (YP0031 and YP0019 from oral samples), and 19 isolates did not exhibit synonymous SNPs (YP0392, YP0326, YP0233, YP0114, YP0093, YP0083, YP0061, YP0050, YP0760, YP0639, YP0493, YP0399, YP0363, YP0047, YP0045 collected from vaginal samples, and YP0103, YP0024 collected from oral samples, and YP0386, YP0176 collected from samples from medical devices. It's important to mention that all the isolates that didn't reveal missense SNPs also exhibited a different pattern of synonymous SNPs and others SNPs frequencies.

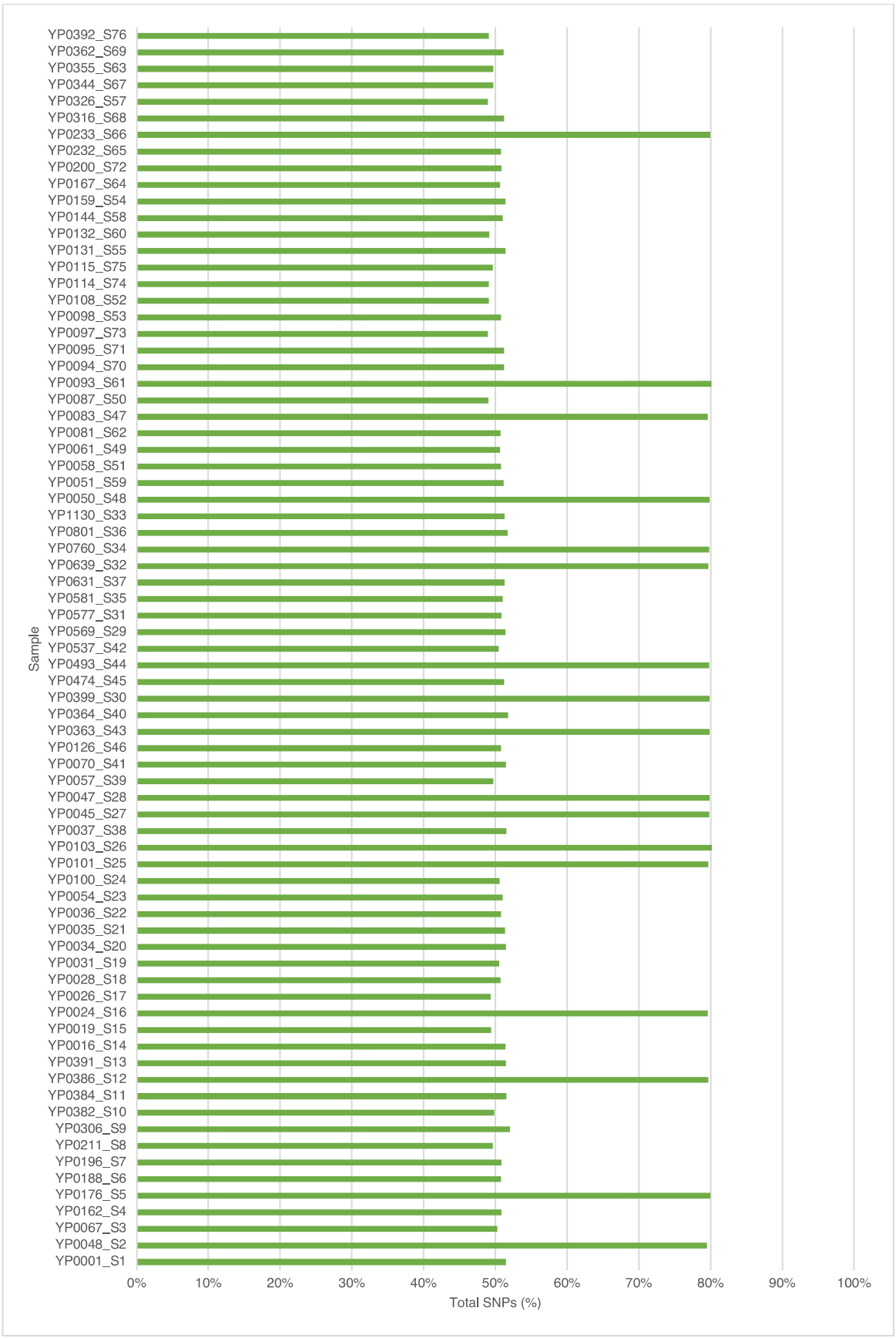


Figure 7: Total of SNPs found in non-coding regions, per sample.

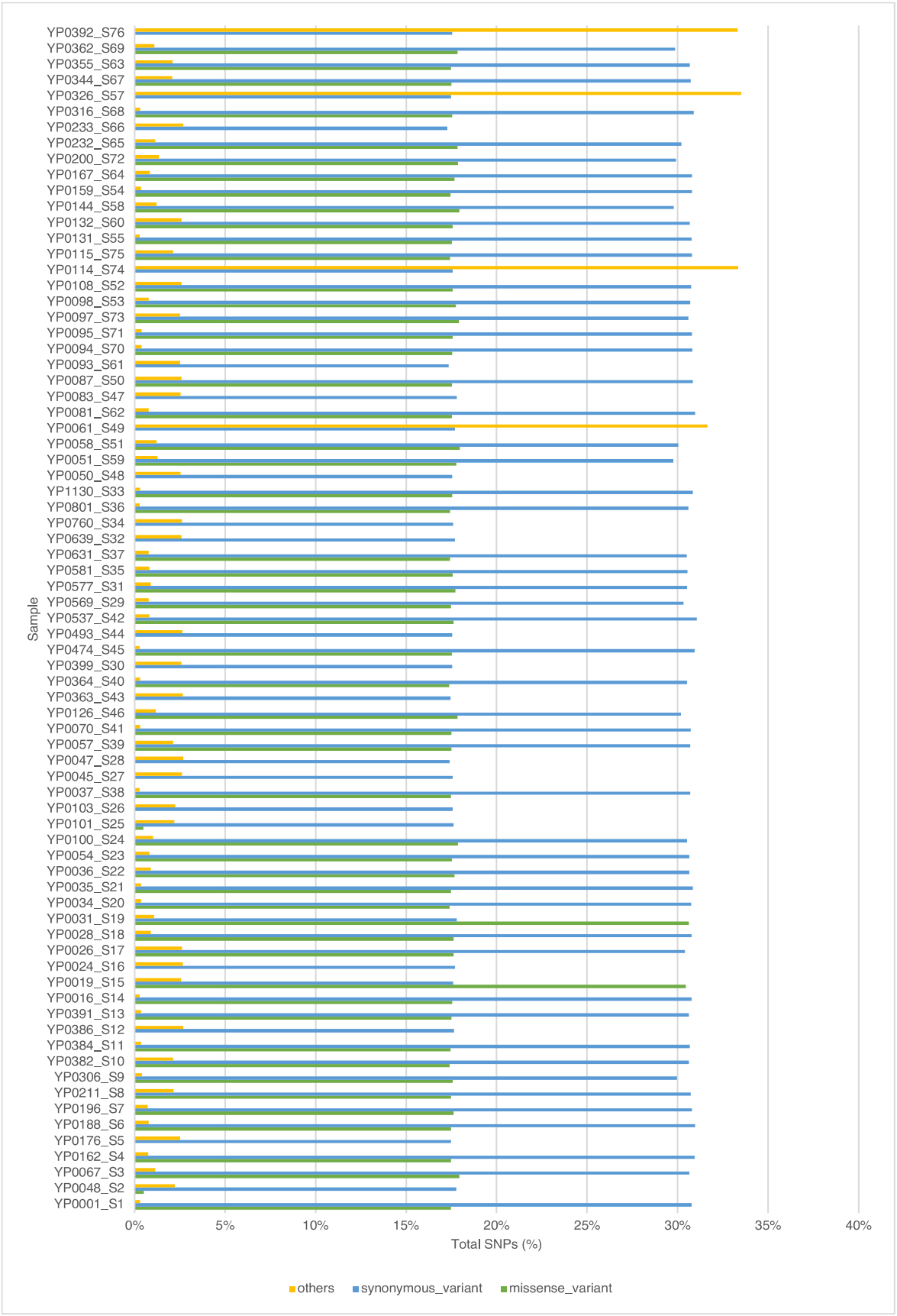


Figure 8: Total of SNPs found in coding regions, per sample.

The isolates studied herein were collected from oral, vaginal and blood samples, and from samples taken from medical devices, constituing samples from four distinct ecological niches.

Among SNPs found in vaginal samples, samples taken from medical devices, oral samples and blood samples, 55%, 58%, 58% and 61% were non-coding SNPs (Figure 9A).

Regarding SNPs in coding regions, isolates from blood registered the highest rate of synonymous SNPs (67%), samples from the oral cavity exhibited the highest rate (36%) of missense SNPs and samples of vaginal origin registered the highest frequency of other SNPs (13%) (Figure 9B).

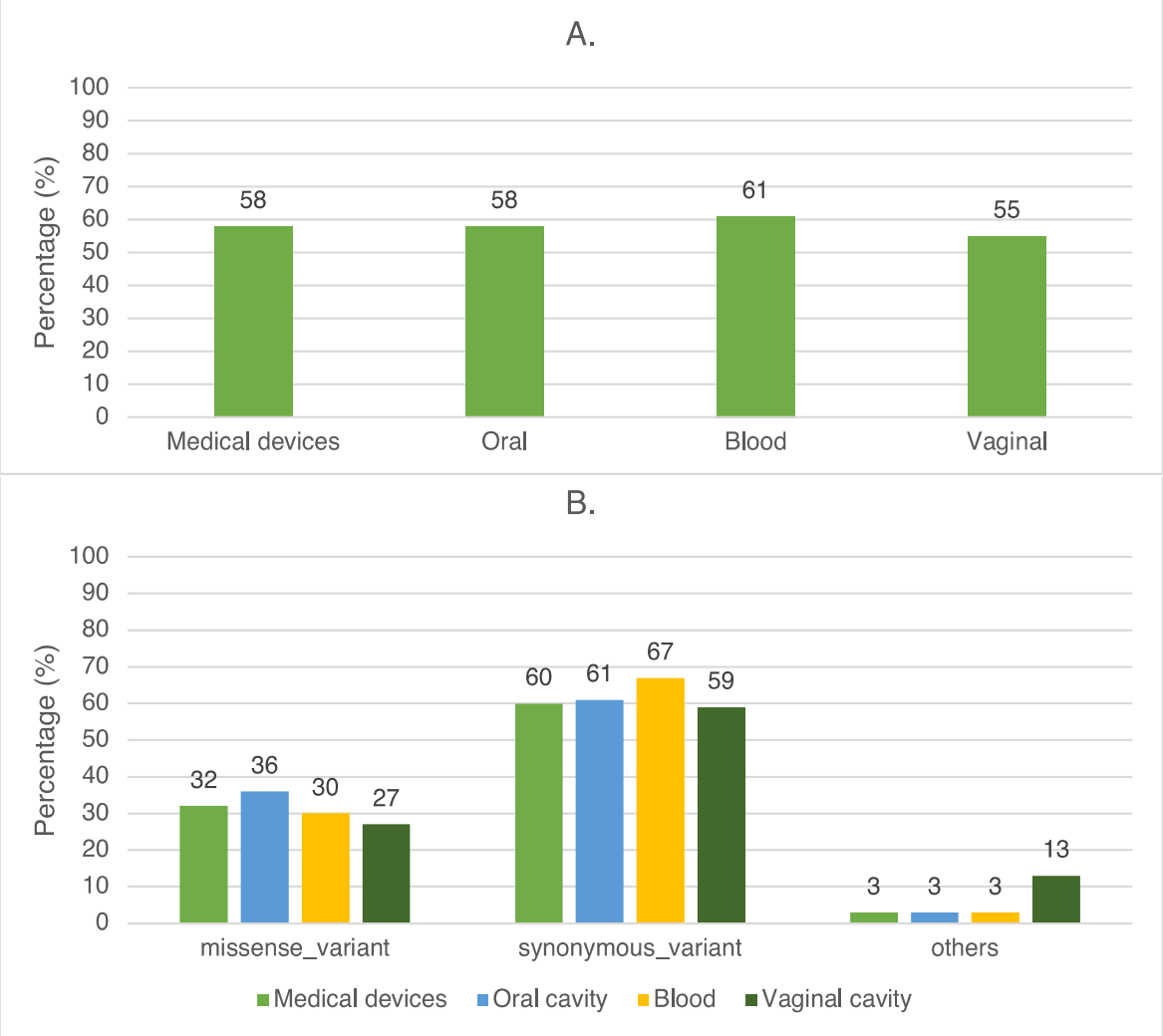


Figure 9: Distribution of found SNPs in A. noncoding and B. coding (missense variant, synonymous variant and others) regions in relation to the total number of SNPs, regarding their origin.

Vaginal samples displayed a higher number of SNPs herein classified as “others” SNPs compared to the remaining sample types. Among these, there were some types that were much higher in vaginal isolates than in the remaining isolates, such as splice_region_variant&intron_variant, splice_region_variant&non_coding_transcript _exon_variant, start_lost, stop_lost, stop_gained&splice_region_variant, splice_region_variant&non_coding_transcript_exon_variant and intron_variant (Table 4).

The splice_region_variant is a sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron sequence, while the intron_variant hits an intron in the transcript. Those effects have a low and modifier impact in the genome sequence, respectively. The non_coding_transcript _exon_variant hits an exon (from a non-coding transcript) or a retained intron. As so, the splice_region_variant& non_coding_transcript _exon_variant effects have a low and modifier impact on the genome sequence, respectively. The splice_region_variant SNPs already descripted and the stop_retained_variant that causes stop codon to be mutated into another stop codon (the new codon produces a different AA) have a low and low impact in the genome sequence, respectively. The start-lost causes a start codon to be mutated into a non-start codon, an effect with high impact in the genome sequence. The stop_lost variant causes stop codon to be mutated into a non-stop codon, an effect that has a high impact in the genome sequence. And lastly, the intron_variant as mentioned above has a modifier impact in the genome sequence[55].

Table 4: Type of coding SNPs among “others” frequency obtained for each sample type.

Type of coding SNPs among "others"	Impact	Medical Devices (n=13)	Oral samples (n=13)	Blood samples (n=20)	Vaginal samples (n=29)
splice_region_variant& intron_variant	Low	13260	7769	15748	**184193**
start_lost	High	704	714	3361	**13398**
splice_region_variant& stop_retained_variant	Low & low	806	6112	4385	1750
stop_lost& splice_region_variant	High &low	941	660	927	1298
stop_lost	High	247	624	599	**1257**
stop_gained& splice_region_variant	High &low	353	594	744	**1226**
stop_gained	High	775	489	1533	1715
initiator_codon_variant	Low	121	343	566	799
stop_retained_variant	Low	224	340	412	400
splice_donor_variant& intron_variant	High & modifier	32	36	75	75
splice_region_variant& non_coding_transcript _exon_variant	Low	3337	4759	7904	**10129**
intron_variant	Modifier	769	756	1579	**4232**
splice_acceptor_variant& Intron_variant	High & modifier	41	47	44	61
Total		21610	23243	37877	220553

In total, 1073 missense SNPs were found in all isolates from all types of samples, 767, 499, 243, and 138 SNPs were exclusively found among all isolates collected from medical devices, blood, oral and vaginal samples, respectively. Figure 10 displays the number of missense SNPs found among the isolates in study according to the ecological niche isolates were collected from: Oral, Vaginal, Blood and Medical Devices.

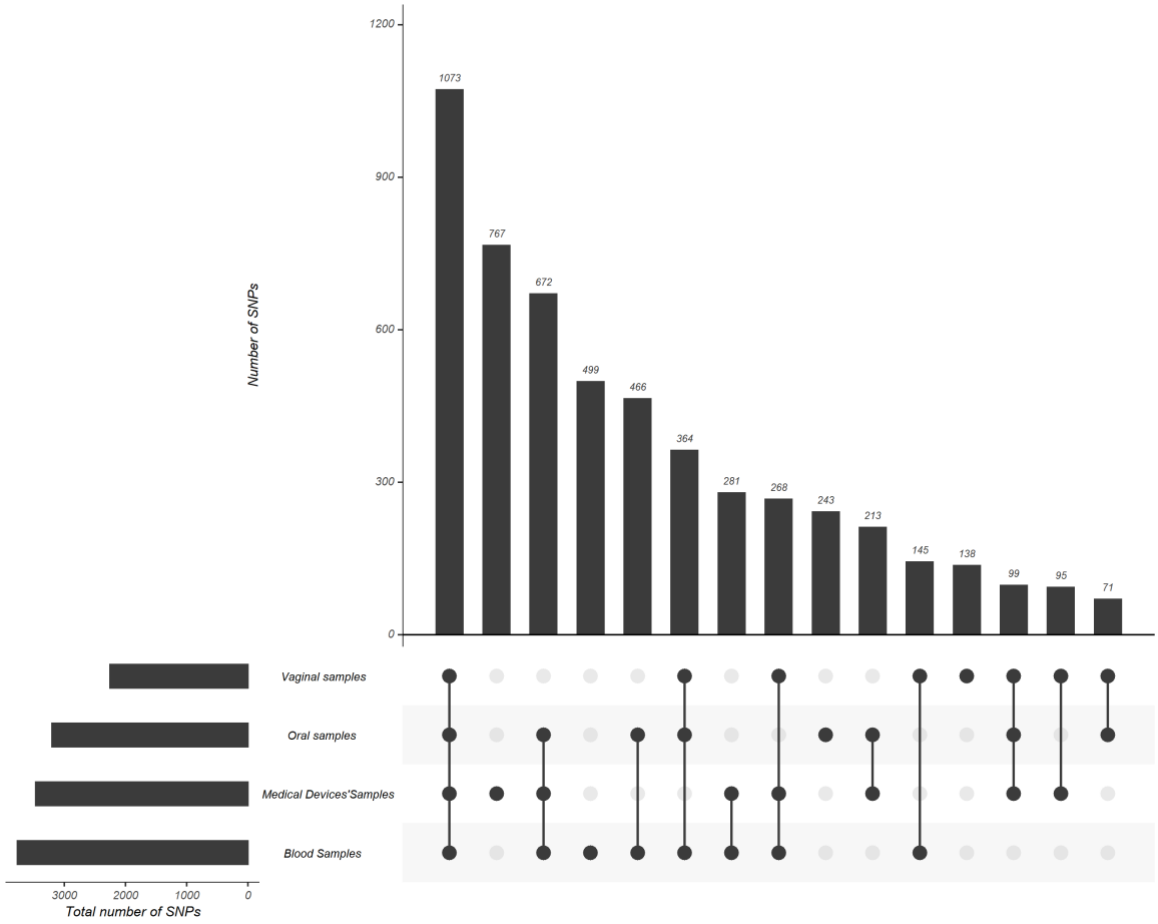


Figure 10: Venn diagram displaying the number of SNPs common to isolates from distinct ecological niches and of SNPs common to isolates from the same ecological niche (Oral; Vaginal; Medical Devices; Blood).

A total of 689 genes were found to have missense SNPs common to all isolates from the distinct niches. Also, 1601, 1282, 1896 and 1673 genes were found to have SNPs common to all isolates of oral, vaginal, blood and medical devices' samples, respectively. Additionally, 170, 100, 308 and 460 genes were found to have SNPs common to and exclusive of all the oral, vaginal, blood and medical devices' isolates, respectively. Figure 11 displays the number of genes bearing SNPs identified among the isolates in study according to the ecological niche isolates were collected from: Oral, Vaginal, Blood and Medical Devices.

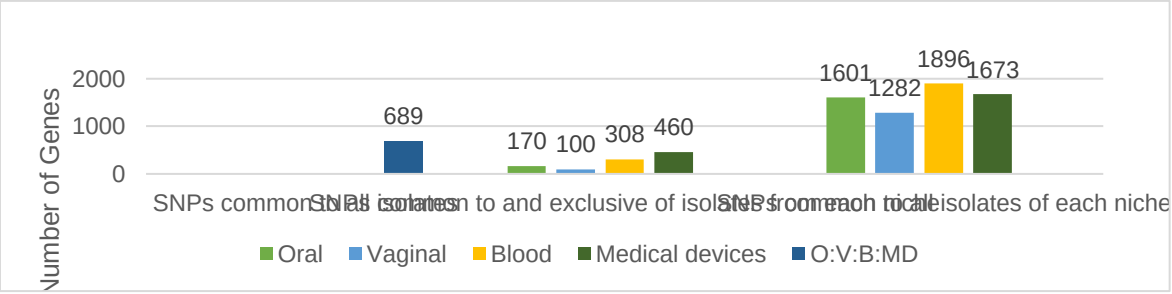


Figure 11: Number of genes with: SNPs common to all isolates, SNPs common to and exclusive of isolates from each niche, and SNPs common to all isolates of each niche.

4. Gene Ontology (GO) analysis of *C. albicans* isolates

Go Slim and GO Term finder analyses were performed to understand the role of the genes where SNPs were identified at the level of molecular function, biological process, or cellular component.

4.1. Genes with SNPs common to all *C. albicans* isolates from all sample types

To understand the variability of the clinical isolates and isolates of the clinical environment in study relatively to the well-studied laboratory *C. albicans* strain SC5314, we selected the genes with SNPs common to all isolates of all sample types (n=689), hypothesizing this variance might be due to the adaptation of the isolates to the human organism and to a hostile environment as the clinical environment is for microorganisms.

The GO Slim analysis mapped (i) genes coding for molecular functions, from which hydrolase activity (14.8%), transferase activity (12.9%) and protein binding (9.3%) were the most common; (ii) genes encoding for cellular components, from which cytoplasm (34.4%), nucleus (26.1%) and membrane (18.9%) were the most common; and (iii) genes encoding for products with a role in biological processes, where response to stress (17.4%), ribosome biogenesis (16.7%) and filamentous growth (13.1%) were the most common (Figure 12 and 13).

The GO Term analysis performed assuming p-value≤0.05, significantly annotated genes to GO terms in the biological process ontology: autophagic mechanism (p-value = 0.2598), biological regulation (p-value = 0.03796), regulation of cellular process (p-value = 0.02558), regulation of biological process (p-value = 0.02360) and autophagy (p-value = 0.00719 (Figure 14).

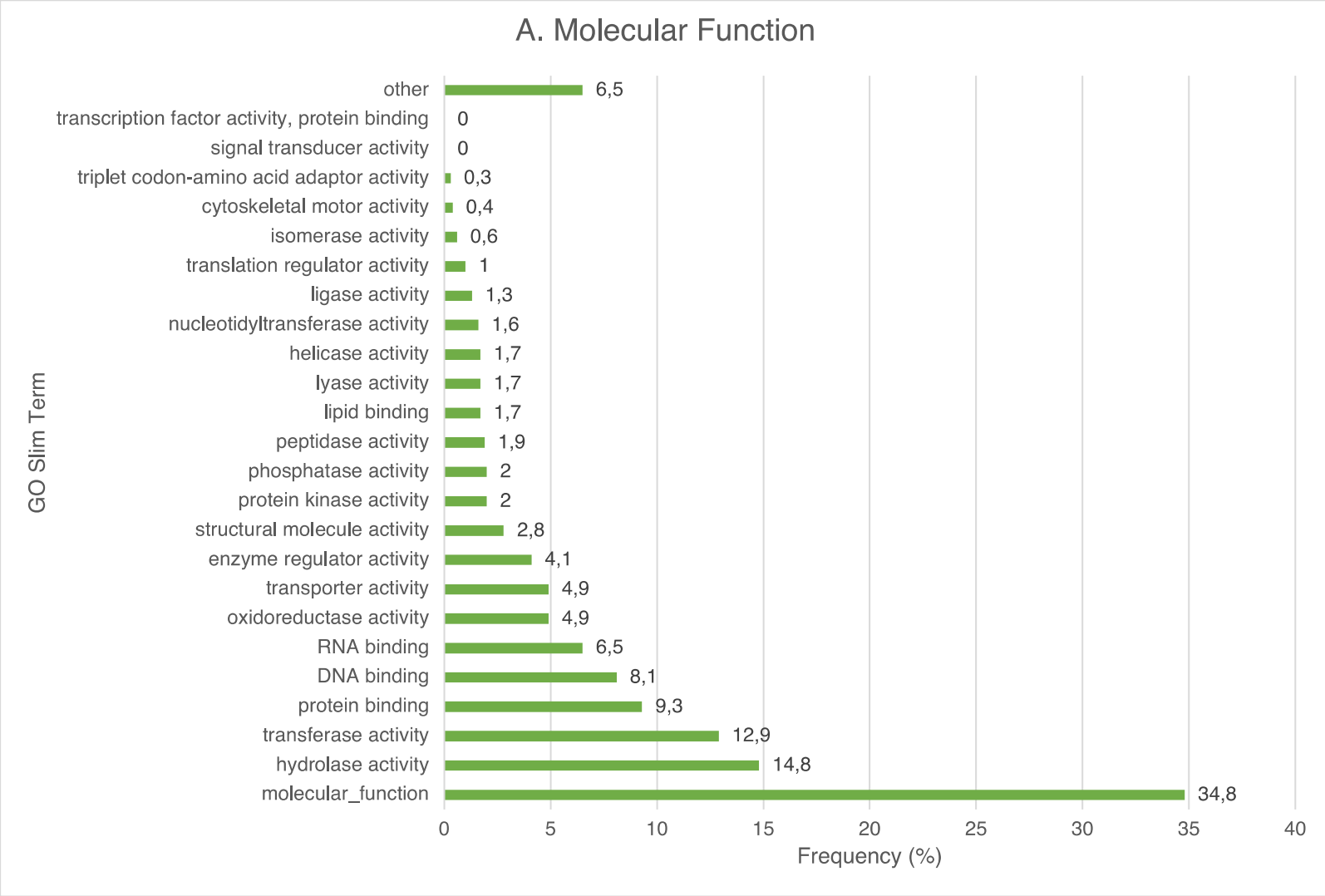


Figure 12: Frequencies of GO Slim terms mapped to genes with missense SNPs common to all isolates of the four origins, regarding A. Molecular Function.

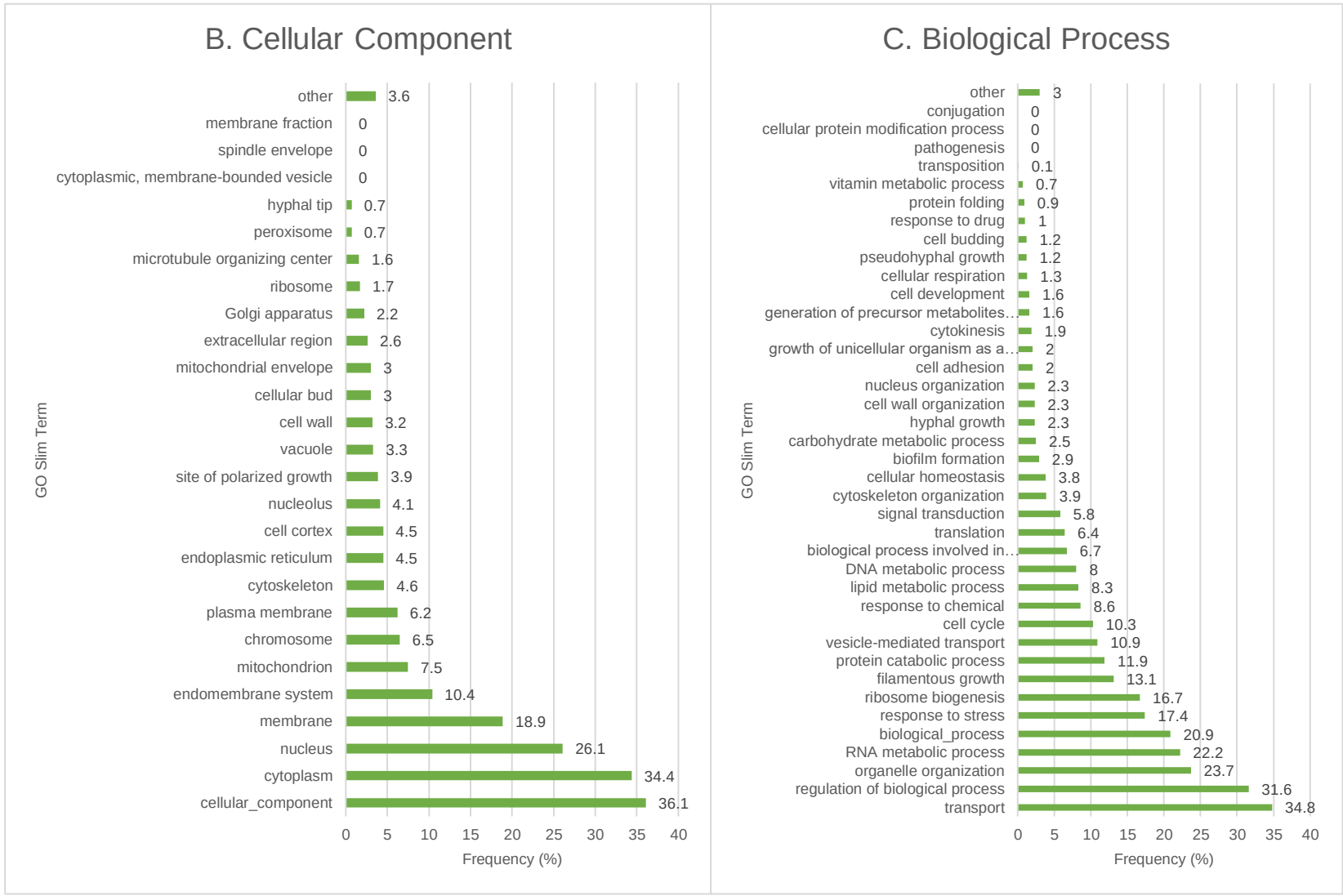


Figure 13: Frequencies of GO Slim terms mapped to genes with missense SNPs common to all isolates of the four origins, regarding B. Cellular Component and C. Biological Process.

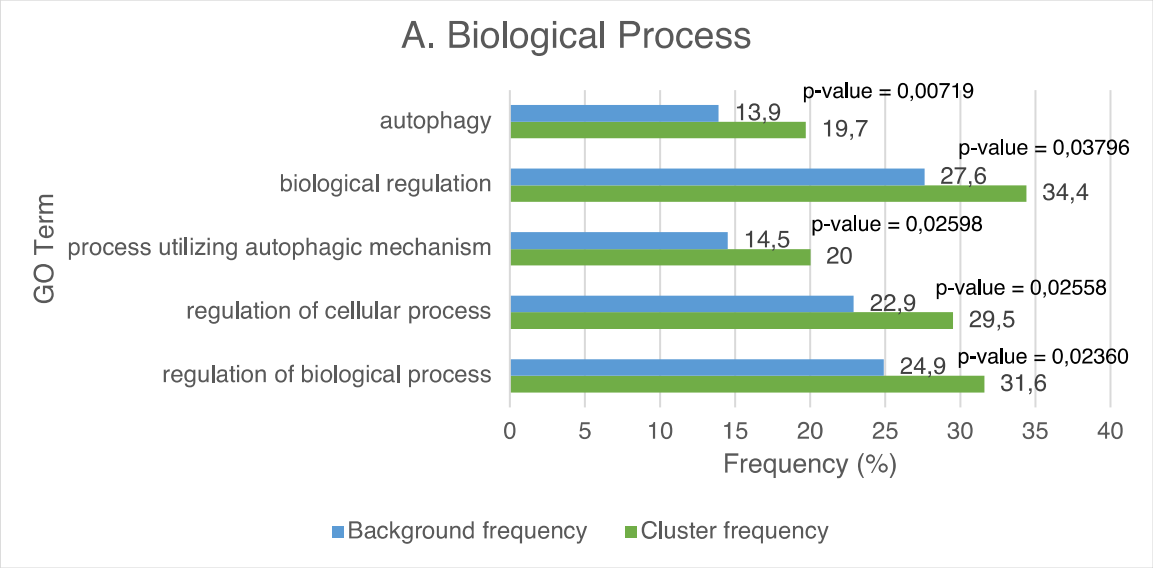


Figure 14: GO Term cluster frequencies of genes with missense SNPs common to all isolates of the four origins.

4.2. Genes with SNPs common to and exclusive of all isolates of each sample type

To understand which was the exclusive genomic variability of the isolates according to their sample origin or ecological niche, we selected genes with SNPs exclusively found in all isolates collected from the same type of sample, with the aim of understanding which were the genes with exclusive variations that might have played a role in the adaptation to a particular ecological human or hospital niche in study.

4.2.1. Medical Devices

In relation to the genes with exclusive SNPs found in isolates from medical devices (n=460), the GO Slim analysis mapped (i) genes coding for molecular function, from which the hydrolase activity (13.9%), transferase activity (13.9%) and DNA binding (8%) were the most common; (ii) genes encoding for cellular components, from which cytoplasm (30.2%), nucleus (24.1%) and membrane (18.4%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which response to stress (17.4%), ribosome biogenesis (16.3%), and filamentous growth (13.1%) were the most common (Figure 15 and 16).

There were no GO terms annotated to the genes of any of the three ontologies using the methods used herein.

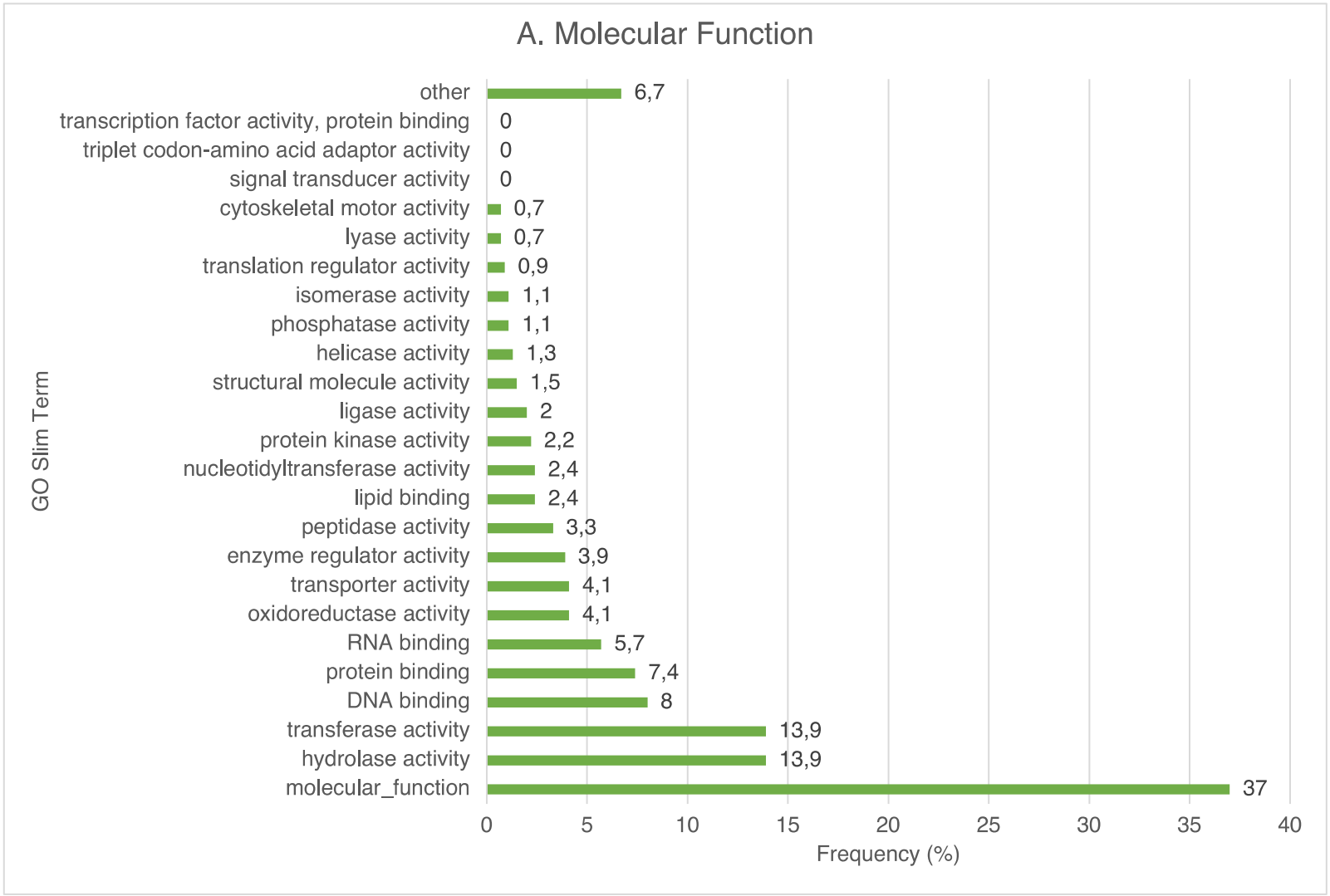


Figure 15: Frequencies of GO Slim terms mapped to genes with missense SNPs exclusively found in isolates from medical devices, regarding A. Molecular Function.

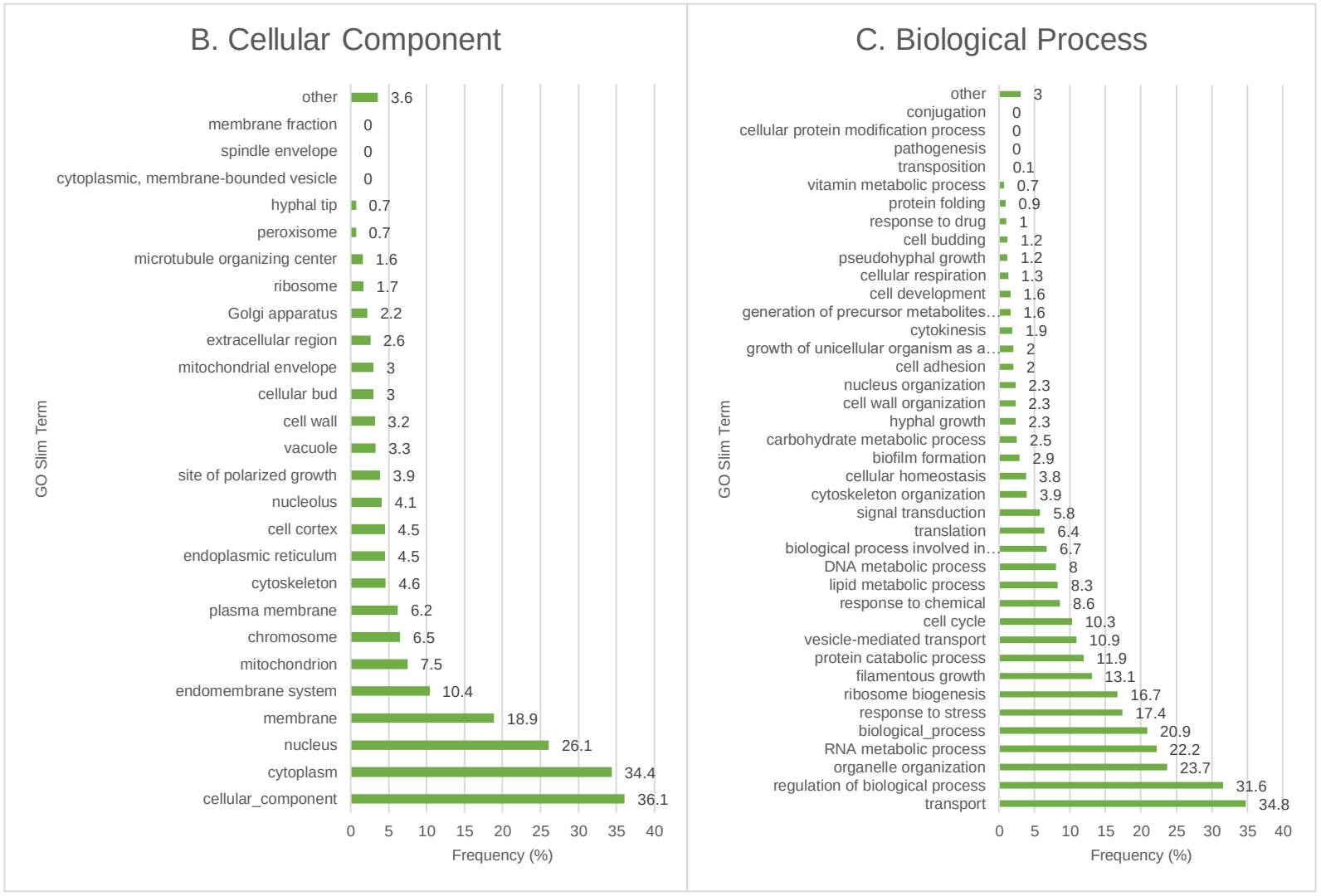


Figure 16: Frequencies of GO Slim terms mapped to genes with missense SNPs exclusively found in isolates from medical devices, regarding B. Cellular Component and C. Biological Process.

4.2.2. Blood

Regarding the genes with exclusive SNPs of isolates from blood samples (n=308), the GO Slim analysis mapped (i) genes coding for molecular function, from which the hydrolase activity (12%), transferase activity (12.7%) and DNA binding (7.5%) were the most common; (ii) genes encoding for cellular components, from which cytoplasm (32.5%), nucleus (26.6%) and membrane (28.8%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which RNA metabolic process (21.4%), organelle organization (21.1%) and ribosome biogenesis (13.6%) were the most common (Figure 17 and 18).

With the GO Term analysis performed, there were no annotated genes with GO Terms of any of the three ontologies.

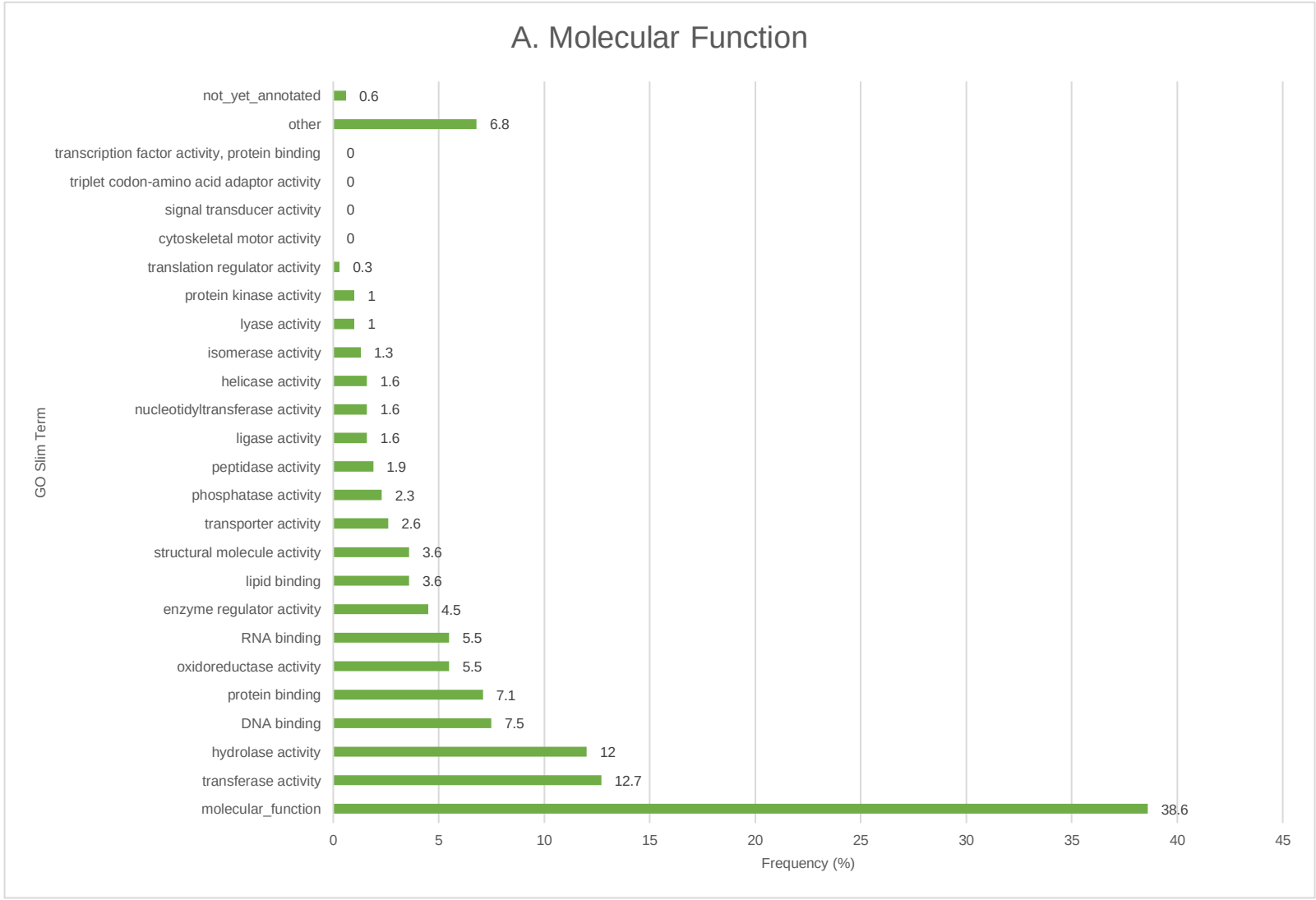


Figure 17: Frequencies of GO Slim terms mapped to genes with missense SNPs that were exclusively found in isolates from blood samples, regarding A. Molecular Function.

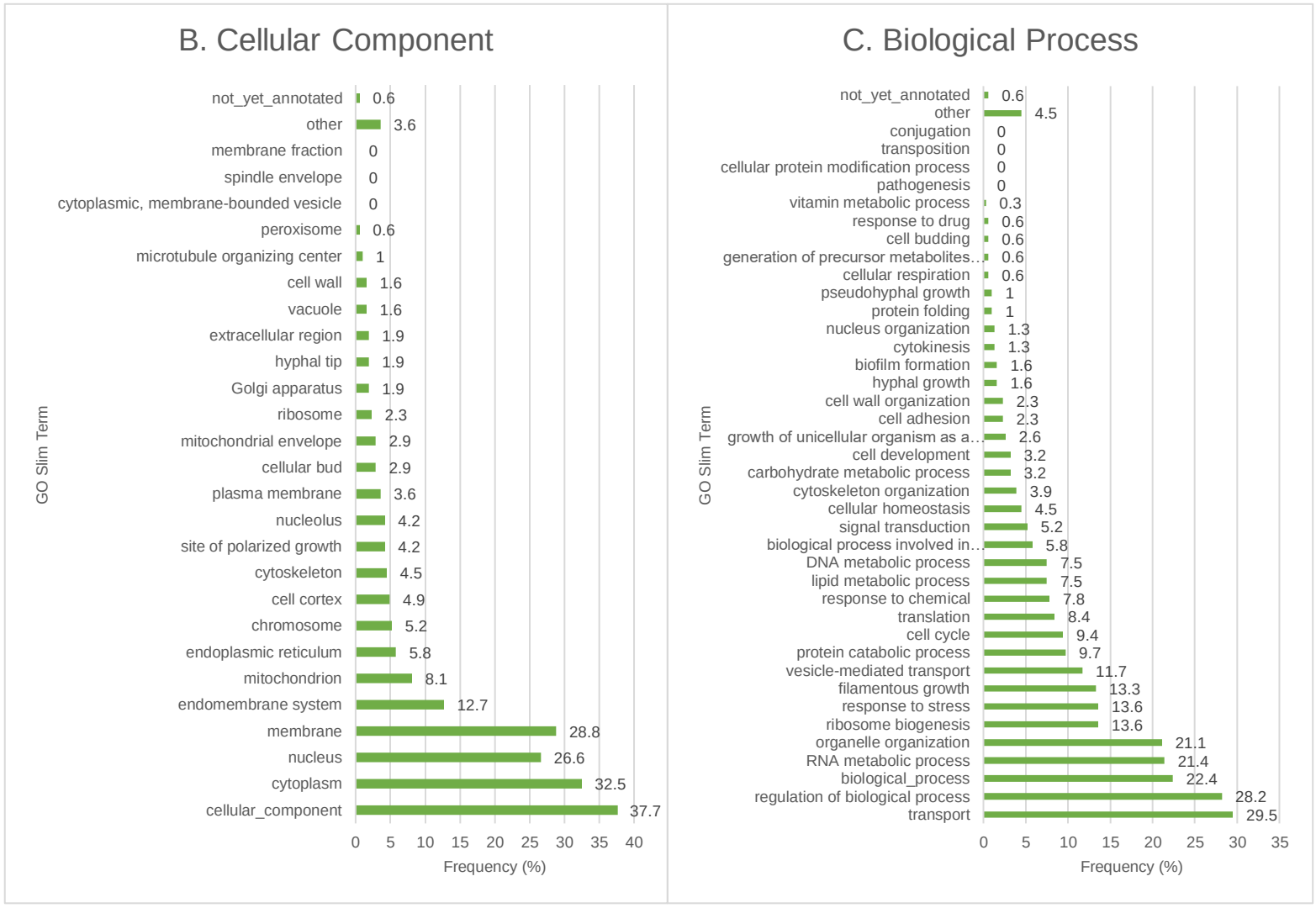


Figure 18: Frequencies of GO Slim terms mapped to genes with missense SNPs that were exclusively found in isolates from blood samples, regarding B. Cellular Component and C. Biological Process.

4.2.3. Oral

In what concerns genes with exclusive SNPs of isolates from oral samples (n=170), the GO Slim analysis mapped (i) genes coding for molecular function, from which the hydrolase activity (11.2%), transferase activity (16.5%) and protein binding (5.9%) were the most common; (ii) genes encoding for cellular components, from which cytoplasm (28.8%), nucleus (21.2%) and membrane (18.8%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which RNA metabolic process (20.6%), organelle organization (21.8%) and response to stress (14.1%) were the most common (Figure 19 and 20).

With the methods used in this study, genes with exclusive SNPs from isolates with oral origin were annotated with GO terms of the biological process ontology: protein localization to chromosome, telomeric region (p-value = 0.04909). Despite the p-value≤0.05, the false discovery rate of this result was 0,06 and so this association was not deemed significant (Figure 21).

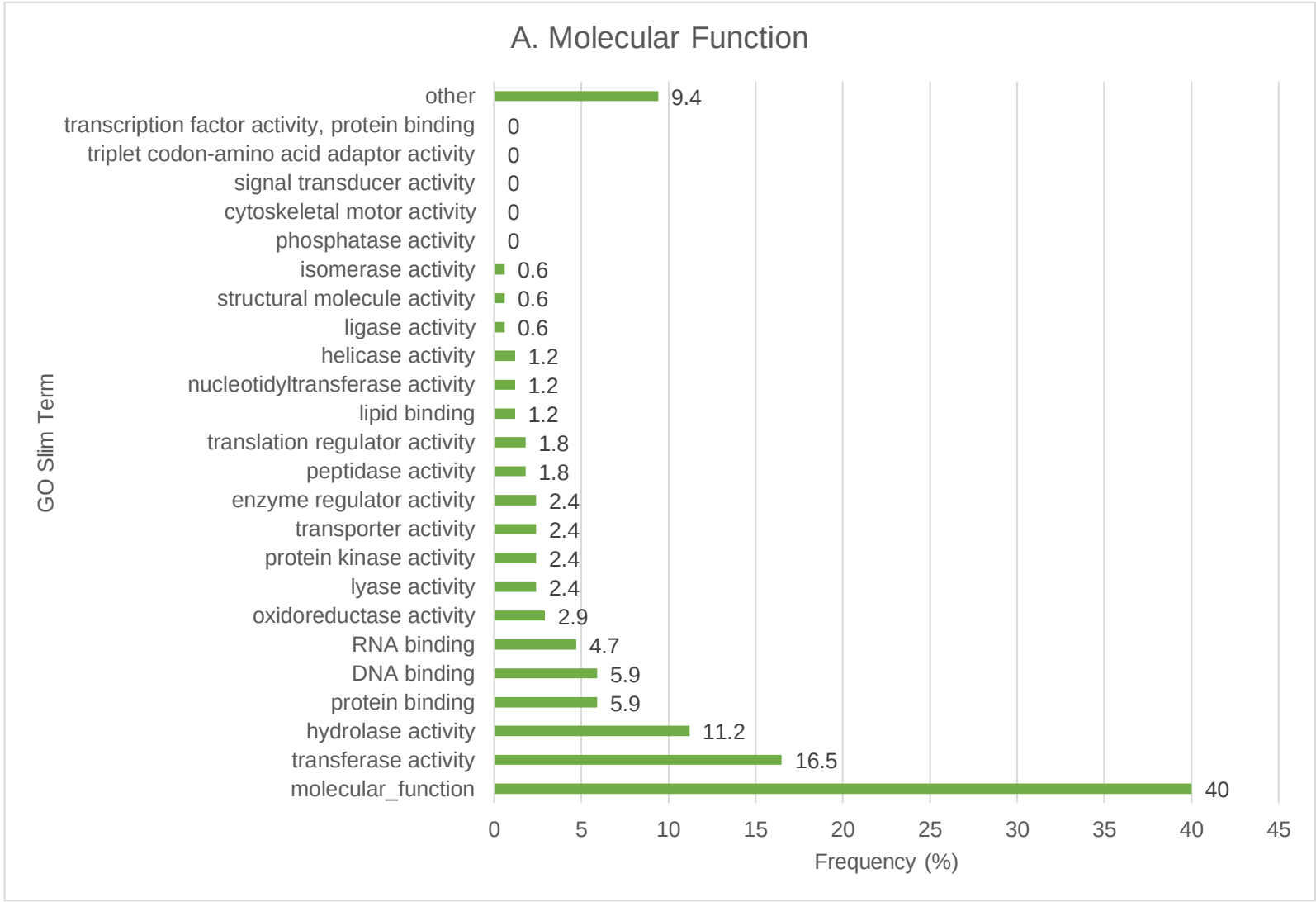


Figure 19: Frequencies of GO Slim terms mapped to genes with missense SNPs that were exclusively found in isolates from oral samples**,** regarding A. Molecular Function

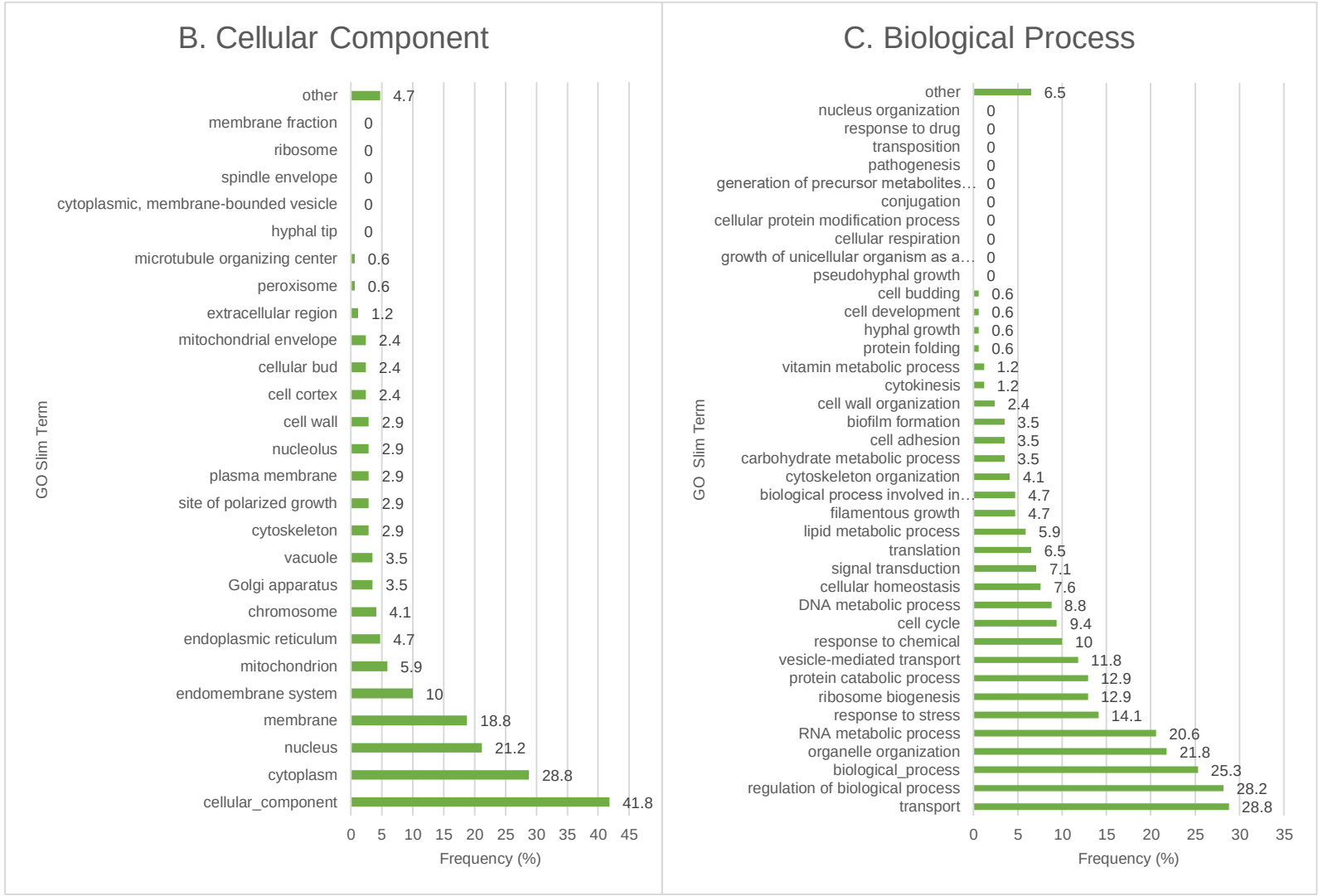


Figure 20: Frequencies of GO Slim terms mapped to genes with missense SNPs that were exclusively found in isolates from oral samples**,** regarding B. Cellular Component and C. Biological Process.

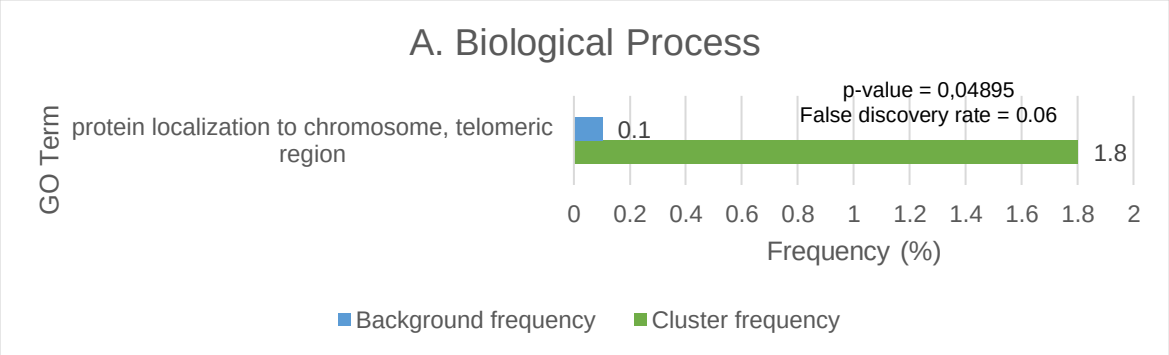


Figure 21: GO Term cluster frequencies of genes with missense SNPs that were exclusively found in isolates from oral samples referent to A. Biological Process.

4.2.4. Vaginal

With regards to the genes with exclusive SNPs of isolates from vaginal samples (n=100), the GO Slim analysis mapped (i) genes coding for molecular function, from which the hydrolase activity (12%), transferase activity (13%) and protein binding (8%) were the most common; (ii) genes encoding for cellular components, from which cytoplasm (29%), nucleus (18%) and membrane (16%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which organelle organization (20%), RNA metabolic process (23%) and regulation of biological process (23%) were the most common (Figure 22 and 23).

The GO Term analysis we performed did not annotate any of these genes to GO Terms of any of the three ontologies.

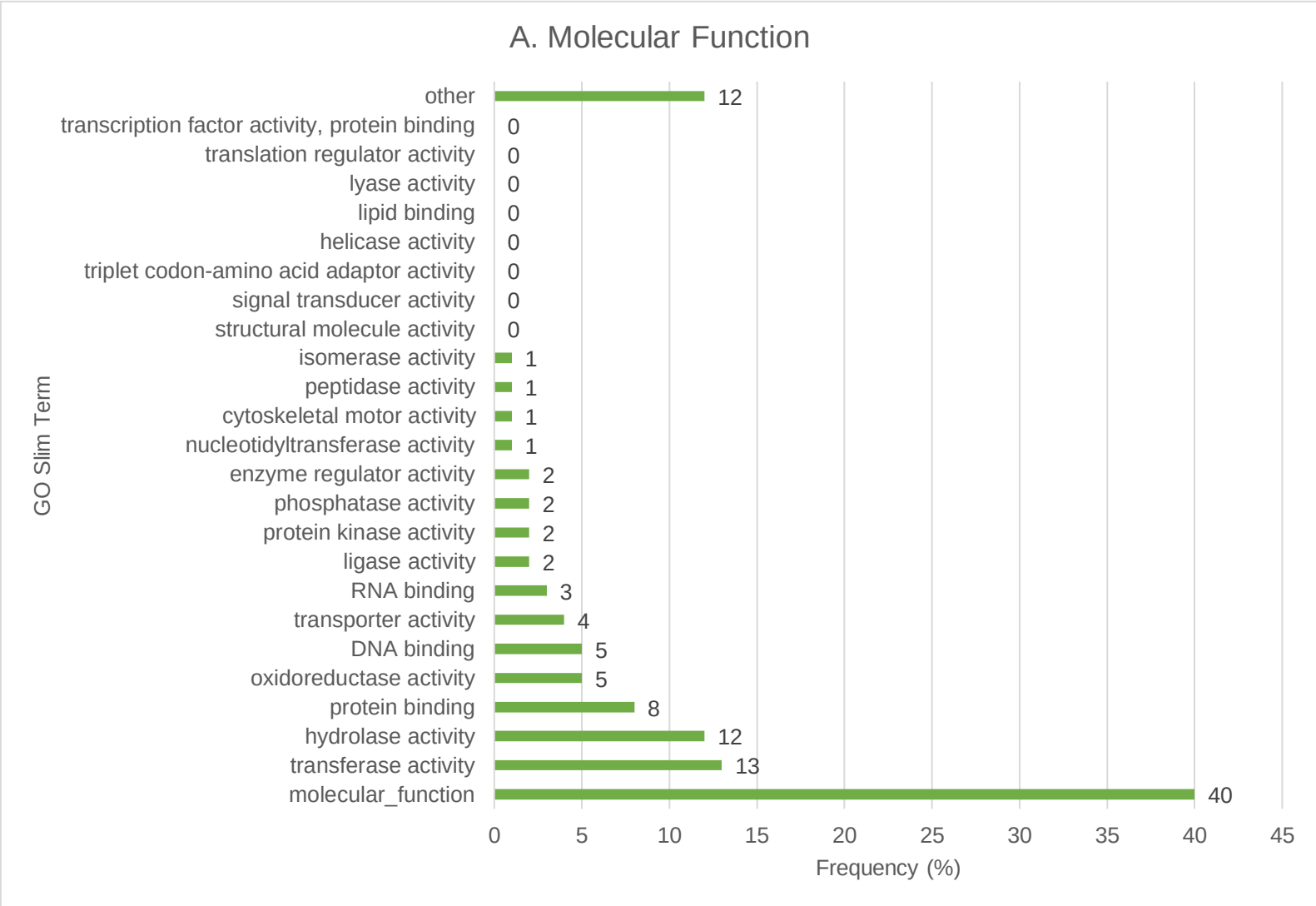


Figure 22: Frequencies of GO Slim terms mapped to genes with missense SNPs that were exclusively found in isolates from vaginal samples, regarding A. Molecular Function.

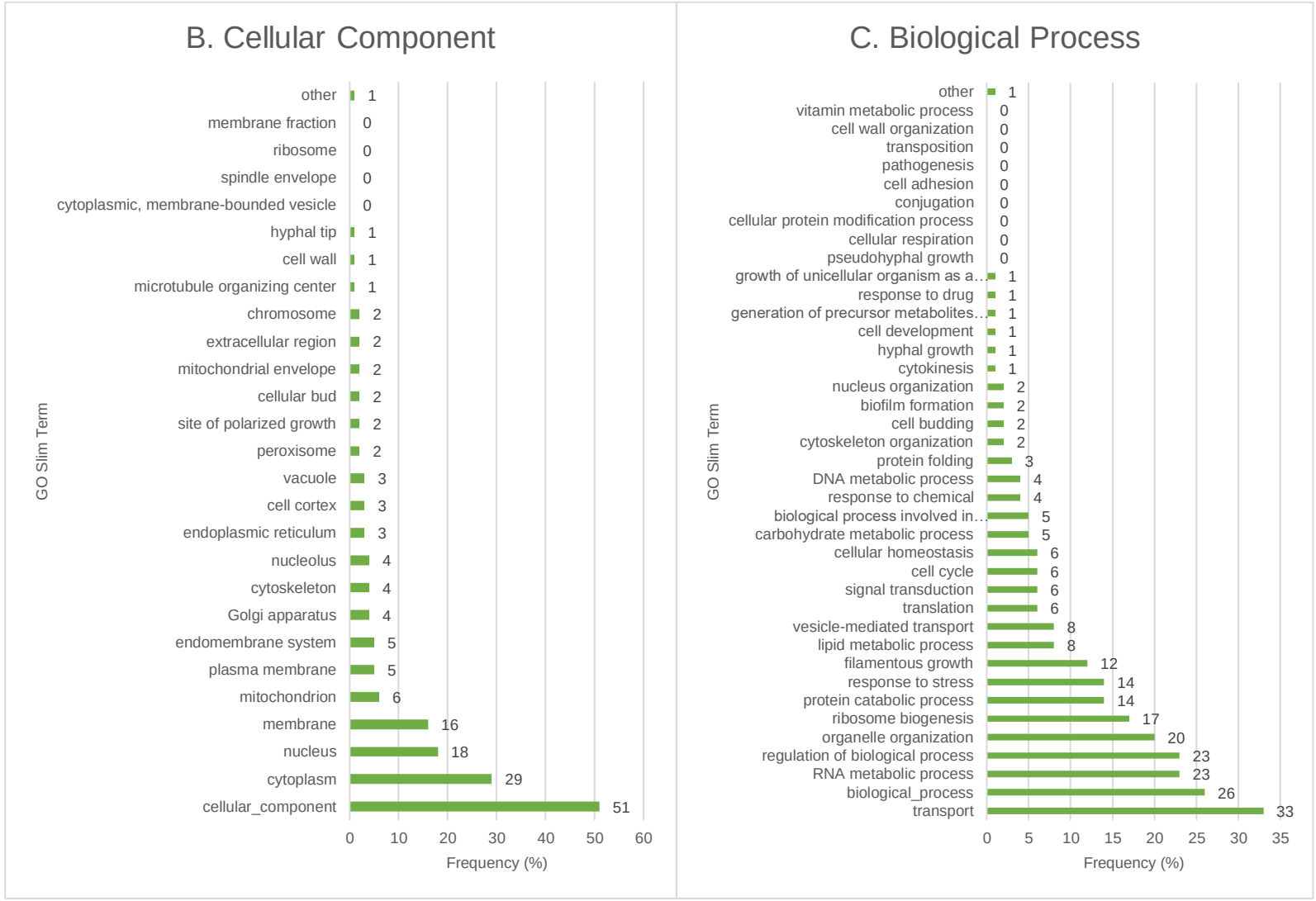


Figure 23: Frequencies of GO Slim terms mapped to genes with missense SNPs that were exclusively found in isolates from vaginal samples, regarding B. Cellular Component and C. Biological Process.

4.3. Genes with SNPs common to all *C. albicans* isolates of each sample type

To understand the overall variability of the isolates according to their sample origin or ecological niche, we selected genes with SNPs found in all isolates collected from the same type of sample, with the aim of understanding which was the group of genes that might have played a role in the adaptation to a particular ecological human or hospital niche in study.

4.3.1 Medical Devices

Regarding the genes with common SNPs found in all isolates from samples taken from medical devices (n=1673), the GO Slim analysis mapped (i) genes coding for molecular function, from which the hydrolase activity (13.2%), transferase activity (13.4%) and DNA binding (9.3%) were the most common; (ii) genes encoding for cellular components, from which cytoplasm (33.5%), nucleus (25.9%) and membrane (18.4%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which response to stress (16.2%), ribosome biogenesis (15.5%) and protein catabolic process (12.2%) were the most common (Figure 24 and 25).

The GO Term analysis performed revealed genes annotated to GO Terms of the molecular function and cellular component ontologies. DNA binding (p-value = 0.01073), catalytic activity, acting on DNA (p-value = 0.02780), and catalytic activity, acting on a nucleic acid (p-value = 0.00410) were GO terms of molecular functions. Nucleus (p-value = 0.00030) was the only GO term of cellular components found (Figure 26 and 27).

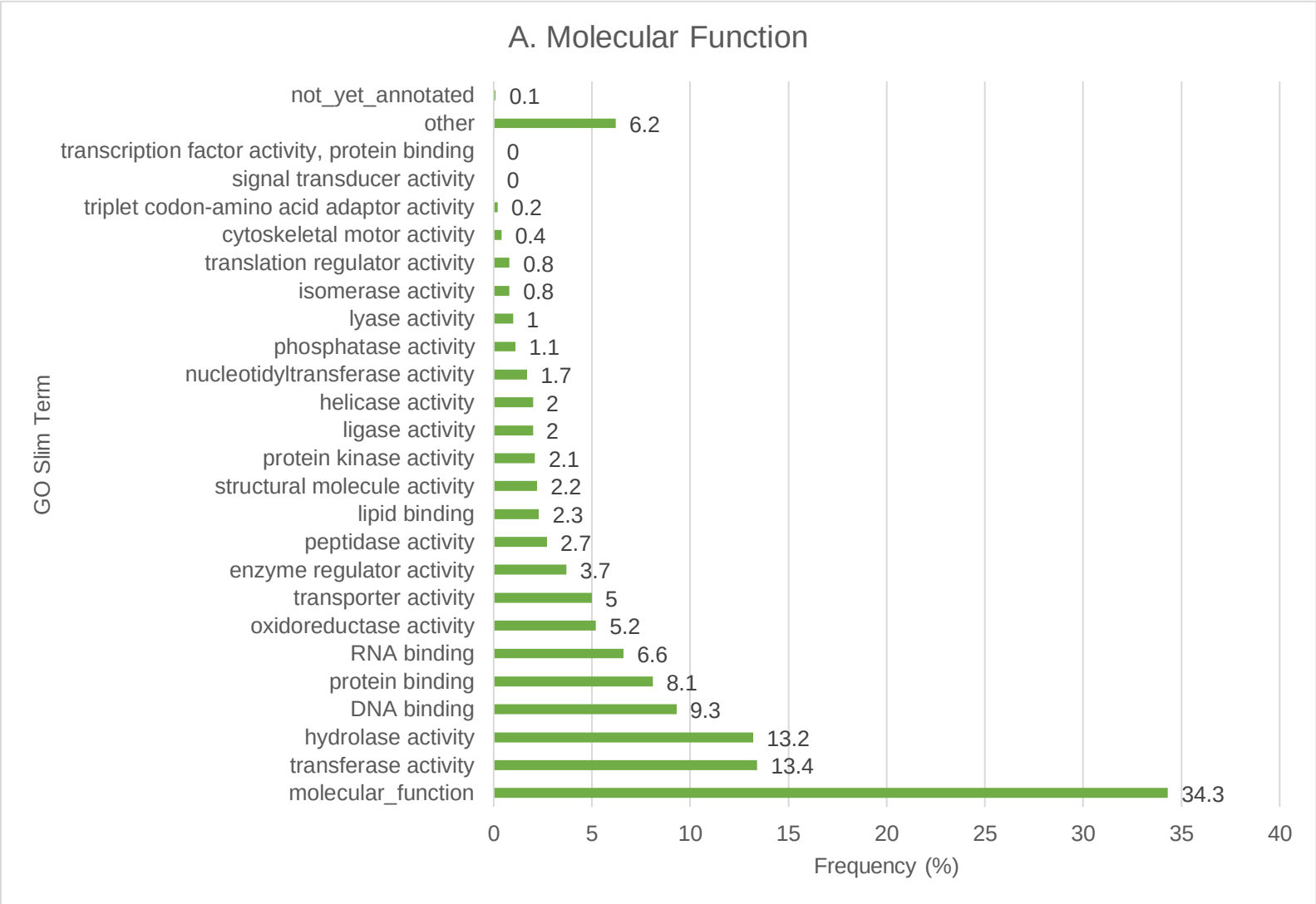


Figure 24: Frequencies of GO Slim terms mapped to genes with missense SNPs that were common to all isolates from medical devices, regarding A. Molecular Function.

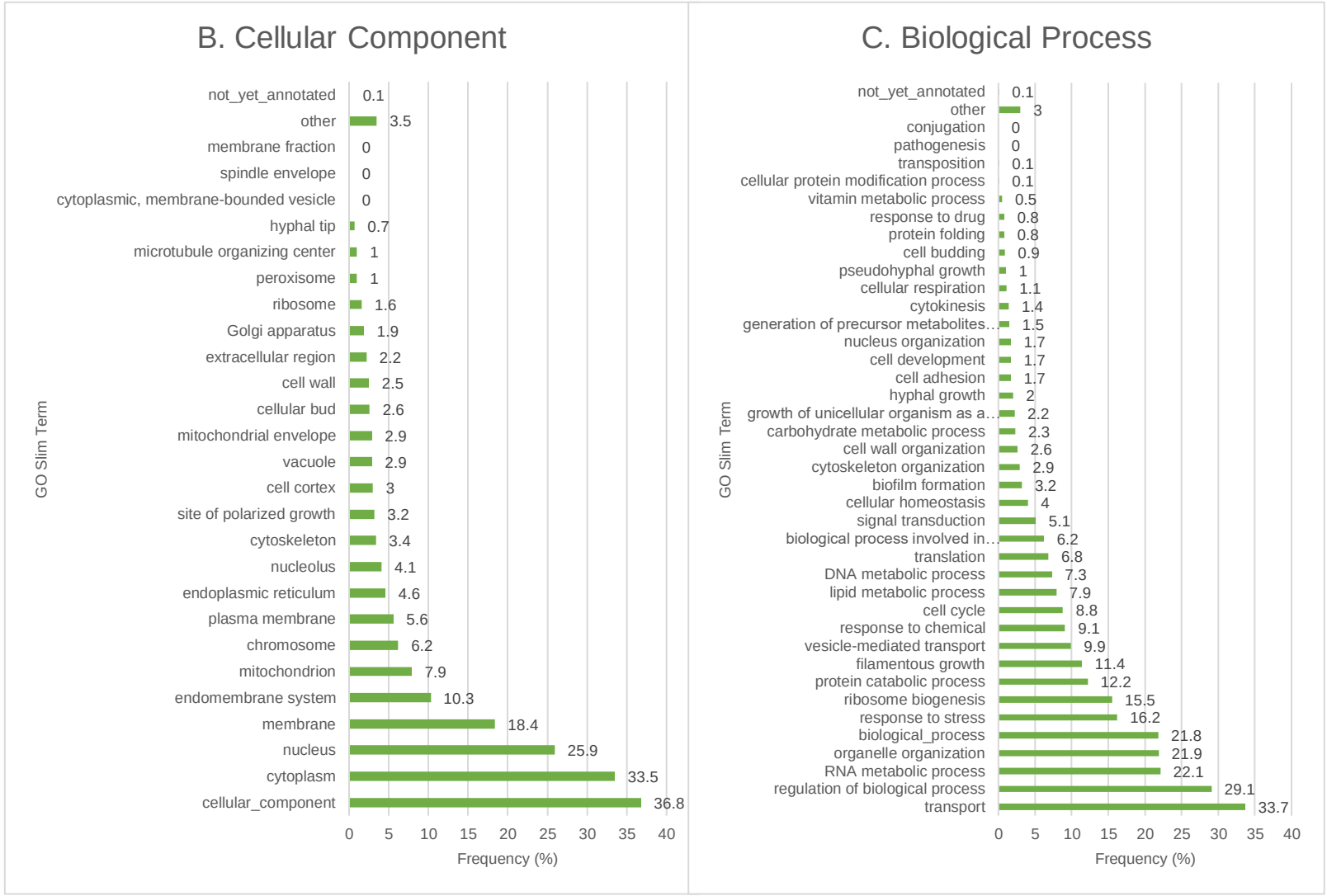


Figure 25: Frequencies of GO Slim terms mapped to genes with missense SNPs that were common to all isolates from medical devices, regarding B. Cellular Component and C. Biologic Process.

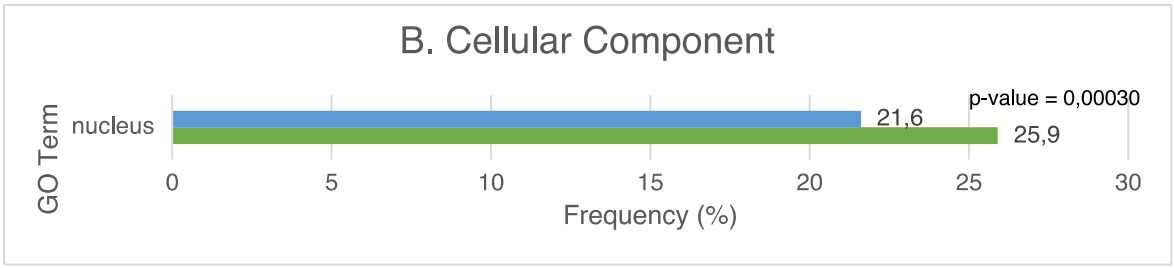


Figure 26: GO Term cluster frequencies of genes with missense SNPs that were common to all isolates from medical devices referent to B. Cellular Component.

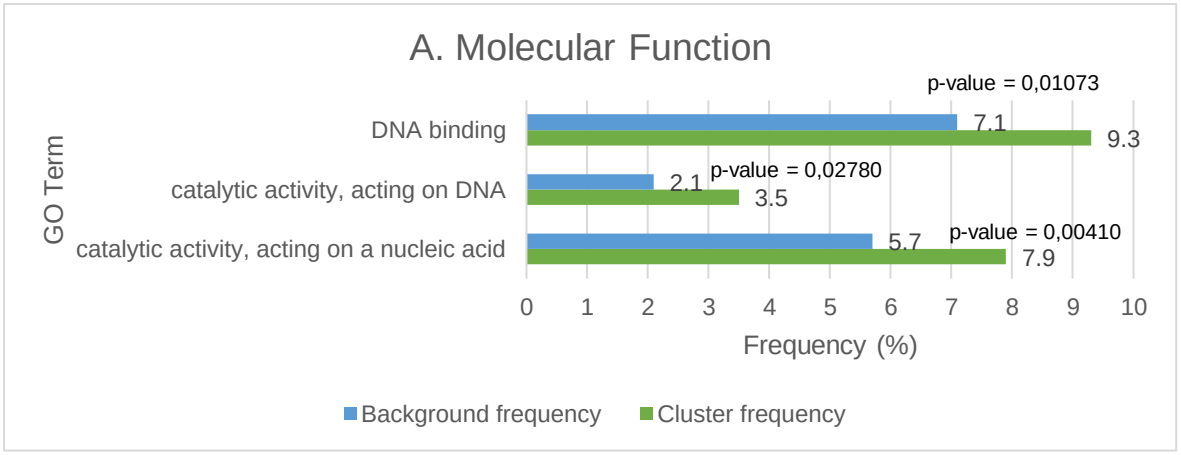


Figure 27: GO Term cluster frequencies of genes with missense SNPs that were common to all isolates from medical devices referent to A. Molecular Function.

4.3.2. Blood

Regarding genes with SNPs common to all blood isolates (n=1896), the GO Slim analysis mapped (i) genes coding for products involved in molecular functions, from which the hydrolase activity (13.1%), transferase activity (12.7%) and DNA binding (8.8%) were the most common; (ii) genes encoding for products related to cellular components, from which cytoplasm (33.9%), nucleus (25.9%) and membrane (18.8%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which RNA metabolic process (21.9%), organelle organization (20.7%) and ribosome biogenesis (15%) were the most common (Figure 28 and 29).

The GO Term analysis revealed genes associated to GO Terms from the three ontologies. The GO term found from the cellular component ontology was nucleus (p-value = 4.11e-05; Figure 30). The GO Term from the molecular function ontology was catalytic activity, acting on a nucleic acid (p-value = 0.04772; Figure 31). The biological processes GO terms associated with these genes were biologic regulation (p-value = 0.04217), establishment of protein localization to vacuole (p-value = 0.03635), vacuolar transport (p-

value = 0.03426), protein targeting to vacuole (p-value = 0.03018), regulation of cellular process (p-value = 0.02905), negative regulation of macromolecule metabolic process (p-value = 0.02611), regulation of nitrogen compound metabolic process (p-value = 0.02521), negative regulation of biologic process (p-value = 0.02134), protein localization to vacuole (p-value = 0.01344), negative regulation of metabolic process (p-value = 0.01043), regulation of biologic process (p-value = 0.00422), regulation of macromolecule metabolic process (p-value = 0.00187), regulation of metabolic process (p-value = 0.00165), negative regulation of gene expression (p-value = 0.00139), cellular catabolic process (p-value = 0.00118), catabolic process (p-value = 0.00112), autophagy of nucleus (p-value = 1.72e-07), lysosomal microautophagy (p-value = 1.45e-07), piecemeal autophagy of the nucleus (p-value = 8.46e-08), process utilizing autophagic mechanism (p-value = 4.76e-08) and autophagy (p-value = 2.12e-08; Figure 32).

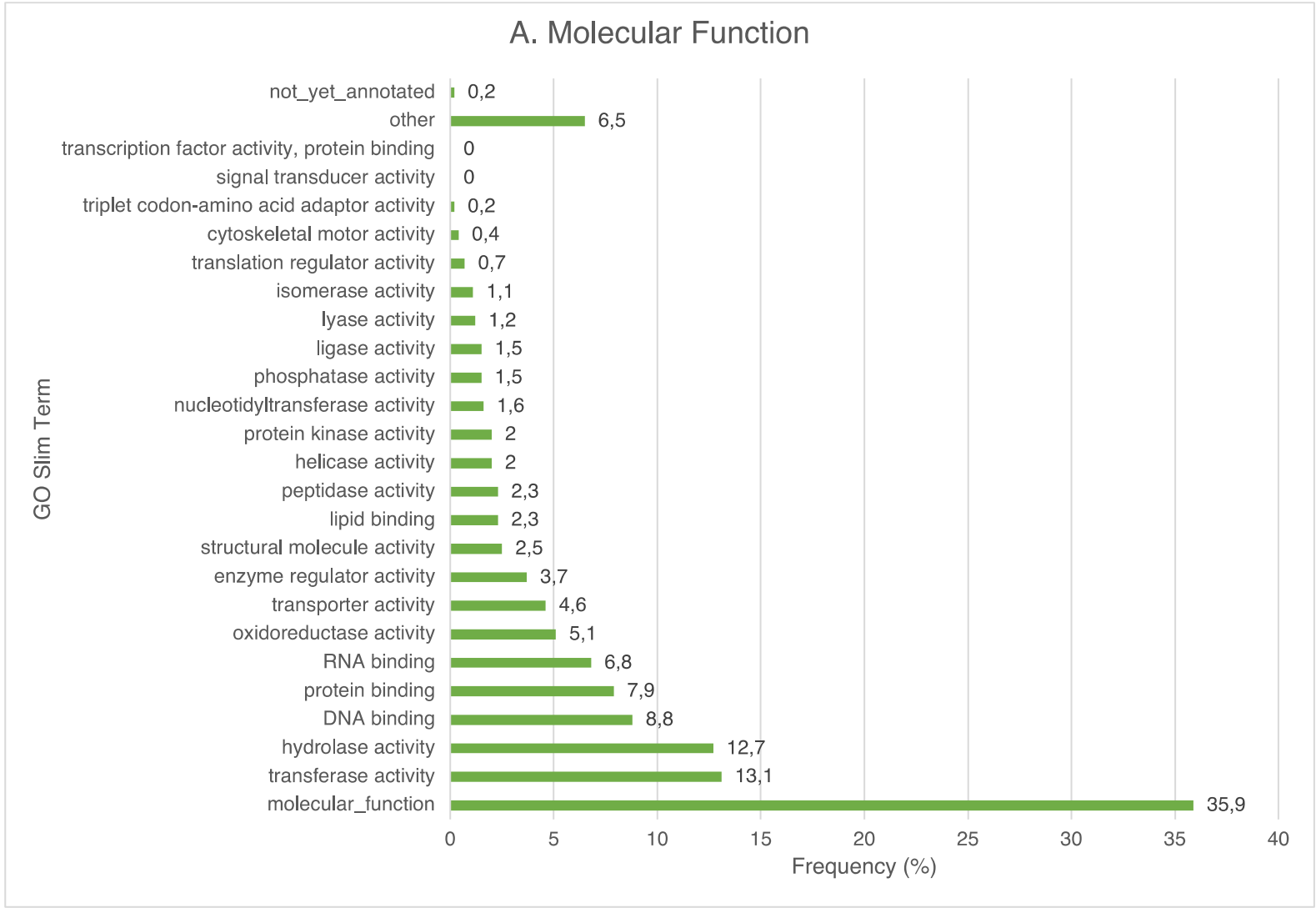


Figure 28: Frequencies of GO Slim terms mapped to genes with missense SNPs that were common to all blood isolates, regarding A. Molecular Function.

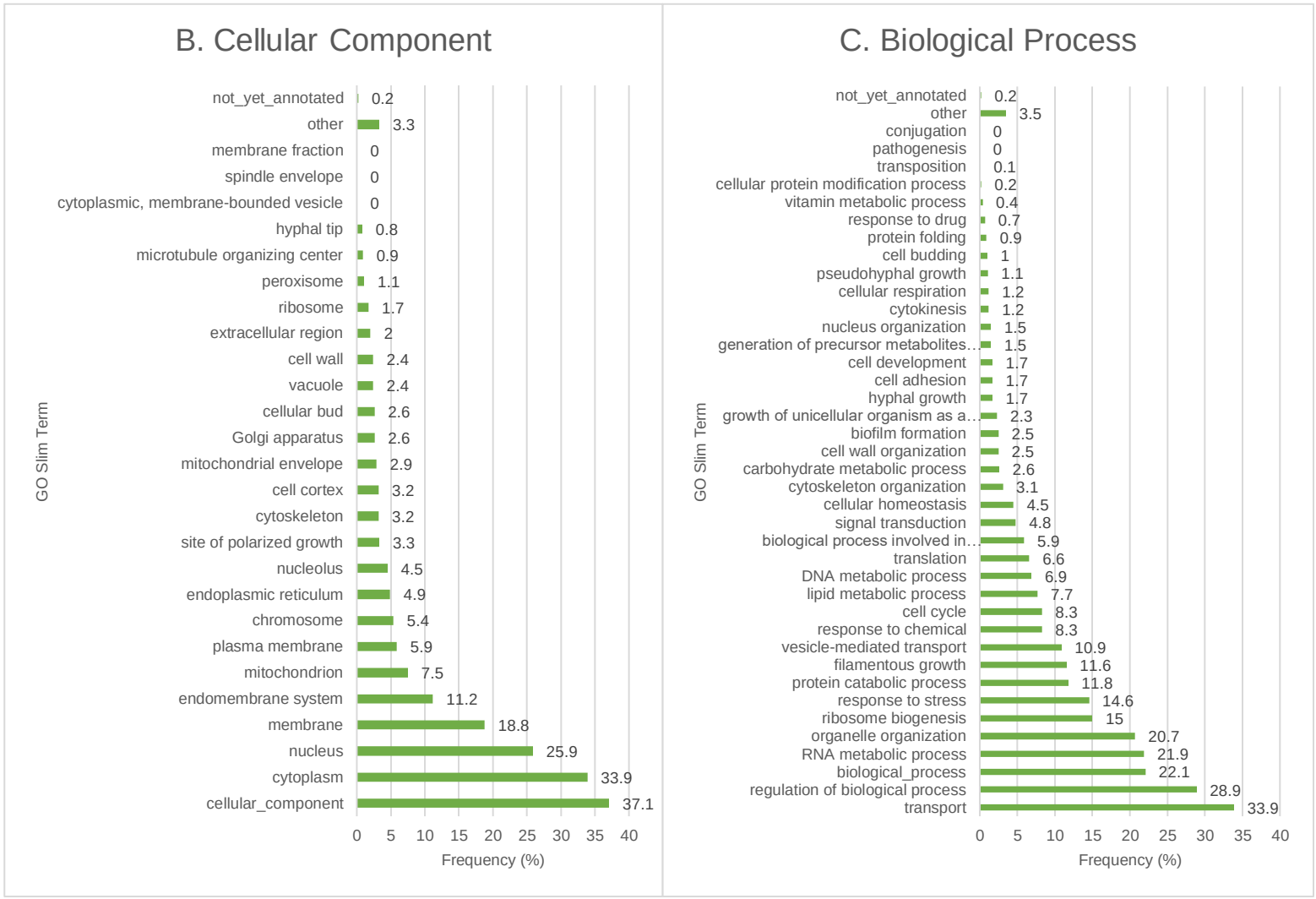


Figure 29: Frequencies of GO Slim terms mapped to genes with missense SNPs that were common to all blood isolates, regarding B. Cellular Component and C. Biologic Process.

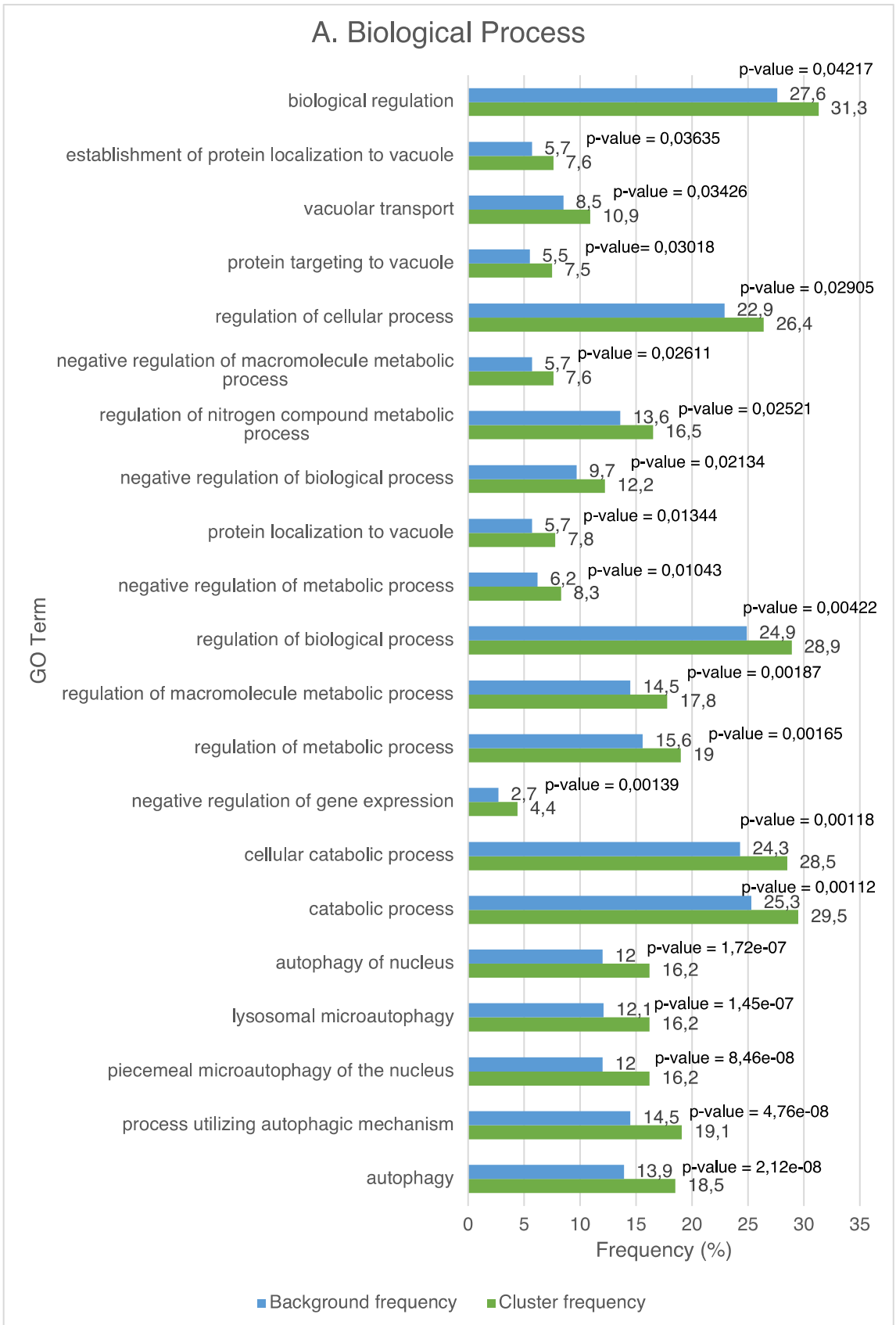


Figure 30: GO Term cluster frequencies of genes with missense SNPs that were common to all blood isolates, referent to A. Biological Process.

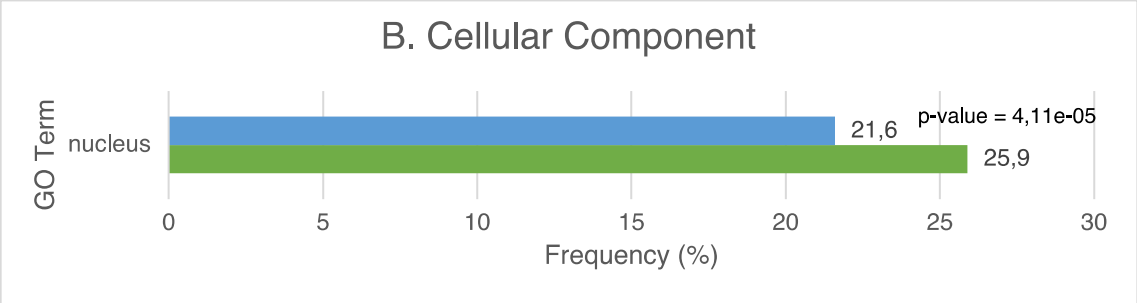


Figure 31: GO Term cluster frequencies of genes with missense SNPs that that were common to all blood isolates, referent to B. Cellular Component.

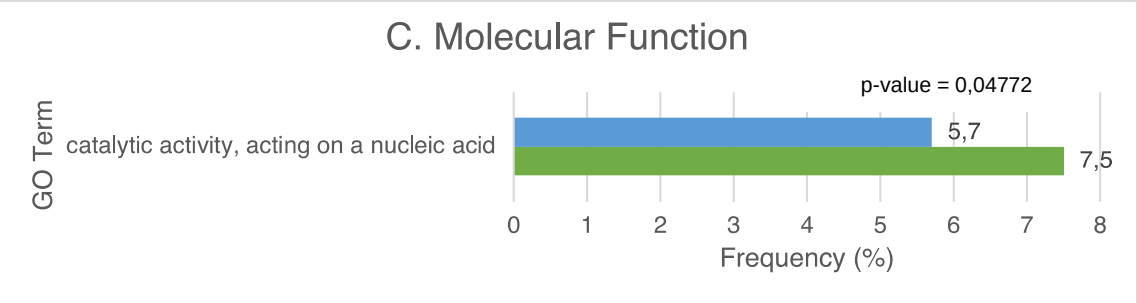


Figure 32: GO Term cluster frequencies of genes with missense SNPs that were common to all blood isolates, referent to C. Molecular Function.

4.3.3. Oral

Concerning genes with SNPs common to all oral isolates (n=1601), the GO Slim analysis mapped (i) genes coding for products involved in molecular functions, from which the hydrolase activity (12.6%), transferase activity (13.9%) and DNA binding (9.1%) were the most common; (ii) genes encoding for products related to cellular components, from which cytoplasm (32.8%), nucleus (24.7%) and membrane (18.1%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which RNA metabolic process (21.5%), organelle organization (20.6%) and response to stress (15.3%) were the most common (Figure 33 and 34).

GO Term analysis revealed genes annotated to the GO Terms DNA-binding transcription factor activity (p-value = 0.04881) and ATP binding (p-value = 0.03389) under the molecular functions ontology. Despite the p-value≤0.05 of the last result, its false discovery rate was >0.05 and so this association was not deemed significant (Figure 35).

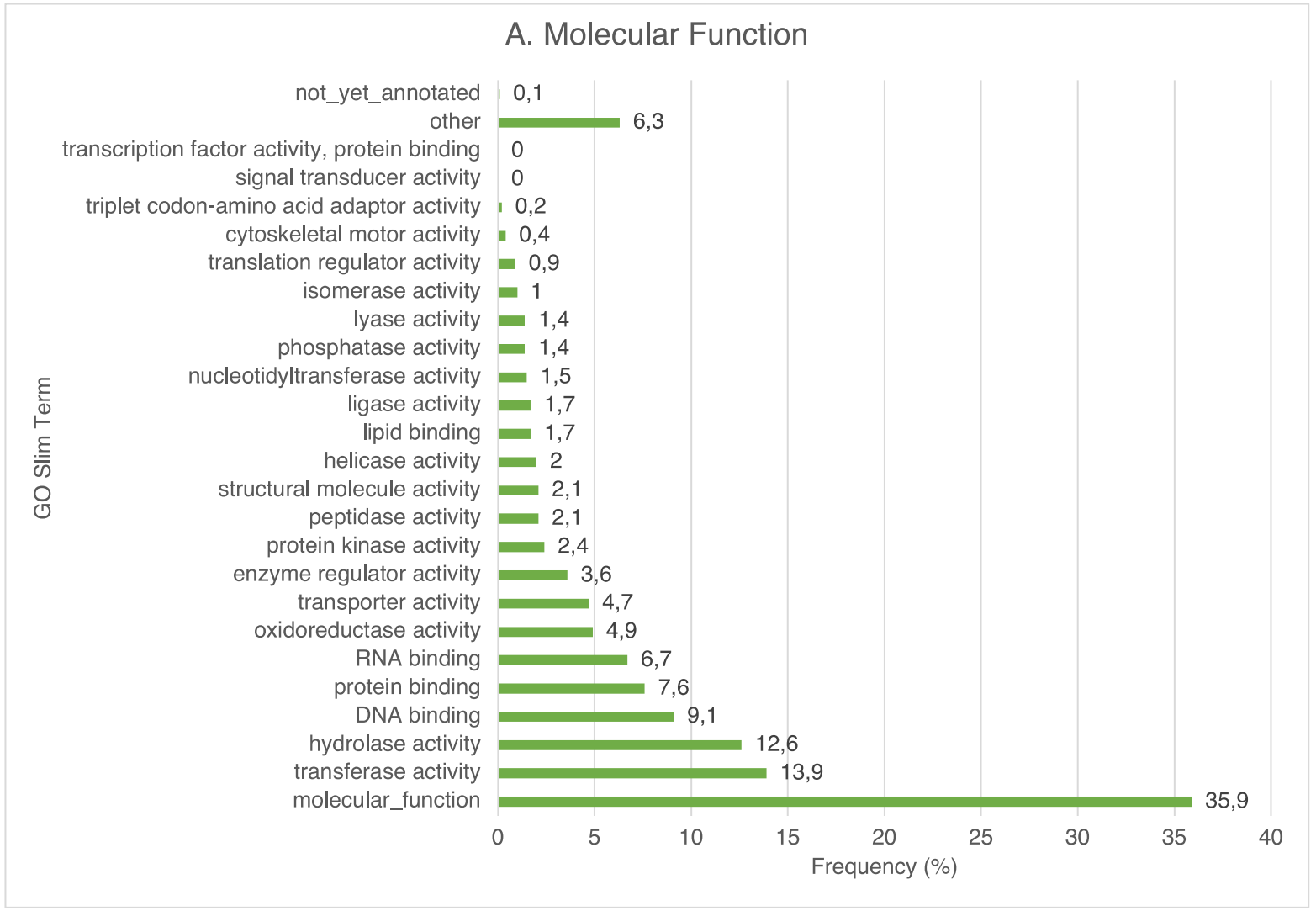


Figure 33: Frequencies of GO Slim terms mapped to genes with missense SNPs common to all oral isolates, regarding A. Molecular Function.

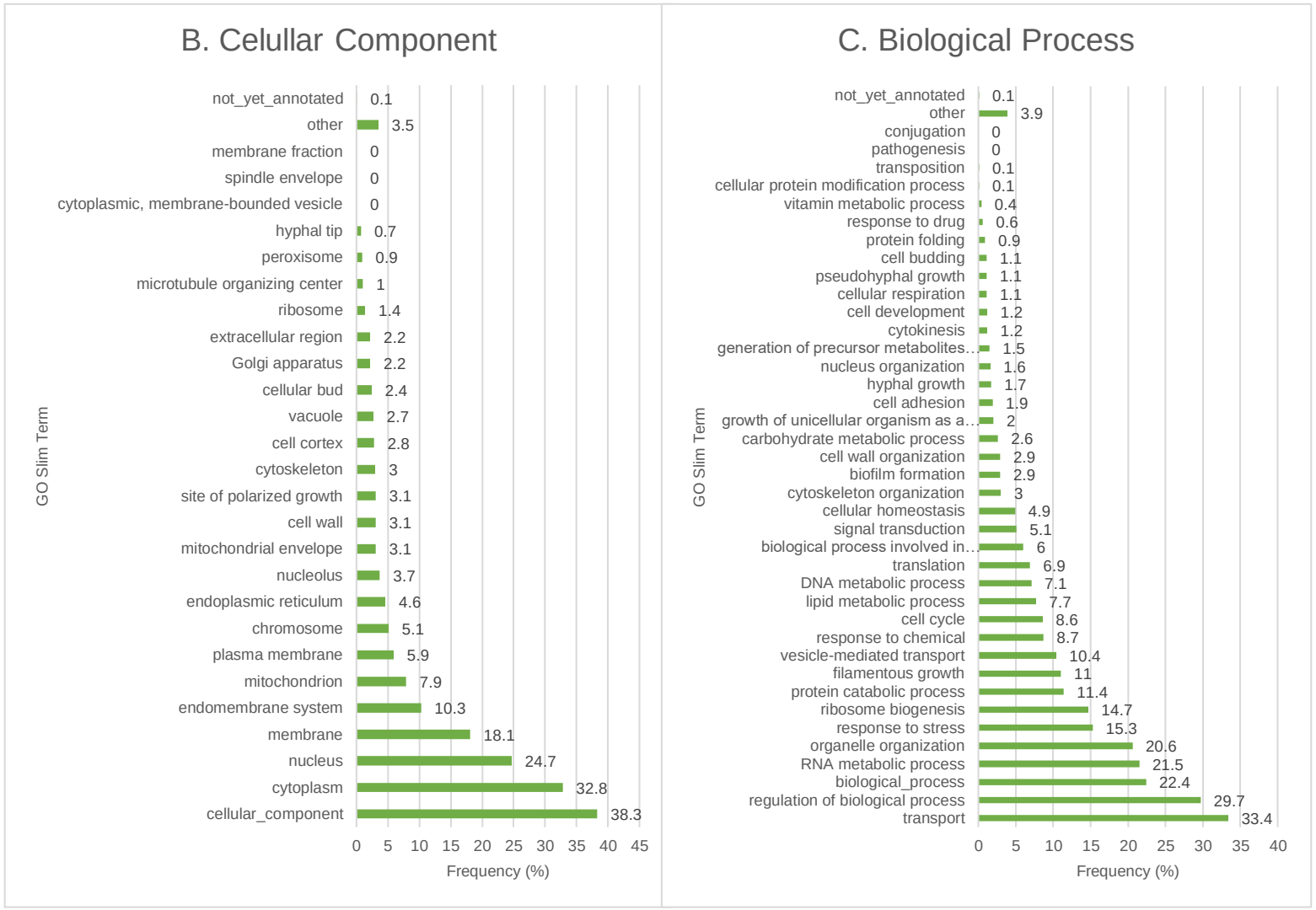


Figure 34: Frequencies of GO Slim terms mapped to genes with missense SNPs common to all oral isolates, regarding B. Cellular Component and C. Biological Process.

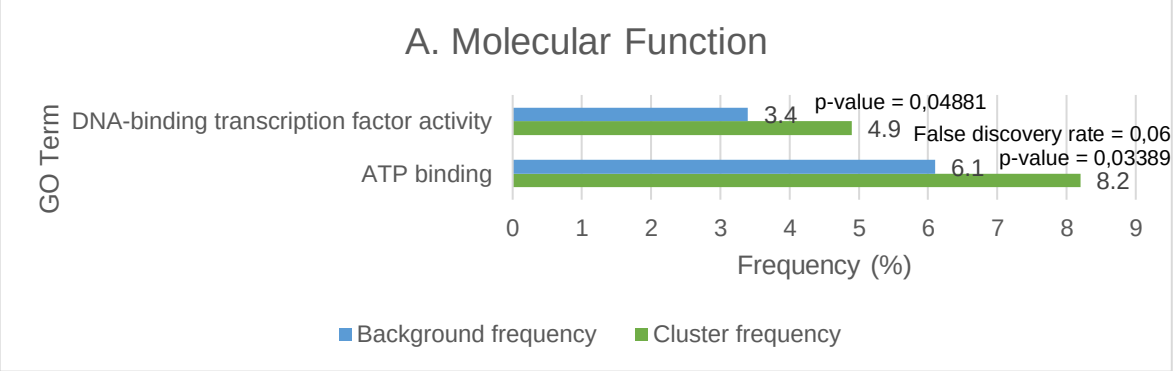


Figure 35: GO Term cluster frequencies of genes with missense SNPs common to all oral isolates, referent to A. Molecular Function.

4.3.4. Vaginal

In what concerns genes with SNPs common to all vaginal isolates (n=1282), the GO Slim analysis mapped (i) genes coding for products involved in molecular functions, from which the hydrolase activity (13.1%), transferase activity (12.6%) and DNA binding (9%) were the most common; (ii) genes encoding for products related to cellular components, from which cytoplasm (33.9%), nucleus (25.6%) and membrane (18.3%) were the most common; and (iii) genes encoding to products with a role in biological processes, from which organelle organization (21.3%), RNA metabolic process (21.2%) and response to stress (17.2%) were the most common (Figure 36 and 37).

GO Term analysis showed genes annotated to GO Terms of the cellular component and biological process ontologies. Nucleus (p-value = 0.04248) was the only GO Term found under cellular components (Figure 38). The GO Terms for biological processes significantly associated to the genes were cellular response to stress (p-value = 0.04445), cellular catabolic process (p-value = 0.04336), cellular response to external stimulus (p-value = 0.04104), cellular response to starvation (p-value = 0.03626), filamentous growth of a population of unicellular organisms (p-value = 0.01653), cell communication (p-value = 0.00591), autophagy of nucleus (p-value = 0.00134), lysosomal microautophagy (p-value =0.00104), piecemeal microautophagy of the nucleus (p-value = 0.00088), process utilizing autophagic mechanism (p-value = 0.00042), autophagy (p-value = 8.47e-05; Figure 39).

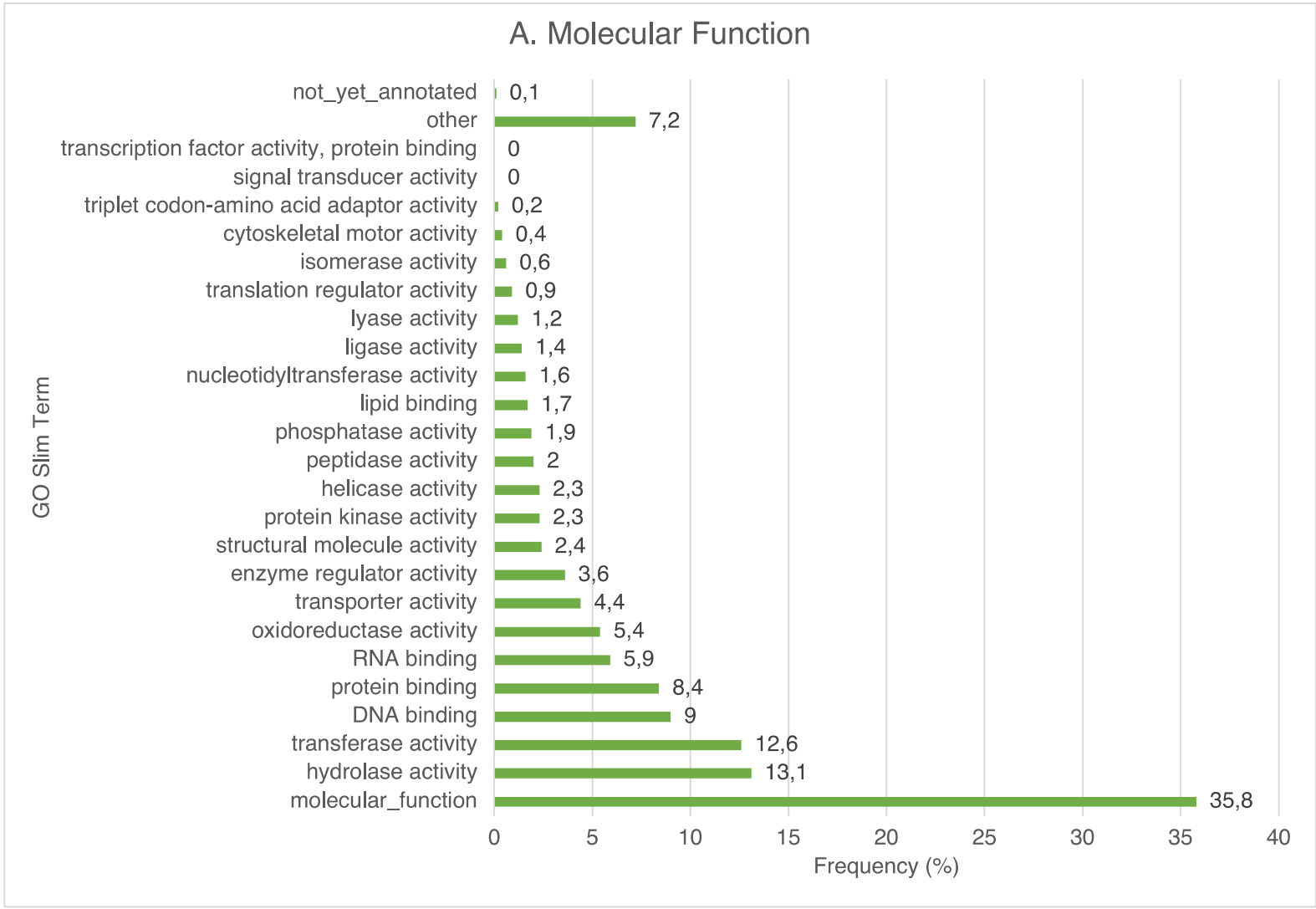


Figure 36: Frequencies of GO Slim terms mapped to genes with missense SNPs common to all vaginal isolates, regarding A. Molecular Function.

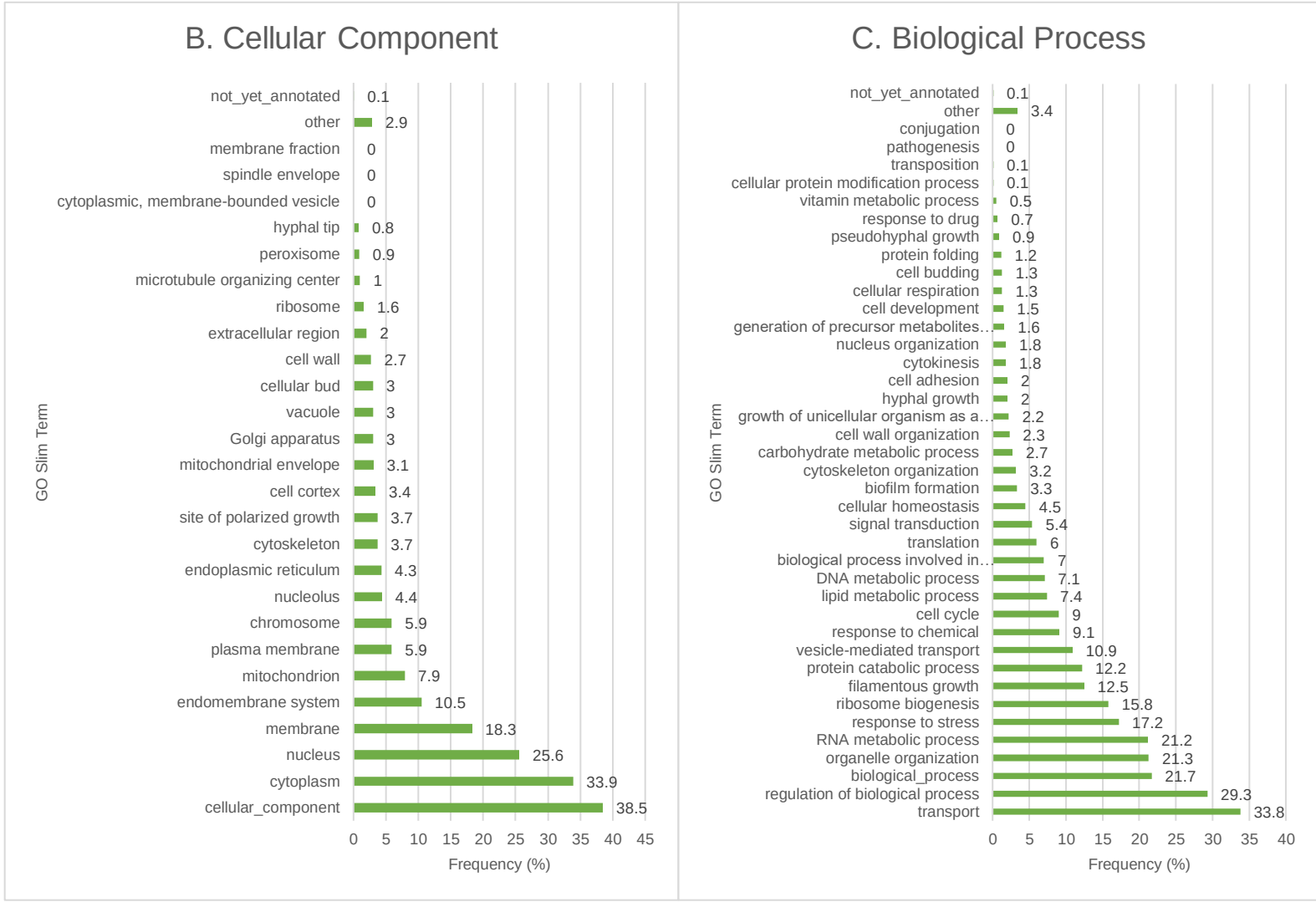


Figure 37: Frequencies of GO Slim terms mapped to genes with missense SNPs common to all vaginal isolates, regarding B. Cellular Component and C. Biologic Process.

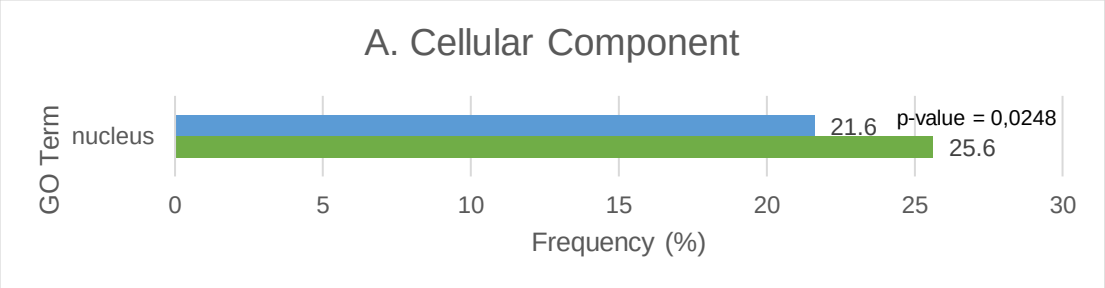


Figure 38: GO Term cluster frequencies of genes with missense SNPs common to all vaginal isolates, referent to A. Cellular Component.

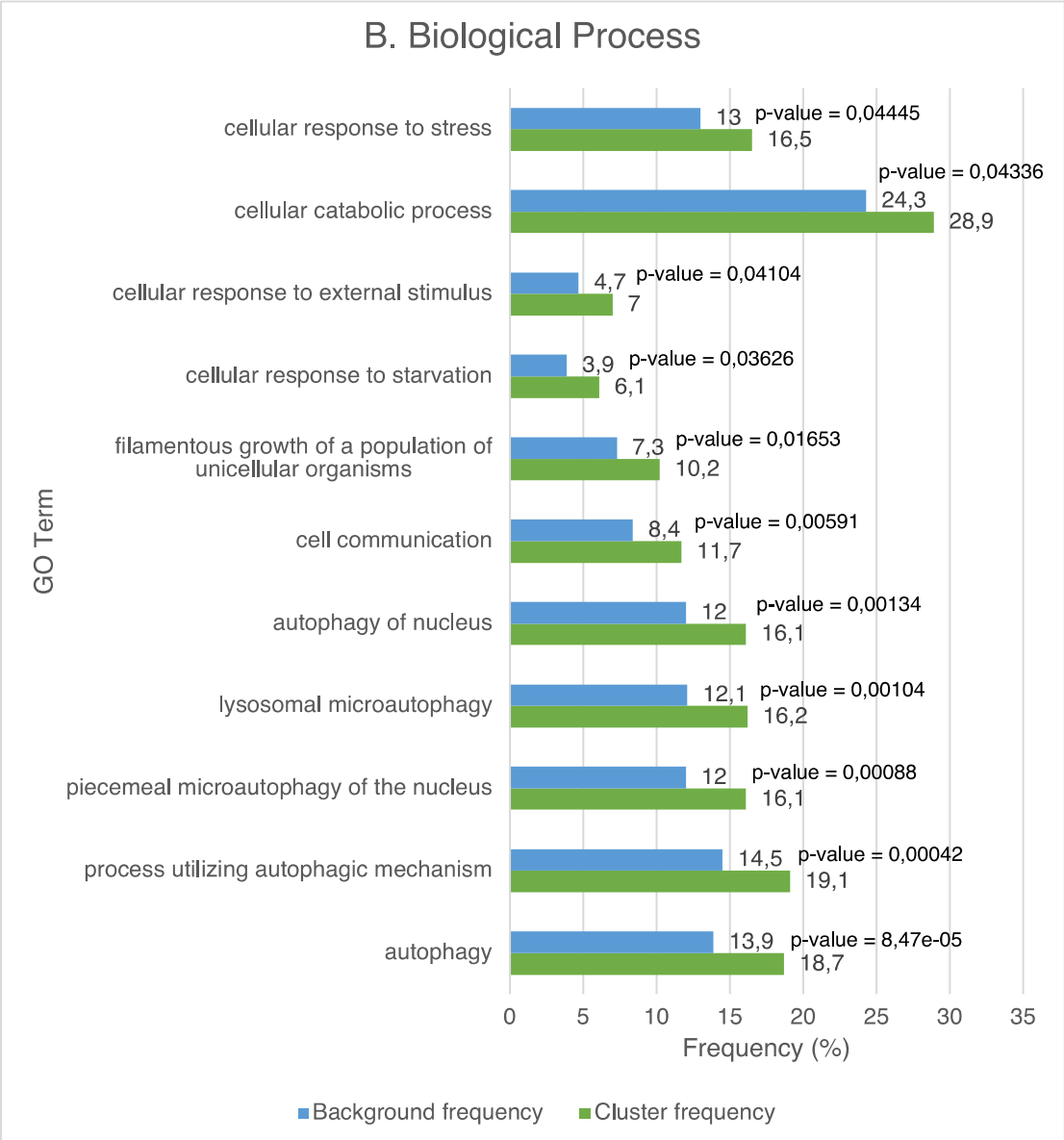


Figure 39: GO Term cluster frequencies of genes with missense SNPs common to all vaginal isolates, referent to B. Biological Process.

IV. DISCUSSION

The genus *Candida* harbors a few opportunistic pathogens species to humans, which represent one of the main causes of morbidity and mortality worldwide[25,56]. Among them, *C. albicans,* a major opportunistic species responsible for mucosal and invasive infections, is the most frequently encountered clinically and the most common cause of life-threatening systemic candidiasis[4,56].

1. Multi Locus Sequence Type (MLST)

The MLST analysis did not allow the identification of a sequence type for any of the *C. albicans* isolates in study. However, several isolates had close matches to sequence types previously found and available through the PubMLST database[57]. Given the genomic variability seen for the isolates in study, it could be that these isolates display allelic profiles for some/all the housekeeping genes within the MSLT scheme that have not yet been described. The nucleotide sequences of these genes have been submitted to the database and are under curation. Once validated, they will represent a significant addition to the database, in representation of Portuguese *C. albicans* clinical strains. Currently, there are 4950 isolates and 3704 sequence types available in the *Candida albicans* PUbMLST, from which only 31 isolates are from Portugal. The addition of our isolates to the database will be extremely valuable for the enrichment of this dataset and important for the surveillance of these organisms worldwide.

2. Single Nucleotide Polymorphisms (SNPs)

Between 75537 and 176406 SNPs were found in the genome of the *C. albicans* isolates under study, which depicts a higher value compared to the 62000 – 70000 SNPs reported for the *C. albicans* reference strain SC5314 genome[58–60], which might be related to the high number of homozygous SNPs seen for these isolates and this warrants further studies to determine LOH events. LOH events are linked to phenotypic changes and host adaptation, and this has been seen for clinical strains[61]. In this way, assuming LOH blocks are common among our isolates, they might be due to the adaptation of these isolates to the ecological niches they were collected from.

The isolate YP0129 was detected as a misidentified *C. albicans*, being posteriorly proven to be a *C. glabrata*. Unlike DNA sequencing methods, as used in this study, classical methods of species identification are unreliable and its use must be avoided, since *Candida*

species misidentification may compromise optimal antifungal therapy, or influence an experiment performance and its results analysis, research wise.

Regarding the *C. glabrata* isolate YP0129 SNPs identification, 80080 SNPs were identified, which was in accordance to the range of 0.04 to 7.23 SNPs/kb of *C. glabrata* reference strain CBS138 genome[62,63].

Among SNPs identification, some types among "others" SNPs were much higher in vaginal isolates than in the remaining niches. Some of these types were variants with a high impact, which are variants that cause changes in the genome sequences and, thus, in the phenotype[55]. Depending on the genes these SNPs occurred, these modifications may be important for the isolates' adaptation to and survival in the environment[55], and these results warrant further investigations.

GO analysis explored the GO terms of biological processes, molecular functions and cellular components associated with the genes with SNPs identified in our isolates in comparison with the *C. albicans* reference strain SC5314 genome.

The genes with missense SNPs common to all origins, demonstrated a significant frequency associated to the biological processes of autophagy, biologic regulation, process utilizing autophagy mechanism, regulation of cellular process and regulation of biological process. Among them, 94 genes were associated to these five GO terms. These 94 genes share several functions that may promote environmental adaptation and survival advantages to the clinical isolates in study. In fact, there were genes which have been described as direct or indirect stakeholders in the hyphal growth process (*C2_02730W_A*, *CDC24*, *FAB1*, *GPI15*, *HST7*, *SNF2* and *SRV2*)[17,54,64–75], a crucial process for the host's epithelial barriers penetration, allowing an optimized infection process in the four ecological niches in study. Also, genes involved in the morphogenesis process (*CDC14*, *RFX2* and *RVS167*)[17,54,76–79] including genes involved in the positive regulation of TORC1 (Target of rapamycin complex 1) signaling that acts as an essential regulator of *C. albicans* morphogenesis and nitrogen acquisition (*C2_03500W_A*, *C3_04460W_A*, *CR_03180W_A* and *GTR1*)[17,54,80–86], which are important in polymorphism virulence factor. Since the isolates in study were collected from human clinical samples the hyphal form of *C. albicans* might have been required to tissue infection, macrophage evasion, host-cell adhesion, and develop biofilm communities in order to adapt to the host niches. Additionally, there were intervening genes in the replication, transcription and translation processes of the cell cycle interphase (*C1_06910C_A*, *C6_03420W_A*, *C7_00840C_A*, *MBP1*, *RPM2*, *SNF2*, *SPT10*, *SRB8*, *TRY6* and *UME7*)[17,54,70,85,87–95] and in the remain cell cycle phases processes

(*C6_04120C_A*, *CDC14*, *CDC15* and *MHP1*)[17,54,77,78,96–99] . There were also genes related to virulence (*C6_03300C_A*, *C6_03320W_A*, *RFX2*, *RPM2*, *RVS167*, *SEC1*, *SRV2*, *VPS34*, *ZNC1*, *ZCF35*, *ZCF31*, *ZCF16* and *TAC1*)[17,54,68,69,76,79,94,95,100–107], which are important during *C. albicans* infection and adaptation to the host niches. Genes associated to antifungal resistance (*GPI15*, *NPR2*, *TSC11*, *ZCF16*, *TAC1*, *ZCF31*, *ZCF35* and *ZNC1*) were also identified[17,54,64,74,94,102,103,108–110]. Altogether, the accumulated genetic variation in these clinical strains and strains from the clinical environment might suggest that they are under selection to modulate their behavior so that strains become more adapted to the host and clinical environment.

Regarding the genes with missense SNPs exclusively found in *C. albicans* oral isolates, three genes were associated with the biological process of protein localization to chromosome (telomeric region). These genes were *C1_05380C_A*, *HST1* and *MEC1. C1_05380C_A* is an ORF of an unknown gene with telomeric DNA binding activity having a role in DNA double-strand break processing, DNA replication initiation, chromatin silencing at silent mating-type cassette, telomere capping, and shelterin complex localization, intervenient processes in the parasexual cycle of *C. albicans* and so, mechanisms that were probably important to enable diversity and promote the adaptation to the oral niche conditions[54,111]. The *HST1* gene is a *SIR2* paralog involved in regulation of white-opaque switching and codes for an enzyme named NAD-dependent histone deacetylase with a possible role in subtelomeric gene expression, cell cycle progression and chromosome stability, processes also related with the parasexual cycle of *C. albicans* and so, might have been important for the isolates adaptation to the oral niche conditions[18,112–114]. The *MEC1* gene codes for a highly conserved serine/threonine protein kinase of the PIKK (phosphatidyl inositol 3-kinase-like kinase) family, named MEC1[112,115]. MEC1 is a central component of the DNA damage checkpoint, which acts as a DNA damage sensor by activating the checkpoint signaling upon genotoxic stresses[112,115]. MEC1 also recognizes the substrate consensus sequence [ST]-Q needed to initiate the DNA repair and phosphorylates histone H2A to form H2AS128ph (gamma-H2A) required for the regulation of DNA damage response mechanism[112]. As so, mutations in *MEC1* affects the G1, S-phase and G2 checkpoints[115]. As these processes are also related to the parasexual cycle of *C. albicans*, they could have been important for the isolates adaptation to the oral niche conditions. Given the false discovery rate of 0.06 obtained for this analysis, the higher frequency of genes involved in the protein localization to chromosome (telomeric region) was not statistically significant. Although, results may be an indication that variants in these genes with these functions such as DNA binding activity, DNA damage checkpoint and

subtelomeric gene expression, are advantageous for isolates adaptation to the oral niche, due to their connection to the parasexual cycle. Perhaps with an increase of the number of samples in the analysis the result would be significant.

The genes with missense SNPs common to all *C. albicans* isolates from medical devices, demonstrated a significant frequency associated to the cellular component nucleus and an association to the molecular functions of DNA binding, catalytic activity acting on DNA and, catalytic activity, acting on a nucleic acid.

Regarding the cellular component annotated GO term, a substantial number of genes were identified as predicted to encode products associated to the nucleus in *C. albicans*, but only a few are proven to have these functions. These genes are *UTP8*, *C1_01160C_A*, *C1_06540_A*, *RAC1*, *C1_09670C_A*, *C4_05870C_A*, CR_01670W_A, *NUP82*, *C1_09840C_A*, *RIX7*, *TSA1B*, *C3_04590W_A* and *TSA1*[54].

Regarding the molecular function GO terms associated to the genes, a number of genes are predicted to encode products with DNA binding activity in *C. albicans*, but only a limited number of genes is proven to have this molecular function, according to the *Candida* genome database, such as *C1_07480C_A, C2_01420C_A, C2_02530W_A* and *C4_04620C_A* genes[54]. Based on the literature, other genes have been described to encode products with DNA binding activity, such as *RFX2*, *ZBP1*, *RFG1*, *EFG1*, *RBF1*, *WOR1* and *NRG1*[116–121]. In order to connect this molecular function to a more specific biological process, we identified shared genes between this association and the biologic process GO slim results. Among them, several genes were associated to the biofilm formation. Products with DNA binding activity have been described to repress or positively regulate several processes. This ability probably allowed the isolates to repress prejudicial processes and positively regulate advantageous processes such as biofilms formation for the isolates adhesion and survival in the surface of medical devices. Five genes were associated to the catalytic activity GO term under molecular function, such as *JAB1, C1_11290W_A, C1_00630W_A, CR_06030C_A* and *EST1* genes[54]. In order to connect catalytic activity to a more specific biological process, we identified shared genes between this association and the biologic process GO slim results. Among them, several genes were associated to filamentous growth. Filamentous growth is a crucial process to *C. albicans* pathogenicity and an important intervenient in biofilms formation. These results can be an indication that variants in genes with those functions are advantageous for the isolates adaptation and survival in medical devices and, thus, the clinical niche.

The genes with missense SNPs common to all *C. albicans* blood isolates exhibited a significant higher frequency associated to several biological processes related with autophagy, vacuoles, metabolic and catabolic processes, distinct processes of biologic regulation, one molecular function of catalytic activity, and one cellular component, nucleus.

Regarding the biological processes identified, autophagy, a highly conserved eukaryotic mechanism that enable cells to recycle cellular elements in order to survive under adverse conditions and dispensable for the virulence of *C. albicans*[122], has approximately 30 genes identified as key components to the autophagy process, collectively termed as *ATG* (Autophagy related genes)[122,123]. These genes involve the former genes named *APG* (autophagy), *AUT* (autophagy), *CVT* (cytoplasm-to-vacuole targeting), *GSA* (glucose-induced selective autophagy), *PAG* (peroxisome degradation), *PAZ* (pexophagy zeocinresistance), and *PDD* (peroxisome degradation-deficient)[123]. Apart from these genes, *Candida* genome database also includes *SSQ1*, *CCZ1*, *CR_03960C_A* and *VAM6* genes to this biological process[17]. Vacuoles are directly associated to the autophagic process, since autophagy facilitates the vacuole bulk degradation of cytoplasmic materials[124]. The genes *CR_03960C_A*, *ATG9* and *VAM6* shared GO terms in biological processes associated to vacuoles and the autophagy process. Additionally, *VPS4*, *BRO1*, *SNF7*, *VPS21*, *VPS41*, *C7_04240C_A* and *C7_03890C_A* were also genes associated to biological processes involving vacuoles[54]. Within the metabolic regulation process, *INO80* and *CR_04300W* were identified as genes encoding products with a role in this biological process[54]. According to literature *EFH1* is also mentioned as an intervenient in the metabolic regulation, as encodes APSES proteins that regulate *C. albicans* metabolism[125]. Lastly, several genes were significantly annotated to the catabolic process GO term, according to GO term analyses performed. Of these, only the *PHMS* gene is proven to have a role in this biological process[17]. Furthermore, the *STP2* gene is responsible for encoding amino acids involved in the catabolic process[126].

Additionally, the molecular function of catalytic activity, acting on a nucleic acid, was significantly associated with modified genes of these isolates. The same happened with the association with the GO term in cellular component, nucleus. Results may indicate that variants in genes involved in this molecular function and cellular component are advantageous for the isolates' adaptation to the blood ecological niche.

Altogether, the mutations of these genes in the isolates in study might enhance the isolates fitness to survive and adapt to the blood ecological niche and for their pathogenicity.

The genes with missense SNPs common to all *C. albicans* oral isolates, exhibited an association to the molecular functions of DNA binding and ATP binding.

Candida genome database identified several predicted genes associated to the ATP binding function. However, only *CDC60* and *SEC18* genes were proven to have this association[54]. According to the literature, *CDR1* and *CDR2* genes, are also associated to this molecular function, since they encode proteins that belong to the superfamily of ATP-binding cassette (ABC) transporters[127]. The ATP molecule is involved in several biological reactions. In particular, this molecule can be important in biofilm formation, virulence and parasexual cycle processes, all advantageous for the isolates adaptation to the oral niche [2]. In order to connect this molecular functions to a more specific biological process, we identified shared genes between this associations and the GO slim results. Among them, several genes were associated to both filamentous growth and biofilm formation. These results allow us to predict that variations in these genes may promote a biofilm formation optimization and consequently enable the adhesion, infection, and adaptation of *C. albicans* isolates to the oral niche.

The genes with missense SNPs common to all *C. albicans* vaginal isolates, exhibited a significant higher frequency associated to the biological processes of autophagy, cell communication, filamentous growth, and cellular response to stress, starvation, and external stimulus. Additionally, the cellular component nucleus was associated with higher frequency of mutated genes in these isolates.

The biological process of cell communication in yeast is represented as diffusible signals between cells which contribute to the community organization so that different regions of a community express different genes and adopt different cell fates, allowing the grow and adaptation of these isolates in the vaginal niche[128]. This process was associated to only one gene in *Candida* genome database, the *C1_01550C_A* gene[54].

The filamentous growth is induced upon exposure of *C. albicans* to a number of host conditions and is required for virulence[129] and according to literature the genes *INT1*, *CZF1*, *EFG1* and *RFG1* are associated to this process.

The cellular response process is associated to a list of predicted genes only. Of those, the genes *CR_01550C_A*, *C1_04970W_A*, *C6_02800W_A*, *C1_01040W_A*, *C5_04420W_A*, *OXR1* and *C4_06430C_A* are associated to the cellular response to stress; the genes *C4_05980C_A*, *C4_06450W_A* and *C3_03680W_A* to the cellular response to starvation; and the genes *C5_04420W_A*, *C2_05810W_A* and *CR_05460W_A* to the cellular response to external stimulus[54]. The vaginal niche is easily affected by external factors and holds challenging conditions such as its acidic pH, as so cellular responses

allow *C. albicans* isolates to adapt to these changes and conditions that act as stress, starvation and external stimulus.

The associations with GO terms of the cellular component, nucleus and the biological process of autophagy and variants in genes associated to these GO terms may be advantageous for the isolates adaptation to the conditions of vaginal niche.

V. CONCLUDING REMARKS AND FUTURE PERSPECTIVES

In this work we studied the genomic diversity of 76 *C. albicans* strains through whole genome sequencing (WGS) and genome analysis. With this, we aimed to identify genomic variations (SNPs) in clinical *C. albicans* strains which could be related to the adaptation to the different ecological niches the isolates were collected from (blood, vaginal, medical devices, oral). We also aimed to contribute to the epidemiologic surveillance of *C. albicans* strains, through MLST analysis.

Whole genome sequencing and further analysis lead us to discover that one of the isolates tested was not a *C. albicans*, but instead a *C. glabrata*. This misidentification corroborates the unreliability of classical methods which evidences the importance of DNA sequencing methods to identify *Candida* isolates to the species level in the clinical routine. *Candida* species misidentification must be prevented owing to the fact that it may compromise optimal antifungal therapy, or influence an experiment performance and its results analysis, research wise.

The distinct GO Term analyses performed revealed that the frequency of genes with missense SNPs involved in different molecular processes and biological functions, was significantly higher than the background frequency in those GO sets. In this way, there are groups of genes accumulating mutations in these isolates whose products are involved in mechanisms that might improve the isolates adaptability and survival ability in the environment they were collected from.

This study is an important first step for the characterization of *C. albicans* clinical isolates, and further investigations are warranted. These future analyses should include the calculation of the sample size necessary to distinguish samples from different niches, considering different factors involving quantitative and qualitative evaluations of the SNPs from each niche. Additionally, results point to large LOH events in these isolates' genomes as the number of homozygous SNPs among most of the samples was high. The detection of LOH events across the genomes of these isolates can be made by defining heterozygous and homozygous blocks as well as blocks with no variation in relation to the reference

genome as previously described[130–133]. LOH events can occur under environmental pressure and might alter the behavior of *C. albicans* during infection and modulate drug resistance[61]. Adding to this, it would also be beneficial to perform an analysis of copy number variation (CNVs) that could influence genetic expression. It has been found that CNVs occur in the presence of azole antifungal drugs. As so, it would be also interesting an analysis of CNVs that could influence the genetic expression, and had been described in the azoles resistance context[14,134]. Also, patterns of the susceptibility to the most used antifungals should be established for the isolates collection and verify if the SNPs or other genomic features associated to these antifungals resistance are present in the strain collection[33].

Nevertheless, this work allowed us to contribute to the in-house pipeline for variant analysis in *C. albicans* strains starting from WGS Illumina runs, that should be automated using bioinformatic code to facilitate future similar analyses. Also, the variants now called, together with structural ones such as CNV or LOH, can be added to phenotypic characterization of these strains, including sample niche origin, growth capacity, catabolic behavior, virulence or drug response, and used to carry out global genomic analysis like genome-wide association studies or polygenic risk scoring. This way, genetic variants of all sorts can be associated to such phenotypes in a much more efficient way. This will also make possible the inclusion of much more strains into the analysis, to gain statistical power so that new causal genetic variants can be discovered and used to increase our knowledge about putative therapeutic targets to tackle *Candida* sp. infections.

VI. BIBLIOGRAPHY

1. Nantel, A. The long hard road to a completed Candida albicans genome. *Fungal Genet. Biol.* **43**, 311–315 (2006).
2. Dadar, M. *et al.* Candida albicans - Biology, molecular characterization, pathogenicity, and advances in diagnosis and control – An update. *Microb. Pathog.* **117**, 128–138 (2018).
3. Todd, R. T., Wikoff, T. D., Forche, A. & Selmecki, A. Genome plasticity in candida albicans is driven by long repeat sequences. *Elife* **8**, 1–33 (2019).
4. D'Enfert, C. *et al. The impact of the fungus-host-microbiota interplay upon Candida albicans infections: Current knowledge and new perspectives*. *FEMS Microbiology Reviews* vol. 45 (2021).
5. Nobile, C. J. & Johnson, A. D. Candida albicans Biofilms and Human Disease. *Annu. Rev. Microbiol.* **69**, 71–92 (2015).
6. Kabir, M. A., Hussain, M. A. & Ahmad, Z. Candida albicans : A Model Organism for Studying Fungal Pathogens . *ISRN Microbiol.* **2012**, 1–15 (2012).
7. McManus, B. A. & Coleman, D. C. Molecular epidemiology, phylogeny and evolution of Candida albicans. *Infect. Genet. Evol.* **21**, 166–178 (2014).
8. Talapko, J. *et al.* Candida albicans-the virulence factors and clinical manifestations of infection. *J. Fungi* **7**, 1–19 (2021).
9. Cottier, F. & Hall, R. A. Face/Off: The Interchangeable Side of Candida Albicans. *Front. Cell. Infect. Microbiol.* **9**, (2020).
10. Sudbery, P., Gow, N. & Berman, J. The distinct morphogenic states of Candida albicans. *Trends Microbiol.* **12**, 317–324 (2004).
11. Noble, S. M., Gianetti, B. A. & Witchley, J. N. Candida albicans cell-type switching and functional plasticity in the mammalian host. *Nat. Rev. Microbiol.* **15**, 96–108 (2017).
12. Witchley, J. N. *et al.* Candida albicans Morphogenesis Programs Control the Balance between Gut Commensalism and Invasive Infection. *Cell Host Microbe* **25**, 432-443.e6 (2019).
13. Odds, F. C., Brown, A. J. P. & Gow, N. A. R. Candida albicans genome sequence: A platform for genomics in the absence of genetics. *Genome Biol.* **5**, 5–7 (2004).
14. Ene, I. V., Bennett, R. J. & Anderson, M. Z. Mechanisms of genome evolution in Candida albicans. *Curr. Opin. Microbiol.* **52**, 47–54 (2019).
15. Lephart, P. R. & Magee, P. T. Effect of the major repeat sequence on mitotic recombination in Candida albicans. *Genetics* **174**, 1737–1744 (2006).
16. Odds, F. C. and Epidemiology of. 67–79 (2010).
17. Skrzypek, M. S. *et al.* The Candida Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.* **45**, D592–D596 (2017).
18. Dunn, M. J. & Anderson, M. Z. To repeat or not to repeat: Repetitive sequences regulate genome stability in candida albicans. *Genes (Basel).* **10**, (2019).
19. Pevzner. 乳鼠心肌提取 HHS Public Access. *Physiol. Behav.* **176**, 139–148 (2017).
20. Prieto, D., Correia, I., Pla, J. & Román, E. Adaptation of Candida albicans to commensalism in the gut. *Future Microbiol.* **11**, 567–583 (2016).
21. Malik, A., Alves, M. & Grohmann, E. *Management of microbial resources in the environment*. *Management of Microbial Resources in the Environment* (2014). doi:10.1007/978-94-007-5931-2.
22. Romani, L., Bistoni, F. & Puccetti, P. Adaptation of Candida albicans to the host environment: The role of morphogenesis in virulence and survival in mammalian hosts. *Curr. Opin. Microbiol.* **6**, 338–343 (2003).
23. Bensasson, D. *et al.* Diverse lineages of Candida albicans live on old oaks. *Genetics* **211**, 277–288 (2019).

24. Schuster, J. & Fisher, B. Candidiasis. in 195-205.e3 (2021). doi:10.1016/B978-0-323-64198-2.00035-X.
25. Santos, G. C. d. O. *et al.* Candida infections and therapeutic strategies: Mechanisms of action for traditional and alternative agents. *Front. Microbiol.* **9**, 1–23 (2018).
26. Schuster, J. E. & Fisher, B. T. *Jennifer E. Schuster, MD, MSCI and Brian T. Fisher, DO, MPH, MSCE*. doi:10.1016/B978-0-323-64198-2.00035-X.
27. Kullberg, Bart Jan; Arendrup, M. C. Invasive Candidiasis. *New Engand J. Med.* **373**, 1445–1456 (2015).
28. Sanguinetti, M., Posteraro, B. & Lass-Flörl, C. Antifungal drug resistance among Candida species: Mechanisms and clinical impact. *Mycoses* **58**, 2–13 (2015).
29. Mayer, F. L., Wilson, D. & Hube, B. Mayer. *Benezit Dict. Artist.* 119–128 (2018) doi:10.1093/benz/9780199773787.article.b00119352.
30. Spampinato, C. & Leonardi, D. Candida infections, causes, targets, and resistance mechanisms: Traditional and alternative antifungal agents. *Biomed Res. Int.* **2013**, (2013).
31. Moudgal, V. & Sobel, J. Antifungals to treat Candida albicans. *Expert Opin. Pharmacother.* **11**, 2037–2048 (2010).
32. Houšť, J., Spížek, J. & Havlíček, V. Antifungal Drugs. *Metabolites* **10**, 106 (2020).
33. Lee, Y., Puumala, E., Robbins, N. & Cowen, L. E. Antifungal Drug Resistance: Molecular Mechanisms in Candida albicans and beyond. *Chem. Rev.* **121**, 3390–3411 (2021).
34. Welch, J. S. & Link, D. C. Genomics of AML : Clinical Applications of Next-Generation Sequencing. 30–35.
35. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177**, 70–84 (2019).
36. Pasquali, F. *et al.* Application of different DNA extraction procedures, library preparation protocols and sequencing platforms: impact on sequencing results. *Heliyon* **5**, e02745 (2019).
37. Illumina. Illumina DNA Prep. *Illumina* 1–4 (2020).
38. Muñoz, R., de las Rivas, B. & Curiel, J. A. *Identification Methods: Multilocus Sequence Typing of Food Microorganisms*. *Encyclopedia of Food Microbiology: Second Edition* vol. 2 (Elsevier, 2014).
39. Maiden, M. C. J. *et al.* MLST revisited: The gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **11**, 728–736 (2013).
40. Odds, F. C. & Jacobsen, M. D. Multilocus sequence typing of pathogenic Candida species. *Eukaryot. Cell* **7**, 1075–1084 (2008).
41. Bougnoux, M. *et al.* Collaborative Consensus for Optimized Multilocus Sequence Typing of Candida albicans. **41**, 5265–5266 (2003).
42. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, S. G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res 45 (D1); D592-D596. http://www.candidagenome.org/ (2017).
43. Draghici, S. No Title. (2018).
44. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
45. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. (2010).
46. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4

(2021).
49. Picard.
50. Auwera, G. A. V. der. *Genomics in the Cloud : Using Docker, GATK, and WDL in Terra*. (O'Reilly, 2020).
51. Caetano-Anolles, D. Hard-filtering germline short variants. *GATK/Technical Documentation/ Algorithms* https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants.
52. Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, X. L. and D. M. R. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Landes Biosci.*
53. Selmecki, A. M. *et al.* Polyploidy can drive rapid adaptation in yeast. *Nature* **519**, 349–351 (2015).
54. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, and S. G. Candida Genome Database.
55. Cingolani, P. Input & output files. *SNPeff* https://pcingola.github.io/SnpEff/se_inputoutput/#eff-field-vcf-output-files.
56. Martin, H., Kavanagh, K. & Velasco-torrijos, T. Targeting adhesion in fungal pathogen Candida albicans. **13**, 313–334 (2021).
57. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Res.* **3**, 1–20 (2018).
58. Forche, A., Magee, P. T., Magee, B. B. & May, G. Genome-wide single-nucleotide polymorphism map for Candida albicans. *Eukaryot. Cell* **3**, 705–714 (2004).
59. Muzzey, D., Schwartz, K., Weissman, J. S. & Sherlock, G. Assembly of a phased diploid Candida albicans genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biol.* **14**, (2013).
60. Ciudad, T., Hickman, M., Bellido, A., Berman, J. & Larriba, G. Phenotypic consequences of a spontaneous loss of heterozygosity in a common laboratory strain of Candida albicans. *Genetics* **203**, 1161–1176 (2016).
61. Liang, S. H. & Bennett, R. J. The impact of gene dosage and heterozygosity on the diploid pathobiont candida albicans. *J. Fungi* **6**, (2020).
62. Carreté, L. *et al.* Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans. *Curr. Biol.* **28**, 15-27.e7 (2018).
63. Ahmad, K. M. *et al.* Genome structure and dynamics of the yeast pathogen Candida glabrata. *FEMS Yeast Res.* **14**, 529–535 (2014).
64. Jain, P. *et al.* Modulation of azole sensitivity and filamentation by GPI15, encoding a subunit of the first GPI biosynthetic enzyme, in Candida albicans. *Sci. Rep.* **9**, 1–16 (2019).
65. Braun, S. & Matuschewski, K. Role of the ubiquitin-selective CDC48UFD1/NPL4 chaperone (segregase) in ERAD of OLE1 and other substrates. *EMBO J.* **21**, 615–621 (2002).
66. Ye, Y., Meyer, H. H. & Rapoport, T. A. The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol. *Nature* **414**, 652–656 (2001).
67. Rape, M. *et al.* Contents, Ed. Board + Forthc. articles. *Trends Biochem. Sci.* **30**, i (2005).
68. Bahn, Y. S. & Sundstrom, P. CAP1, an adenylate cyclase-associated protein gene, regulates bud-hypha transitions, filamentous growth, and cyclic AMP levels and is required for virulence of Candida albicans. *J. Bacteriol.* **183**, 3211–3223 (2001).
69. Rocha, C. R. C. *et al.* Signaling through adenylyl cyclase is essential for hyphal growth and virulence in the pathogenic fungus Candida albicans. *Mol. Biol. Cell* **12**, 3631–3643 (2001).

70. Mao, X., Cao, F., Nie, X., Liu, H. & Chen, J. The Swi/Snf chromatin remodeling complex is essential for hyphal development in Candida albicans. *FEBS Lett.* **580**, 2615–2622 (2006).
71. Clark, K. L. *et al.* Constitutive activation of the Saccharomyces cerevislae mating response pathway by a MAP kinase kinase from Candida albicans. *MGG Mol. Gen. Genet.* **249**, 609–621 (1995).
72. Priya, A. & Pandian, S. K. Piperine Impedes Biofilm Formation and Hyphal Morphogenesis of Candida albicans. *Front. Microbiol.* **11**, 1–18 (2020).
73. Cheng, S. *et al.* Identification of Candida albicans genes induced during thrush offers insight into pathogenesis. *Mol. Microbiol.* **48**, 1275–1288 (2003).
74. Nobile, C. J. *et al.* A recently evolved transcriptional network controls biofilm development in Candida albicans. *Cell* **148**, 126–138 (2012).
75. Augsten, M. *et al.* Defective hyphal induction of a Candida albicans phosphatidylinositol 3-phosphate 5-kinase null mutant on solid media does not lead to decreased virulence. *Infect. Immun.* **70**, 4462–4470 (2002).
76. Douglas, L. M., Martin, S. W. & Konopka, J. B. BAR domain proteins Rvs161 and Rvs167 contribute to Candida albicans endocytosis, morphogenesis, and virulence. *Infect. Immun.* **77**, 4150–4160 (2009).
77. Clemente-Blanco, A. *et al.* The Cdc14p phosphatase affects late cell-cycle events and morphogenesis in Candida albicans. *J. Cell Sci.* **119**, 1130–1143 (2006).
78. Jiménez, J., Cid, V. J., Nombela, C. & Sánchez, M. A single-copy suppressor of the Saccharomyces cerevisae late-mitotic mutants cdc15 and dbf2 is encoded by the Candida albicans CDC14 gene. *Yeast* **18**, 849–858 (2001).
79. Hao, B. *et al.* Candida albicans RFX2 encodes a DNA binding protein involved in DNA damage responses, morphogenesis, and virulence. *Eukaryot. Cell* **8**, 627–639 (2009).
80. Mayer, F. L. *et al.* The novel Candida albicans transporter Dur31 is a multi-stage pathogenicity factor. *PLoS Pathog.* **8**, (2012).
81. Nakashima, N., Noguchi, E. & Nishimoto, T. Saccharomyces cerevisiae putative G protein, Gtr1p, which forms complexes with itself and a novel protein designated as Gtr2p, negatively regulates the Ran/Gsp1p G protein cycle through Gtr2p. *Genetics* **152**, 853–867 (1999).
82. Tor-activating, T. C. crossm. (2017).
83. Tor-activating, T. C. crossm. **2**, 1–12 (2017).
84. Gonzalez, S. & Rallis, C. The TOR signaling pathway in spatial and temporal control of cell size and growth. *Front. Cell Dev. Biol.* **5**, 1–6 (2017).
85. Péli-Gulli, M. P., Sardu, A., Panchaud, N., Raucci, S. & De Virgilio, C. Amino Acids Stimulate TORC1 through Lst4-Lst7, a GTPase-Activating Protein Complex for the Rag Family GTPase Gtr2. *Cell Rep.* **13**, 1–7 (2015).
86. Zakikhany, K. *et al.* In vivo transcript profiling of Candida albicans identifies a gene essential for interepithelial dissemination. *Cell. Microbiol.* **9**, 2938–2954 (2007).
87. Hanaoka, N. *et al.* Identification of the putative protein phosphatase gene PTC1 as a virulence-related gene using a silkworm model of Candida albicans infection. *Eukaryot. Cell* **7**, 1640–1648 (2008).
88. Uppuluri, P. & Chaffin, W. L. J. Defining Candida albicans stationary phase by cellular and DNA replication, gene expression and regulation. *Mol. Microbiol.* **64**, 1572–1586 (2007).
89. Wang, J. M. *et al.* Intraspecies transcriptional profiling reveals key regulators of candida albicans pathogenic traits. *MBio* **12**, (2021).
90. Eriksson, P. R. *et al.* Global Regulation by the Yeast Spt10 Protein Is Mediated through Chromatin Structure and the Histone Upstream Activating Sequence Elements. *Mol. Cell. Biol.* **25**, 9127–9137 (2005).
91. Ernsting, B. R. & Dixon, J. E. The PPS1 gene of Saccharomyces cerevisiae codes

for a dual specificity protein phosphatase with a role in the DNA synthesis phase of the cell cycle. *J. Biol. Chem.* **272**, 9332–9343 (1997).
92. Bonhomme, J. *et al.* Contribution of the glycolytic flux and hypoxia adaptation to efficient biofilm formation by Candida albicans. *Mol. Microbiol.* **80**, 995–1013 (2011).
93. Hussein, B. *et al.* G1/S transcription factor orthologues Swi4p and Swi6p are important but not essential for cell proliferation and influence hyphal development in the fungal pathogen Candida albicans. *Eukaryot. Cell* **10**, 384–397 (2011).
94. Finkel, J. S. *et al.* Portrait of Candida albicans adherence regulators. *PLoS Pathog.* **8**, (2012).
95. Stribinskis, V., Heyman, H.-C., Ellis, S. R., Steffen, M. C. & Martin, N. C. Rpm2p, a Component of Yeast Mitochondrial RNase P, Acts as a Transcriptional Activator in the Nucleus. *Mol. Cell. Biol.* **25**, 6546–6558 (2005).
96. MHP1(1).pdf.
97. Bates, S. Candida albicans Cdc15 is essential for mitotic exit and cytokinesis. *Sci. Rep.* **8**, 1–11 (2018).
98. Uhl, M. A., Biery, M., Craig, N. & Johnson, A. D. Haploinsufficiency-based large-scale forward genetic analysis of filamentous growth in the diploid human fungal pathogen C.albicans. *EMBO J.* **22**, 2668–2678 (2003).
99. Irminger-Finger, I., Hurt, E., Roebuck, A., Collart, M. A. & Edelstein, S. J. MHP1, an essential gene in Saccharomyces cerevisiae required for microtubule function. *J. Cell Biol.* **135**, 1323–1339 (1996).
100. Blankenship, J. R., Fanning, S., Hamaker, J. J. & Mitchell, A. P. An extensive circuitry for cell wall regulation in Candida albicans. *PLoS Pathog.* **6**, (2010).
101. Raimund, E., Bruckmann, A., Wetzker, R. & Künkel, W. A phosphatidylinositol 3-kinase of Candida albicans: Molecular cloning and characterization. *Yeast* **16**, 933–944 (2000).
102. Schillig, R. & Morschhäuser, J. Analysis of a fungus-specific transcription factor family, the Candida albicans zinc cluster proteins, by artificial activation. *Mol. Microbiol.* **89**, 1003–1017 (2013).
103. Maicas, S. *et al.* In silico analysis for transcription factors with Zn(II)2C6 binuclear cluster DNA-binding domains in Candida albicans. *Comp. Funct. Genomics* **6**, 345–356 (2005).
104. Günther, J. *et al.* Generation and functional in vivo characterization of a lipid kinase defective phosphatidylinositol 3-kinase Vps34p of Candida albicans. *Microbiology* **151**, 81–89 (2005).
105. Sun, L. *et al.* phz1 contributes much more to phenazine-1-carboxylic acid biosynthesis than phz2 in Pseudomonas aeruginosa rpoS mutant. *J. Basic Microbiol.* **59**, 914–923 (2019).
106. Modify, B. *et al.* crossm.
107. Rollenhagen, C. *et al.* The Role of Secretory Pathways in Candida albicans Pathogenesis. *J. Fungi* **6**, 26 (2020).
108. Andes, D., Lepak, A., Pitula, A., Marchillo, K. & Clark, J. A simple approach for estimating gene expression in Candida albicans directly from a systemic infection site. *J. Infect. Dis.* **192**, 893–900 (2005).
109. Mukhopadhyay, K. *et al.* Membrane Sphingolipid-Ergosterol Interactions Are Important Determinants of Multidrug Resistance in Candida albicans. *Antimicrob. Agents Chemother.* **48**, 1778–1787 (2004).
110. Kukurudz, R. J. *et al.* Acquisition of cross-Azole tolerance and aneuploidy in Candida albicans strains evolved to posaconazole. *G3 Genes, Genomes, Genet.* **12**, (2022).
111. Marton, T., Feri, A., Commere, PH, Maufrais, C., d'Enfert, C., y Legrand, M. crossm Identification of Recessive Lethal Alleles in the Diploid Genome of a Candida albicans Laboratory Strain Unveils a. *Host-Microbe Biol.* 1–17 (2019).
112. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic*

Acids Res. **49**, D480–D489 (2021).
113. Zhao, G. & Rusche, L. N. Genetic Analysis of Sirtuin Deacetylases in Hyphal Growth of Candida albicans . *mSphere* **6**, 1–15 (2021).
114. Hnisz, D., Sehwarzmüller, T. & Kuchler, K. Transcriptional loops meet chromatin: A dual-layer network controls white-opaque switching in Candida albicans. *Mol. Microbiol.* **74**, 1–15 (2009).
115. Melanie Legrand, Christine L. Chan, Peter A. Jauert, D. T. K. The contribution of the S-phase checkpoint genes MEC1 and SGS1 to genome stability maintenance in Candida albicans. *ungal Genet. Biol.* **48**, 823–830 (2011).
116. Leng, P., Lee, P. R., Wu, H. & Brown, A. J. P. Efg1, a morphogenetic regulator in Candida albicans, is a sequence-specific DNA binding protein. *J. Bacteriol.* **183**, 4090–4093 (2001).
117. Glazier, V. E. EFG1, Everyone's Favorite Gene in Candida albicans: A Comprehensive Literature Review. *Front. Cell. Infect. Microbiol.* **12**, 1–12 (2022).
118. Ishii, Nobuya, Yamamoto, Matumi, Lahm, Hans-Wener, Lizumi, Shinnji, Yoshihara, Fumie, Nakayama, Hironobu, Arisawa, Mikio, Aoki, Y. A DNA-binding protein from. *Microbiology* **2**, 417–427 (1997).
119. Khalaf, R. A. & Zitomer, R. S. The DNA binding protein Rfg1 is a repressor of filamentation in Candida albicans. *Genetics* **157**, 1503–1512 (2001).
120. Banoth, B. *et al.* ZBP1 promotes fungi-induced inflammasome activation and pyroptosis, apoptosis, and necroptosis (PANoptosis). *J. Biol. Chem.* **295**, 18276–18283 (2020).
121. Lohse, M. B., Zordan, R. E., Cain, C. W. & Johnson, A. D. Distinct class of DNA-binding domains is exemplified by a master regulator of phenotypic switching in Candida albicans. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14105–14110 (2010).
122. Palmer, G. E., Askew, D. S. & Williamson, P. R. The diverse roles of autophagy in medically important fungi. *Autophagy* **4**, 982–988 (2008).
123. Klionsky, D. J. *et al.* A unified nomenclature for yeast autophagy-related genes. *Dev. Cell* **5**, 539–545 (2003).
124. Su, T. *et al.* Autophagy: An Intracellular Degradation Pathway Regulating Plant Survival and Stress Response. *Front. Plant Sci.* **11**, 1–16 (2020).
125. Doedt, T. *et al.* APSES proteins regulate morphogenesis and metabolism in Candida albicans. *Mol. Biol. Cell* **15**, 3167–3180 (2004).
126. Miramón, P., Pountain, A., van Hoof, A. & Lorenz, M. C. crossm The Paralogous Transcription Factors Stp1 and Stp2 of Acquisition and Host Interaction. 1–18 (2020).
127. Franz, R., Michel, S. & Morschhäuser, J. A fourth gene from the Candida albicans CDR family of ABC transporters. *Gene* **220**, 91–98 (1998).
128. Honigberg, S. M. Cell signals, cell contacts, and the organization of yeast communities. *Eukaryot. Cell* **10**, 466–473 (2011).
129. Johnson, D. K. and A. D. Induction of the Candida albicans Filamentous Growth Program by Relief of Transcriptional Repression: A Genome-wide Analysis. *Mol. Biol. Cell* **Vol. 16**, 2903–2912 (2005).
130. Mixão, V., Saus, E., Boekhout, T. & Gabaldón, T. Extreme diversification driven by parallel events of massive loss of heterozygosity in the hybrid lineage of Candida albicans. *Genetics* **217**, (2021).
131. Pryszcz, L. P., Németh, T., Gácser, A. & Gabaldón, T. Genome comparison of candida orthopsilosis clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol. Evol.* **6**, 1069–1078 (2014).
132. Pryszcz, L. P. *et al.* The Genomic Aftermath of Hybridization in the Opportunistic Pathogen Candida metapsilosis. *PLoS Genet.* **11**, 1–29 (2015).
133. Mixão, V. *et al.* Whole-genome sequencing of the opportunistic yeast pathogen candida inconspicua uncovers its hybrid origin. *Front. Genet.* **10**, 1–11 (2019).
134. Du Toit, A. Copy-number variation. *Nat. Rev. Microbiol.* **18**, 542 (2020).

VII. ANNEXES

Annex A: Genomic DNA (gDNA) concentrations per isolate.

Isolates	gDNA ng/µl (dil 1/10)	gDNA libraries ng/µl
YP0001	24	16,5
YP0048	10,8	14,7
YP0067	13,1	15,8
YP0162	15,2	15,7
YP0176	17,3	14,6
YP0188	14,6	13,2
YP0196	21,4	18,3
YP0211	17,1	15,8
YP0306	14,2	14,2
YP0382	8,17	18,3
YP0384	14,8	16,2
YP0386	13,5	18
YP0391	12,6	17,5
YP0016	14,6	17,1
YP0019	14,8	19,2
YP0024	15,9	18,7
YP0026	16	15,4
YP0028	18,8	16,9
YP0031	13,3	17,2
YP0034	14	17
YP0035	15,1	15,8
YP0036	10,6	16,6
YP0054	22,8	13,6
YP0100	21,5	16,7
YP0101	12,5	15,8
YP0103	18,2	16,1
YP0045	17,3	16,2
YP0047	14,5	16
YP0569	17,8	15,5
YP0399	26,5	16,7
YP0577	18,2	15,6
YP0639	15,2	16
YP1130	12	17,7
YP0760	14	17,2

YP0581	12,5	11,3
YP0801	1,72	14,3
YP0631	12,5	12,3
YP0037	14,8	14,8
YP0057	14,2	18,1
YP0364	13,9	11
YP0070	6,89	16,8
YP0537	17,3	18
YP0363	24,6	17,1
YP0493	10,7	15,8
YP0474	18,1	17,6
YP0126	8,63	14,5
YP0083	16,8	17,7
YP0050	30,2	16,8
YP0061	8,38	17,3
YP0087	15,5	19
YP0058	19	18,3
YP0108	11,7	17,7
YP0098	2,51	17,6
YP0159	9,02	18,4
YP0131	13,8	17,5
YP0129	5,43	18,4
YP0326	6,49	16,9
YP0144	11,6	18,3
YP0051	12,6	17,9
YP0132	17,8	18,6
YP0093	14,9	17
YP0081	18,3	16,9
YP0355	15,5	16,9
YP0167	17,9	20
YP0232	3,91	18,3
YP0233	20,4	18,6
YP0344	11,9	14,9
YP0316	21,1	17,6
YP0362	11,1	17,5
YP0094	4,48	17,2
YP0095	15,4	19
YP0200	16,4	16,8
YP0097	17,6	18
YP0114	13,2	20,7

YP0115	16,4	17,6
YP0392	8,72	18

Annex B: Number of SNPs before and after filtering for alternate allele support and allelic frequency as described in[53].

Isolates	Number of SNPs before filter	Number of SNPs after filter
YP0001	147604	147312
YP0016	146478	146234
YP0019	144755	144383
YP0024	134126	133812
YP0026	143426	143139
YP0028	128651	128381
YP0031	80817	80693
YP0034	130365	130089
YP0035	138356	138008
YP0036	123817	123540
YP0037	153131	152590
YP0045	132188	131526
YP0047	139061	138409
YP0048	130800	130482
YP0050	138858	138557
YP0051	81332	81181
YP0054	128492	128254
YP0057	130153	129857
YP0058	83630	83520
YP0061	129582	129134
YP0067	82453	82328
YP0070	151851	151253
YP0081	131066	130818
YP0083	145289	144976
YP0087	136383	136057
YP0093	141255	140886
YP0094	126190	125872
YP0095	126173	125866
YP0097	135560	135255
YP0098	138207	137962
YP0100	89978	89849
YP0101	124615	124318
YP0103	134609	134175
YP0108	141952	141616

YP0114	131410	131096
YP0115	135780	135476
YP0126	89653	89554
YP0131	145083	144800
YP0132	141980	141626
YP0144	80779	80644
YP0159	136933	136661
YP0162	147616	147275
YP0167	130396	130111
YP0176	147589	147197
YP0188	142814	142488
YP0196	148827	148370
YP0200	80493	80374
YP0211	134441	134224
YP0232	91020	90930
YP0233	144589	144250
YP0306	75665	75546
YP0316	150645	150358
YP0326	133189	132846
YP0344	138744	138475
YP0355	134690	134306
YP0362	86292	86159
YP0363	133913	133616
YP0364	165065	164630
YP0382	130559	130213
YP0384	135015	134664
YP0386	129187	128863
YP0391	132689	132371
YP0392	127132	126852
YP0399	123332	123118
YP0474	149099	148834
YP0493	146877	146642
YP0537	136318	136033
YP0569	153243	152663
YP0577	132480	131707
YP0581	176862	176415
YP0631	168279	167790
YP0639	131822	131191
YP0760	145837	145499
YP0801	157941	157358

YP1130	149843	149452

Annex C: Number of homozygous and heterozygous SNPs per sample.

Sample	Homozygous SNPs	Heterozygous SNPs
YP0001	96863	50440
YP0016	93731	52268
YP0019	96027	47214
YP0024	76903	55543
YP0026	94520	47441
YP0028	67208	59788
YP0031	57888	21929
YP0034	84091	44500
YP0035	96666	39763
YP0036	61392	60572
YP0037	114290	37867
YP0045	81906	47976
YP0047	99908	36823
YP0048	73942	54904
YP0050	87149	49780
YP0051	65325	14951
YP0054	80629	46131
YP0057	81236	47117
YP0058	74557	8026
YP0061	69346	58397
YP0067	67767	13619
YP0070	109893	40941
YP0081	72155	57246
YP0083	93720	49440
YP0087	93202	41269
YP0093	79569	59451
YP0094	72664	51703
YP0095	72663	51719
YP0097	81532	51967
YP0098	90230	46001
YP0100	81380	7464
YP0101	64946	57817
YP0103	80900	51599
YP0108	89422	50899
YP0114	73362	56298
YP0115	92764	41133

YP0126	83058	5501
YP0131	91749	52751
YP0132	89462	50846
YP0144	66994	12756
YP0159	94139	40899
YP0162	94642	50846
YP0167	71887	56825
YP0176	94534	50852
YP0188	84680	56001
YP0196	101877	45173
YP0200	65933	13555
YP0211	88170	44537
YP0232	84448	5420
YP0233	87828	54404
YP0306	61165	13485
YP0316	102597	47495
YP0326	76409	54980
YP0344	80603	56585
YP0355	86639	46073
YP0362	75302	9886
YP0363	80586	51523
YP0364	139604	24318
YP0382	78723	49979
YP0384	92523	40583
YP0386	68937	58203
YP0391	86949	43934
YP0392	66901	58583
YP0399	57440	64286
YP0474	97873	50668
YP0493	97234	48096
YP0537	83384	51185
YP0569	123274	27615
YP0577	80521	49526
YP0581	148943	26457
YP0631	139855	27192
YP0639	79048	50716
YP0760	94653	49528
YP0801	126123	30658
YP1130	100386	48779