



Universidade do Minho

Escola de Engenharia

Departamento de Informática

Ana Patrícia Mota de Oliveira

**Development of a Framework
for Identification of *Candida* Species
and Detection of Antifungal Resistance**

December 2019



Universidade do Minho

Escola de Engenharia

Departamento de Informática

Ana Patrícia Mota de Oliveira

**Development of a Framework
for Identification of *Candida* Species
and Detection of Antifungal Resistance**

Master Dissertation

Master Degree in Bioinformatics

Dissertation supervised by

Miguel Francisco Almeida Pereira Rocha

Ana Sofia da Quinta e Costa Neves de Oliveira

December 2019

Despacho RT - 31 /2019 - Anexo 3

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição
CC BY

<https://creativecommons.org/licenses/by/4.0/>

ACKNOWLEDGEMENTS

Firstly, I would to thank my supervisor Sofia Oliveira for the opportunity to work in collaboration with her and for all the inspiring conversations. To my supervisor, professor Miguel Rocha, thank you for all technical support.

Secondly, I would to thank my mom, my dad and my little sister, Esmeralda, Jorge and Matilde, for all the emotional support, and all the moments of relaxing and happiness. And mainly, for the social values transmitted during my entire life, and the education that allowed me to get here.

Thirdly, I would to thank my friends for all the support, moments and smiles.

A special thanks to my boyfriend Tiago. For all conversations related to software engineering and computer science. For all persistence and patience. For all happiness, inspiration, motivation and support. For all the love. Without you, it wouldn't have been possible to do this work. I love you.

I would like to dedicate this master thesis to my grandmother Fernanda. Thanks for the care since of my three months of life. For all the pride you've always shown. And for all the moments, conversations, smiles and teachings. Now, looking out for me from the stars.

Despacho RT - 31 /2019 - Anexo 4**STATEMENT OF INTEGRITY**

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Invasive fungal infections affect millions of people around the world, and they are associated with high mortality rates, which in some cases can reach 90%. These infections have a high incidence in immunosuppressed patients. The most frequent pathogens that cause this type of infections belong to the *Candida* genus. Most of them have the ability to develop molecular mechanisms to fight antifungal treatment, which leads to the increase of antifungal resistance.

The remarkable evolution in speed and cost of sequencing technologies, such as Next-Generation Sequencing, allowed whole genomes to be sequenced in a single run, becoming a solution with advantages over traditional methods. In the last years, the number of sequenced genomes increased dramatically. However, the number of bioinformatics systems to treat and analyze these data did not follow this growth, especially in clinical mycology.

So, to address the lack of bioinformatics solutions applied to clinical mycology, a framework has been developed to allow the identification and detection of antifungal resistance in different species of the *Candida* genus. The framework is prepared to receive sequencing files from Illumina platforms and has the ability to identify a high range of fungi species. The detection of antifungal resistance is available for *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*, regarding ERG11 and FKS1 genes. The framework developed can be a great solution for any microbiology laboratory equipped with Next-Generation Sequencing platforms.

Keywords: Antifungal resistance detection, *Candida* species, Fungi identification, Next-Generation Sequencing

RESUMO

Em todo o mundo milhares de pessoas são afetadas por infecções fúngicas invasivas. Estas infecções estão associadas a altas taxas de mortalidade, que em alguns casos pode ultrapassar os 90%. A maioria das pessoas infetadas são doentes imunodeprimidos. O patógeno mais frequente causador deste tipo de infecções pertence às espécies do género *Candida*. A maioria tem a capacidade de desenvolver mecanismos moleculares para o tratamento com antifúngicos, o que tem levado a um aumento da resistência aos antifúngicos.

Com o crescimento das tecnologias de sequenciação, como é o caso de *Next-Generation Sequencing*, permitiu a sequenciação do genoma inteiro numa única reação, tornando-se uma solução com grandes vantagens em relação aos métodos tradicionais. Nos últimos anos, o número de genomas sequenciados aumentou exponencialmente. No entanto, o número de sistemas bioinformáticos de tratamento e análise deste tipo de dados não seguiu a mesma tendência, especialmente em micologia clínica.

De forma a combater a ausência de soluções bioinformáticas aplicadas à micologia clínica, foi desenvolvida uma *framework* que permite a identificação e a deteção de resistência aos antifúngicos em espécies do género *Candida*. A *framework* recebe ficheiros de sequenciação provenientes da plataforma Illumina e tem a capacidade de identificar um grande número de espécies de fungos. A deteção de resistência aos antifúngicos está disponível para *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* e *Candida tropicalis* relativamente aos genes ERG11 e FKS1. A *framework* desenvolvida poderá ser uma boa solução para qualquer laboratório de microbiologia equipado com plataformas de *Next-Generation Sequencing*.

Palavras-Chave: Espécies de *Candida*, Identificação de fungos, *Next-Generation Sequencing*, Resistência aos antifúngicos

CONTENTS

1	INTRODUCTION	1
1.1	Context and Motivation	1
1.2	Objectives	2
1.3	Document Organization	3
2	THE CANDIDA GENUS	4
2.1	Invasive Fungal Infections	4
2.2	Overview of the Candida Genus	5
2.2.1	The Biology	5
2.2.2	Mechanisms of Infection	6
2.2.3	Epidemiology	6
2.3	Methods for Identification	8
2.4	Antifungal Agents	9
2.4.1	Pyrimidines	9
2.4.2	Polyenes	10
2.4.3	Azoles	10
2.4.4	Echinocandins	10
2.5	Antifungal Resistance and Mechanisms	10
2.5.1	Against Pyrimidines	11
2.5.2	Against Polyenes	11
2.5.3	Against Azoles	12
2.5.4	Against Echinocandins	13
2.6	Methods for Detection of Antifungal Resistance	14
2.7	Biological Databases	15
2.7.1	GenBank	16
2.7.2	Sequence Read Archive	16
2.7.3	Mycology Antifungal Resistance Database	16
3	NEXT-GENERATION SEQUENCING	17
3.1	Overview of Next-Generation Sequencing	17
3.2	Technical Procedures	20
3.2.1	Library Preparation	21
3.2.2	Base Calling	21
3.3	Sequence Data Formats	22
3.3.1	FASTQ	22
3.3.2	FASTA	23

3.4	Data Preprocessing	24
3.4.1	Quality Analysis	24
3.4.2	Undesired Sequences Removal	24
3.5	Data Analysis	25
3.5.1	Read Mapping	25
3.5.2	Detection of Single-Nucleotide Polymorphism	26
3.6	Application in Clinical Mycology	27
3.7	Bioinformatics Tools for Clinical Mycology	28
3.7.1	ITSx	29
3.7.2	CLoVR-ITS	29
3.7.3	FHiTINGS	29
3.7.4	PIPITS	30
3.7.5	FindFungi	30
3.7.6	Comparison of Tools	31
4	FRAMEWORK GENESIS	33
4.1	Conception	33
4.1.1	Requirements	33
4.1.2	Domain Model	35
4.1.3	Architecture	36
4.2	Development	38
4.2.1	Package Structure	38
4.2.2	Application Start	39
4.2.3	Application Run	41
4.2.4	Process Run	42
4.3	Demonstration	47
4.3.1	Console Mode	47
4.3.2	Graphical User Interface Mode	48
5	FRAMEWORK TESTING	50
5.1	Identification of Specie	50
5.2	Detection of Antifungal Resistance	51
6	CONCLUSION	53
6.1	Final Considerations	53
6.2	Future Prospects	54
A	SEQUENCING DATA FORMATS	70

LIST OF FIGURES

Figure 1	The technical procedures of Next Generation Sequencing (NGS) – library preparation and base-calling. Adapted from Harvard Chan Bioinformatics Core (2019) .	20
Figure 2	Example of the FASTQ file. Adapted from Zhang (2016) .	23
Figure 3	Example of the FASTA file format. Adapted from Zhang (2016) .	23
Figure 4	Representation of the functional requirements using use case diagram. Designed in Enterprise Architecture.	34
Figure 5	Representation of the domain model. Designed in Enterprise Architecture.	36
Figure 6	Logic view. Designed in Enterprise Architecture.	37
Figure 7	Packages view of the framework. Designed in Enterprise Architecture.	38
Figure 8	Sequence diagram of the application start. Designed in Enterprise Architecture.	40
Figure 9	Class diagram of the application start. Designed in Enterprise Architecture.	40
Figure 10	Sequence diagram of the application run. Designed in Enterprise Architecture.	41
Figure 11	Class diagram of the application run. Designed in Enterprise Architecture.	42
Figure 12	Sequence diagram of the process run. Designed in Enterprise Architecture.	44
Figure 13	Class diagram of the specie identification process. Designed in Enterprise Architecture.	45
Figure 14	Class diagram of the antifungal resistance detection process. Designed in Enterprise Architecture.	46
Figure 15	An example of interaction in console mode of the framework.	48
Figure 16	An example of interaction in Graphical User Interface (GUI) mode of the framework.	49
Figure 17	An example of the species identification output.	50
Figure 18	Snippet of outputs generated by the framework.	52
Figure 19	Example of SAM file format with the header and alignment sections. Adapted from Zhang (2016) .	70

Figure 20	Example of the GFF/GTF file format. Adapted from Zhang (2016) .	73
Figure 21	Example of the BED file format with the track line. Adapted from Zhang (2016) .	74
Figure 22	Example of the VCF file format. Adapted from Zhang (2016) .	75

LIST OF TABLES

Table 1	Main risk factors for infection caused by <i>Candida</i> species (Centonze et al., 2013; Concia et al., 2009; Fridkin, 1996; Muskett et al., 2011; Pfaller and Diekema, 2007; Costa-De-Oliveira et al., 2008).	7
Table 2	Main characteristics of the sequencing platforms (Ansorge, 2010).	19
Table 3	Analysis of the available tools for taxonomic identification.	32

LIST OF ABBREVIATIONS

- A** Adenine. 21–23
- ABC** ATP-Binding Cassette. 12
- ASCII** American Standard Code for Information Interchange. 22
- AST** Antifungal Susceptibility Testing. 1, 14
- AUC** Area Under Curve. 15
-
- BLAST** Basic Local Alignment Search Tool. 25, 29, 30
- BLAST_n** BLAST Nucleotide. 29
- bp** Base-Pairs. 18
-
- C** Cytosine. 21–23
- cDNA** Complementary DNA. 18
- CHCA** α -cyano-4-hydroxycinnamic Acid. 8
- CLoVR** Cloud Virtual Resource. 29
- CLSI** Clinical and Laboratory Standards Institute. 14
-
- DDBJ** DataBank of Japan. 16
- DHB** 2,5-dihydroxybenzoic Acid. 8
- DNA** Deoxyribonucleic Acid. 3, 9, 11, 14, 16–18, 20–22, 26, 29, 36, 43, 51
- dNTP** Deoxyribonucleotide Triphosphate. 9
-
- ENA** European Nucleotide Archive. 16
- EUCAST** European Committee on Antimicrobial Susceptibility Testing. 14
-
- FDA** Food and Drug Administration. 28
-
- G** Guanine. 21–23
- GC** Guanine-Cytosine. 21
- GUI** Graphical User Interface. viii, 41, 47–49
-
- HIV** Human Immunodeficiency Virus. 1, 5, 12, 27
- HOG** High Osmolarity Glycerol. 14
-
- IC** Invasive Candidiasis. 1, 6, 7
- ICU** Intensive Care Unit. 6, 7
- IFI** Invasive Fungal Infection. 1, 2, 4–6, 10, 53, 54
- INSDC** International Nucleotide Sequence Database Collaboration. 16

- ITS** Intergenic Transcribed Spacer. 8, 27, 29–31, 51
- LCA** Lowest Common Ancestor. 30
- MALDI-TOF MS** Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass spectrometry. 8, 15
- MARDY** Mycology Antifungal Resistance Database. 16
- MB** Megabyte. 18
- MBT ASTRA** MALDI Biotyper-Antibiotic Susceptibility Test Rapid Assay. 15
- MIC** Minimum Inhibitory Concentration. 14
- mRNA** Messenger Ribonucleic Acid. 12
- NCBI** National Center for Biotechnology Information. 16, 51
- NGS** Next Generation Sequencing. viii, 2, 3, 8, 9, 15–18, 20–29, 50, 51, 53, 70, 72
- OTU** Operational Taxonomic Unit. 29, 30
- PCR** Polymerase Chain Reaction. 15, 21, 24
- PGM** Personal Genome Machine. 18
- PKC** Protein Kinase C. 14
- RDP** Ribosomal Database Project. 30, 31
- RG** Relative Growth. 15
- RNA** Ribonucleic Acid. 9, 11, 18, 20, 21, 24, 27
- RNA-Seq** RNA Sequencing. 18
- rRNA** Ribosomal Ribonucleic Acid. 8, 24
- SBL** Sequencing By Ligation. 17, 18
- SBS** Sequencing By Synthesis. 17, 18
- SDA** Sabouraud Dextrose Agar. 5
- SNP** Single-Nucleotide Polymorphism. 25–27, 74
- SOLiD** Sequencing by Oligo Ligation Detection. 18
- SOP** Standard Operating Procedure. 29
- SRA** Sequence Read Archive. 16
- T** Thymine. 21–23
- TFA** Trifluoroacetic Acid. 8
- UML** Unified Modeling Language. 33, 35, 38
- WGS** Whole Genome Sequencing. 18, 21, 27, 28

INTRODUCTION

The incidence of **Invasive Fungal Infections (IFIs)** in critically ill patients is increasing. The use of antineoplastic and immunosuppressive drugs, empiric antifungal therapy or prophylaxis and indwelling medical devices are risk factors that contribute to this trend. The increase of fungal infections is associated with the high rates of morbidity and mortality. Patients with burns, neutropenia, **Human Immunodeficiency Virus (HIV)** infection and pancreatitis are more prone to **IFIs** (Yapar, 2014; Enoch et al., 2006).

Candida species are commonly found as human commensals on the mucosal surfaces of gastrointestinal and genitourinary tracts and skin. They are versatile organisms and their success in the infection process is related with their pathogenicity and virulence mechanisms, as well as the immunologic condition of the patient. They are opportunistic pathogens, representing a serious health care problem with mortality rates ranging 50%-75% (Enoch et al., 2006). More than 90% of **IFIs** are caused by *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis* (Sardi et al., 2013).

The taxonomic identification and **Antifungal Susceptibility Testing (AST)** of the fungi species involved on **IFIs** are crucial for the effectiveness of antifungal treatment. Thus, the current work presents a solution for identification and detection of antifungal resistance in species of the genus *Candida*, in order to help clinicians choose the most adequate treatment.

The current chapter addresses the context and motivations of the present work, the objectives defined, and the document organization.

1.1 CONTEXT AND MOTIVATION

The rapid and correct identification of *Candida* species can play an important role in infection management, decreasing mortality rates. There are a diversity of methods for identifying yeast isolates from clinical samples. Conventional methods are based on morphological and physiological attributes and are time-consuming. Thus, the development of new technologies for diagnosing **Invasive Candidiasis (IC)** is of foremost importance (Urban et al., 2014).

Antimicrobial susceptibility testing plays an important role in clinical microbiology to detect antifungal resistance and to guide clinicians in therapy approaches. With the emergence of antifungal resistance, methods to detect and quantify resistance of clinically important yeasts are imperative. The gold standard antimicrobial susceptibility testing is mostly performed using methods relying on microbial growth, which are long-lasting, hence more concise alternatives have been actively sought for (Rex et al., 2001).

With the growth of molecular technologies and the emergence of new approaches of functional genomics, such as Next Generation Sequencing (NGS), it has been possible to get the sequencing of the whole genome of numerous pathogens in one sequence run. The application of NGS has revealed several advantages compared with traditional methods, such as the improvement of quality of the sequencing, reduced costs and a decreasing of sequencing time. This technology has been proven to be useful in medical microbiology and the next step will be the introduction of these new techniques into routine microbiological diagnostics (Deurenberg et al., 2017; Zoll et al., 2016).

Due to the lack of bioinformatics solutions applied to clinical mycology, especially related to the treatment and analysis of data produced by NGS assays, in this work, a framework was developed with two main purposes: the identification and detection of antifungal resistance in *Candida* species. The solution is intended to be user-friendly, and the goal is, in the future, its use in any microbiology laboratory equipped with NGS platforms.

1.2 OBJECTIVES

The objective of this thesis is the development of an integrated solution for treatment and analysis of data obtained from NGS assays. This solution is a framework that allows, simultaneously, the identification of the most important medical *Candida* species and the detection of the antifungal resistance, for the main antifungal drugs.

Considering the development of the framework, the following scientific/technological objectives were identified:

1. Review the relevant literature about IFIs caused by *Candida* species, the application of the NGS to improve the identification and detection of antifungal resistance in clinical mycology and the tools created for analysis of NGS data;
2. Study the available data sources of NGS data for *Candida* species;
3. Design, develop and validate the framework for the identification of *Candida* species, through the conserved regions on the genome;
4. Design, develop and validate the framework for the detection of antifungal resistance in *Candida* species, through the identification of mutations in genes associated with antifungal resistance for the main antifungals.

1.3 DOCUMENT ORGANIZATION

This work is divided in the following chapters:

- The second chapter will address the *Candida* genus, presenting an overview about the biology of its species, mechanisms of infection and epidemiology. It will also address antifungal agents and the resistance mechanisms developed by *Candida* species. Finally, it will cover methods for the identification and detection of antifungal resistance, and biological databases;
- The third chapter will aim to expose [NGS](#), presenting an overview about the technologies, the steps and tools of bioinformatics for analysis of data generated by these assays. Ultimately, it will discuss the applicability of [NGS](#) in clinical mycology and the tools available for the identification of the main pathogenic *Candida* species and detection of their antifungal resistance;
- The fourth chapter will attend to the specifications of the solution, describing the requirements, the domain involved, the architecture, the details of implementation and the demonstration of the solution;
- The fifth chapter will show the testing process of the framework, through the use of example datasets and [Deoxyribonucleic Acid \(DNA\)](#) sequences;
- The sixth chapter will present the final conclusions about the solution developed and the future prospects.

THE CANDIDA GENUS

This chapter addresses the *Candida* genus by describing its role in IFIs, and presenting an overview of the biology of its species, mechanisms of infection and epidemiology. It also presents the antifungals agents and the mechanisms of resistance developed by *Candida* species. Ultimately, methods for identification and detection of antifungal resistance, and biological databases are exposed.

2.1 INVASIVE FUNGAL INFECTIONS

Fungi are the biggest lineage of eukaryotic organisms and include molds, yeasts, mushrooms, polypore's, plant-parasitic rusts and smuts. Along the recent advances in sequencing methods, mycologists have predicted the existence of 5.1 million species of fungi. These organisms can grow in almost all habitats on Earth and live in a wide range of abiotic (temperature, pH, salinity, etc.) and biotic (interaction with all major groups of organisms) factors (Blackwell, 2011). Among the estimated 5.1 million fungal species, only about 600 species can cause disease in humans and very few can affect healthy people (Ko et al., 2015). There are four types of fungal infections: superficial and cutaneous infections (e.g., athlete's foot, nail infections and ringworm), mucocutaneous infections (e.g., oral, vaginal and esophageal), allergic infections (e.g., asthma and chronic sinusitis) and invasive infections (e.g., candidosis and aspergillosis) (Garber, 2001).

IFIs are the most serious clinical events regarding fungal infections. They have significantly increased in the last decades, especially due to the enlargement in the immunocompromised population, and they are associated with high morbidity and mortality rates (Costa-De-Oliveira et al., 2008). Species of the genera *Candida* and *Aspergillus* are the main organisms isolated in immunocompromised patients. The rate of mortality ranges 50% to 75% for *Candida* species (Costa-De-Oliveira et al., 2008). Although not so prevalent, the IFI by *Aspergillus* species has a mortality rate of 60% to 90% (Alangaden, 2011; Garbee and Manning, 2017; Paramythiotou et al., 2014; Pfaller et al., 2004; Costa-De-Oliveira et al., 2008).

The immunosuppression and breakdown of anatomical barriers, especially due to gastrointestinal surgery, are the major risk factors for the development of IFIs (Costa-De-Oliveira et al., 2008). It includes patients with neutropenia, haematological malignancies, bone marrow transplantation, solid organ transplantation, prolonged treatment with corticosteroids, prolonged stays in intensive care, chemotherapy, HIV infection, invasive medical procedures, and the newer immune suppressive agents (Ko et al., 2015; Richardson, 2005; Costa-De-Oliveira et al., 2008).

2.2 OVERVIEW OF THE CANDIDA GENUS

The species of the genus *Candida* are human commensal commonly found on the mucosal surfaces and skin. However, these species become opportunistic pathogens in immunocompromised patients, causing cutaneous/mucosal infections to systemic infections (Miceli et al., 2011; Negri et al., 2012; Sardi et al., 2013).

The genus contains over 150 heterogeneous species, but only 17 different *Candida* species are known to be pathogenic. Nevertheless, more than 90% of invasive infections are caused by *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis* (Sardi et al., 2013; Pfaller, 1996; Fridkin, 1996).

2.2.1 The Biology

The *Candida* genus contains an heterogeneous group of species that can all grow as yeast morphology. Colonies of *Candida* species can grow in Sabouraud Dextrose Agar (SDA) medium, after a 48-72 hours incubation at 35°C (Dadar et al., 2018; Pesti et al., 1999; Silva et al., 2012).

Microscopically, under standard conditions with optimal nutrients, yeasts grow in log phase as budding cells (blastoconidia), which are spherical to oval and have a size ranging from 1 to 8 μm (Dadar et al., 2018; Silva et al., 2012).

Moreover, certain species, such as *C. albicans* and *C. dubliniensis*, can produce a filamentous type of growth, such as true hyphae or more frequently, pseudohyphae (Dadar et al., 2018; Silva et al., 2012).

There is a diversity of genome characteristics in *Candida* species and the main distinguishing one is the haploid or diploid genome. *C. albicans*, *C. tropicalis* and *C. parapsilosis* have a diploid genome, in contrast with *C. glabrata* that has a haploid genome (Silva et al., 2012).

2.2.2 Mechanisms of Infection

Candida species are commonly found as human commensals in healthy patients. Yet, in specific situations, such as immunosuppression and/or breakdown of anatomical barriers, they can cause systemic infections denominated by IC (Enoch et al., 2006).

IC includes candidemia, disseminated haematogeneous infections, involvement of a single deep organ site (e.g. peritonitis, other abdominal infections, meningitis and infective endocarditis) and chronic hepatosplenic candidosis (Paramythiotou et al., 2014; Tragiannidis et al., 2013).

The source of *Candida* infections can be from endogenous origin, from gastrointestinal flora or mucocutaneous colonization and exogenous, from the hands of health care workers, use of the indwelling catheter devices or contaminated intravenous solutions (Erlendsdo et al., 1999).

The success of the infection process of the *Candida* species is related with their pathogenicity. These fungi have a wide range of virulence factors and an extraordinary ability to adapt to many sites of the human body. The most important virulence factors are the morphological transitions between yeast and hyphal forms, the adherence to host tissues and medical devices, the biofilm formation, the phenotypic switching and the secretion of hydrolytic enzymes (e.g. proteases, phospholipases and haemolysins) (Sardi et al., 2013; Mayer et al., 2013).

2.2.3 Epidemiology

C. albicans is the most frequently strain isolated in patients with IC, followed by *C. parapsilosis*, *C. glabrata*, and *C. tropicalis* (Paramythiotou et al., 2014; Sardi et al., 2013; Costa-De-Oliveira et al., 2008). However, in the last decade, the isolation of the non-*albicans* *Candida* species are increasing. This trend has been observed in the USA, Europe, Latin America and Africa, and could be associated with severe immunosuppression, prematurity, exposure to broad-spectrum antibiotics and older patients (Paramythiotou et al., 2014; Sardi et al., 2013; Concia et al., 2009; Lamoth et al., 2018; Costa-De-Oliveira et al., 2008).

Candida species are the most frequent cause of infection in Intensive Care Units (ICUs) worldwide. The main risk factors associated with the development of IC in patients in ICUs are the presence of central venous catheters, the treatment with broad-spectrum antibiotics, multifocal *Candida* colonization, gastrointestinal surgery, pancreatitis, parenteral nutrition, hemodialysis, mechanical ventilation and the prolonged ICU stay (Lamoth et al., 2018; Paramythiotou et al., 2014; Blumberg et al., 2001; Vincent et al., 2009).

IC is the second cause of IFIs in allogeneic haematopoietic stem cell transplant recipients and patients with hematological malignancies. The most common species isolated are *C.*

krusei and *C. glabrata*. The main risk factors for IC in patients with hematological malignancies are the neutropenia, the corticosteroid therapy and the presence of central venous catheters (Lamoth et al., 2018; Kontoyiannis et al., 2010; Pagano et al., 2006).

Table 1 summarizes the main risk factors for infection caused by *Candida* species.

<i>C. albicans</i>	Prolonged ICU stay
	Treatment with corticosteroids
	Diabetes mellitus
	Advanced age
	Central venous catheter
	Gastrointestinal surgery
	Total parental nutrition
	Prolonged antimicrobial use
	Pancreatitis
	Immunosuppressive agents
	Chemotherapy
	Neutropenia
	Renal replacement therapy
	Malnutrition
	Multiple site colonization
Burns over 50% of body sites	
Major trauma	
<i>C. glabrata</i>	Elderly patients
	Patients with malignancies
	Use of the specific antibiotics (piperacillin/tazobactam, vancomycin)
	Patients under total parenteral nutrition and with central venous catheters
	Solid organ transplantation
<i>C. parapsilosis</i>	Fluconazole prophylaxis
	Nosocomial outbreaks
	Formation of biofilms in central venous catheters
<i>C. tropicalis</i>	Implanted devices
	Total parenteral nutrition
	Less susceptible to echinocandins
<i>C. tropicalis</i>	Hematological malignancies
	Neutropenia

Table 1.: Main risk factors for infection caused by *Candida* species (Centonze et al., 2013; Concia et al., 2009; Fridkin, 1996; Muskett et al., 2011; Pfaller and Diekema, 2007; Costa-De-Oliveira et al., 2008).

2.3 METHODS FOR IDENTIFICATION

The identification of *Candida* species can play an important role in infection management, decreasing mortality rates. There is a diversity of methods for identifying yeasts from clinical samples (Urban et al., 2014).

The conventional methods are based on morphological and physiological attributes, for example, the germ-tube test, morphology studies, carbohydrate utilization, enzymatic and fluorogenic tests, and molecular typing techniques. These methods are time-consuming, contributing to a delay in microbiology diagnostics (Urban et al., 2014).

Nowadays, there are advanced technologies that allow the correct and rapid identification of fungal species, such as Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass spectrometry (MALDI-TOF MS), immunological identification, microarrays and molecular fingerprinting by NGS (Lion, 2017).

For the last years, MALDI-TOF MS has become a widely used method for microbial identification in clinical laboratories. This technology starts with the transfer of microbial colonies from agar plate to a slide that will be soaked in a matrix solution. The matrix solution is composed by α -cyano-4-hydroxycinnamic Acid (CHCA), the solvents acetonitrile, ethanol and methanol, an organic acid such as Trifluoroacetic Acid (TFA), and 2,5-dihydroxybenzoic Acid (DHB). These compounds allow the extraction of proteins from cells and the identification of differences in mass spectral patterns, through the variation of relative intensities of individual peaks. Before matrix solution preparation, the laser focus scans the sample in a predefined pattern and accumulates a mass spectrum from a defined number of laser pulse cycles. The raw spectrum is processed to yield a mass fingerprinting that contains information about peak apex m/z values. The last step of this procedure is the comparison of the mass fingerprinting of the sample with a database that contains reference mass fingerprints. The data acquisition from MALDI-TOF MS is performed in an automated manner. The great success of this method is due to the simplicity of the sample preparation, the accuracy and the speed in obtaining results (Lion, 2017; Welker, 2011).

Immunodetection is a technique that allows the detection of antigens using specific antibodies. This technique uses the monoclonal antibodies that have an affinity with the specific markers of fungal species or genera (Lion, 2017).

Microarrays are a set of probes that enable the parallel analysis of several compounds such as nucleic acids, proteins, and carbohydrates. For identification of microbial species, specific molecules are used to detect pathogens. The probes can be immobilized on a solid support (glass, silicon or nylon), and the hybridization of the microarrays leads to a specific molecular interaction, that can be detected through fluorescence. For fungal diagnostics arrays, the 28S fungal Ribosomal Ribonucleic Acid (rRNA) gene or the Intergenic Transcribed Spacer (ITS) are used (Lion, 2017).

The advent of high-throughput sequencing technologies, such as NGS, allowed the increase of the number of sequenced genomes. The general principle of the NGS technology is the incorporation of fluorescently marked Deoxyribonucleotide Triphosphates (dNTPs) through DNA polymerase, for each specific base in consecutive cycles. When incorporated during each cycle, the nucleotides are identified by their specific fluorophore excitation. The data resulting is submitted to adequate bioinformatics tools to obtain the genome assembled. Using these workflows, the genomic content can be characterized and annotated, and used for microbial identification through specific bioinformatics tools. The usage of NGS for microbial identification has several advantages such as the reduced costs, the high accuracy and quality of reads, and it is faster than other sequencing technologies (Lion, 2017).

2.4 ANTIFUNGAL AGENTS

Over the last decades, the incidence of fungal infections in critically ill patients increased. The number of antifungal drugs available is limited, especially when compared with antibacterial drugs. The antifungal agents can be divided in four groups (Kathiravan et al., 2012; Perea and Patterson, 2002), based on their mode of action:

- group I: inhibition of Ribonucleic Acid (RNA) and/or DNA synthesis (pyrimidines);
- group II: alteration of the membrane permeability (polyenes);
- group III: alteration of cell wall biosynthesis by inhibition of β -(1,3)-glucan synthase (echinocandins);
- group IV: inhibition of lanosterol demethylase in ergosterol biosynthesis (azoles).

2.4.1 *Pyrimidines*

The flucytosine (5-fluorocytosine) is a fluorinated derivative of the pyrimidine cytosine and it is the only drug of the pyrimidine class used in treatment of candidosis. Flucytosine is taken up by cytosine permease and transported into fungal cells. This compound is enzymatically transferred into 5-fluorouracil by cytosine deaminase. Additionally, the uracil phosphoribosyltransferase transforms 5-fluorouracil into 5-fluorouridine monophosphate. This compound changes protein synthesis by incorporating into RNA, which results in the cell death (Bondaryk et al., 2013; Kathiravan et al., 2012).

2.4.2 Polyenes

The polyenes are composed by a heterocyclic amphipathic molecule. Nystatin, natamycin, and amphotericin B are the three polyene drugs used in candidosis treatment. The ring of polyene molecules, resulting from the binding polyene-ergosterol, is interleaved with the fungal membrane. The result of this is the formation of aqueous pores, through which the cellular contents leak, that leads to the increase of permeability of the membrane and oxidative damage (Bondaryk et al., 2013; Kathiravan et al., 2012; Pemán et al., 2009).

2.4.3 Azoles

The azoles are a class of heterocyclic synthetic compounds and they include fluconazole, itraconazole, voriconazole, and posaconazole. Azoles compounds act by blocking the ergosterol biosynthetic pathway, binding to 14- α -lanosterol demethylase, which leads to the inhibition of the enzyme, therefore disturbing ergosterol biosynthesis. Due to accumulated intermediates of ergosterol biosynthesis, a subsequent mechanism of sterol metabolism mediated by C5,6 desaturase enzyme is activated, which results in toxic methylated sterols leading to growth inhibition (Bondaryk et al., 2013; Kathiravan et al., 2012; Pemán et al., 2009).

2.4.4 Echinocandins

The echinocandins are a class of semisynthetic lipopeptide antifungal compounds. They are the first-line treatment of invasive infections caused by *Candida* species. This class of antifungals is composed by caspofungin, micafungin and anidulafungin. The target of echinocandins is the β -(1,3)-glucan synthase, that is responsible for synthesis of β -(1,3)-glucan, which is a main component of the cell wall. So, through inhibition of synthesis of glucan, the cell can not control the osmotic pressure, leading to cell lysis (Bondaryk et al., 2013; Kathiravan et al., 2012; Pemán et al., 2009).

2.5 ANTIFUNGAL RESISTANCE AND MECHANISMS

Antifungal resistance is becoming a serious problem in the management of IFIs. The number of antifungals available is limited and patients with high risk often have co-occurrence of two or more medical conditions, that affect the effectiveness of the therapy (Wiederhold, 2017). The main factors that influence antifungal resistance and consequently clinical resistance are:

- Antifungal susceptibility of the fungal isolate, that depends of the expression of their virulence factors and its interaction with the host and the antifungal agents;
- Characteristics of antifungal agents like pharmacokinetic variability, dose regimen and drug penetration, stability and interactions with other drug;
- Factors of the host such as immune response, severity and site of infection and underlying disease.

These factors taken together contribute to a poor outcome (Pemán et al., 2009). Antifungal resistance can be acquired, after exposure to antifungal drugs or intrinsic (innate) (Perea and Patterson, 2002).

2.5.1 Against Pyrimidines

The prevalence of flucytosine resistance in *Candida* species is low, and the percentage of resistant-species is less than 2% (Sanglard and Odds, 2002).

The flucytosine is a base pyrimidine analog that is incorporated during DNA and RNA synthesis and results in cell death. The mechanisms of resistance are associated with changes in the enzyme purine-cytosine permease, which is responsible for the uptake of the drug into the cell; in the enzyme cytosine deaminase, that is responsible for the conversion to 5-fluorouracil; or in the enzyme uracil phosphoribosyltransferase, which is responsible for the conversion to 5-fluorouracil (Pemán et al., 2009; Espinel-Ingroff, 2008).

2.5.2 Against Polyenes

The occurrence of polyene resistance in *Candida* species is limited. *C. lusitaniae* shows intrinsic resistance to amphotericin B (Pfaller et al., 1994). *C. glabrata* and *C. krusei* are more susceptible to amphotericin B than *C. albicans* (Pfaller et al., 2004; Canuto and Rodero, 2002).

Resistance to amphotericin B is quite rare. However, the resistance is related to mutations in the ERG3 gene (which encodes a C-5 sterol desaturase, an enzyme involved in ergosterol biosynthesis), which leads to low concentration of ergosterol in the fungal membrane. Resistance to amphotericin B may also be mediated by increased catalase activity, with decreasing susceptibility to oxidative damage. In polyene-resistant *Candida* species, the cell membrane has low ergosterol content (Kanafani and Perfect, 2010; Pemán et al., 2009).

2.5.3 Against Azoles

In HIV infected patients with oropharyngeal and esophageal candidosis, the number of fluconazole-resistant *C. albicans* is high. Nonetheless, the percentage of azole-resistant isolates in the most invasive *Candida* infections is low, and represents 1.0%-2.1% in *C. albicans*, 0.4-4.2% in *C. parapsilosis*, 1.4-6.6% in *C. tropicalis* and 7-12% in *C. glabrata* (Pfaller et al., 2006). *C. krusei* is intrinsically resistant to fluconazole, and several studies have associated the innate azole resistance to a reduced drug accumulation (Katiyar and Edlind, 2001; Marichal et al., 1995).

Candida species developed multiple mechanisms of antifungal tolerance to azoles, mainly fluconazole. These mechanisms include alterations at molecular level such as changes in ergosterol biosynthetic enzymes, drug uptake and efflux systems, chromosomal abnormalities, biofilms, and Messenger Ribonucleic Acid (mRNA) stability (Pemán et al., 2009; Sanglard et al., 2014).

In *Candida* species resistant to azoles, point mutations in the gene ERG11 (encodes the 14- α -lanosterol demethylase) have been identified, that prevent binding of azoles to the target site. These amino acid substitutions result in a decrease of susceptibility to azoles and have been observed in three critical *hot spot* regions of ERG11 alleles (Marichal et al., 1999). Another mechanism of azole resistance, also associated with the gene ERG11, is the increase of its expression due to mutations in the gene encoding the zinc-cluster transcriptional regulator Upc2p. The expression of UPC2 alleles was found to be associated with increased of ERG11 expression, increased ergosterol production, and decreased fluconazole susceptibility (Pemán et al., 2009; Whaley et al., 2017; MacPherson et al., 2005).

The activation of efflux pumps is another mechanism of azole-resistance and allows the decrease of intracellular drug concentration. This mechanism is regulated by two specific pumps: the ATP-Binding Cassette (ABC) transporters encoded by CDR genes and the major facilitators encoded by MDR genes. CDR gene upregulation confers resistance to all azoles and MDR-encoded efflux pumps are specific for fluconazole (Pemán et al., 2009). An additional mechanism of resistance is the CDR1 mRNA stability, where the polyA tail is hyperadenylated in resistance isolates, and this is associated with the lifetime of mRNA, where is threefold higher (Sanglard et al., 2014; Manoharlal et al., 2008).

Chromosomal abnormalities, including loss of heterozygosity of specific genomic regions, an increase of chromosome copy number and segmental or chromosomal aneuploidies, are associated with azole resistance. ERG11, CDR and MDR are related to chromosomes where the alteration in copy number confers resistance (Sanglard et al., 2014; Coste et al., 2007).

Biofilms are an organized three-dimensional structure comprised of a dense network of yeast and filamentous cells embedded in an exopolymeric matrix consisting of carbohydrates, proteins, and nucleic acids. The matrix production is highly regulated, and the

key constituent is β -(1,3)-glucan, which is produced by glucan synthase. *Candida* biofilms are intrinsically resistant to azoles, and the mechanisms of resistance are multifactorial, involving induction of drug efflux transporters and drug sequestration within the extensive matrix structure. The drug sequestration within the extracellular matrix is the largest determinant of the multidrug-resistance phenotype (Fanning and Mitchell, 2012; Ramage et al., 2005).

2.5.4 Against Echinocandins

Echinocandin-resistant *Candida* species are rare, and the percentage of resistant-species is low. Considering *C. albicans*, the rates were found to be 0.0-0.4% to anidulafungin, 0.2-0.4% to caspofungin and 0.1-0.3% to micafungin, while for *C. glabrata* these values are 1.8-4.6% to anidulafungin, 1.6-2.5% to caspofungin and 0.9-1.2% to micafungin (Grossman et al., 2014; Perlin, 2015). *C. parapsilosis* and *C. guilliermondii* are intrinsically resistant to echinocandins and this resistance was associated with the natural occurrence of proline-to-alanine substitution at amino acid position 660 (Cantón et al., 2006; Garcia-Effron et al., 2008).

The mechanisms of echinocandin resistance include mutations in genes that encode FKS subunits, biofilms, adaptive cellular factors and chitin synthesis (Pemán et al., 2009; Sanglard et al., 2014; Costa-De-Oliveira et al., 2011).

The target of echinocandin drugs is the β -(1,3)-glucan synthase, that is encoded by different FKS genes (FKS1 and FKS2). The echinocandin-resistant *Candida* species exhibit point mutations in highly conserved regions of FKS genes. These amino acid changes lead to the decrease of echinocandin susceptibility (Pemán et al., 2009; Sanglard et al., 2014; Wiederhold, 2016; Costa-De-Oliveira et al., 2011). For *C. albicans*, amino acid change at FKS1 Ser645 is the most common. In *C. glabrata*, mutations occur in both FKS1 and FKS2, and the amino acid change at FKS2 Ser663 (equivalent to *C. albicans* Ser645) are the most frequent (Costa-De-Oliveira et al., 2011). Moreover, amino acid substitutions at FKS1 Ser629 and at FKS2 Phe659 may also occur. However, the *hot spot* mutations that occur in FKS1 or FKS2 depend on the relative expression of these genes (Garcia-Effron et al., 2009; Katiyar et al., 2012; Perlin, 2015).

As it happens in azole resistance, the biofilms have an important role in echinocandin resistance. These can sequester echinocandin agents, preventing them from reaching to the cell membrane. Even so, the echinocandins change the synthesis of β -(1,3)-glucan, the key constituent of biofilms, decreasing drug sequestration, which make biofilms susceptible to antifungal agents. The protein Smi1 has been associated with the production of the glucan in biofilms. The interaction between Smi1, Rlm1 (transcription factor) and glucan synthase FKS1 allow the production of drug-sequestering biofilm β -glucan (Desai et al., 2013).

Candida species have adaptive cellular factors that confer protection against cellular stress, like those caused by echinocandins. These mechanisms are associated with the activation of pathways such as cell wall integrity, Protein Kinase C (PKC), Ca^{2+} /Calcineurin/Crz1, and High Osmolarity Glycerol (HOG) cascades. The activation of these pathways induces over-expression of chitin synthase gene (Chs2 and Chs8), and leads to the increase of chitin content, that is directly related with echinocandin resistance (Munro et al., 2007; Walker et al., 2013, 2015).

2.6 METHODS FOR DETECTION OF ANTIFUNGAL RESISTANCE

Antimicrobial susceptibility testing plays an important role in clinical microbiology allowing the detection of antifungal resistance and guide clinicians in therapy approach. With the emergence of antifungal resistance, methods to detect and quantify resistance of clinically important yeasts are imperative (Rex et al., 2001).

AST comprehends a set of methods that can be used for detection and evaluation of antifungal resistance. They are based on the measurement of microbial growth resulting from the exposure to an inhibitor. Various testing procedures include broth microdilution (gold standard), agar and disk diffusion and Etest® (Rex et al., 2001; Sanguinetti and Posteraro, 2018).

The methods based on broth microdilution have two standard methods, the Clinical and Laboratory Standards Institute (CLSI) (CLSI, 2012) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (Arendrup et al., 2015). During 24 hours, the yeasts are incubated in growth medium supplemented with different concentrations of antifungal agents. At the end of the incubation time, the antifungal activity is measured through the determination of the Minimum Inhibitory Concentration (MIC) of an antifungal drug, that is the minimal drug concentration that inhibits fungal growth. Both CLSI and EUCAST have different breakpoint values, that allow the classification of the isolate as susceptible, susceptible dose-dependent or resistant (Sanguinetti and Posteraro, 2018).

In the Etest®, the MIC values are determined from the point of intersection of a growth inhibition zone with a calibrated strip impregnated with a gradient of antifungal concentration, and placed on an agar plate lawned with the fungal isolate under test. This methodology has the standard method and reference MIC values for classification of isolates adapted to each antifungal agents (Rex et al., 2001; Song et al., 2015).

Flow cytometry has been pointed out as a promising technology to measure the effect of the antifungal agents in the membrane, the alteration in metabolic activity resulting in membrane damage and the uptake of DNA binding dye in the yeast cells. This method uses sodium deoxycholate for permeability and propidium iodide to detect the increased permeability of the cell membrane after antifungal treatment. The yeasts are incubated

with growth medium and antifungal agent. After incubation, the dyes are applied in the preparation and placed on the equipment of the flow cytometry. Flow cytometry antifungal susceptibility testing allows to obtain results in a short time (2 to 6h) and reproducible results (Rex et al., 2001; Ramani and Chaturvedi, 2000; Pina-Vaz et al., 2001, 2005).

The MALDI-TOF MS technology, as described above, can be used for identification of microbial species, but over the last years, its potential in testing resistance and susceptibility of microorganisms to antimicrobial agents has also been unveiled. The MALDI Biotyper-Antibiotic Susceptibility Test Rapid Assay (MBT ASTRA) is a MALDI-TOF MS-based semi-quantitative technique designed for rapid antibiotic susceptibility testing in bacteria, which has also been tested in *C. albicans* and *C. glabrata*. This method compares the growth of microorganisms after incubation in the absence or in the presence of different concentrations of antimicrobial drugs. The cell growth is inferred from the comparison of the Area Under Curve (AUC) of the MALDI-TOF MS spectra for the different incubation setup (with or without antimicrobial drug). The Relative Growth (RG) is calculated for each concentration of antimicrobial agent as the ratio of the AUC observed with or without drug exposure, and a RG cutoff discriminating resistance from susceptibility is determined experimentally. In case strains show RG above the cutoff, they are considered resistant, otherwise are considered susceptible. However, it is necessary to optimize MALDI-TOF MS-based assays to obtain timely, accurate and reliable results (Florio et al., 2018).

Recent years have witnessed the growth of molecular biology technologies, ideally suited, not only for fungal identification, but also for the assessment of drug resistance mechanisms. Real-time Polymerase Chain Reaction (PCR) and sequencing techniques have been widely used for the quantification of gene expression and search for transcriptional regulator mutations involved in the evolution of antifungal drug resistance (Costa-De-Oliveira et al., 2011; Castanheira et al., 2010; Walker et al., 2009).

New technologies such as NGS will integrate sequencing and data analysis in one efficient workflow that will be crucial for routine microbiology laboratories. This technology is extensively discussed in the next chapter.

2.7 BIOLOGICAL DATABASES

A database is a computational library for storage, searching, and representation of several types of data. A database has the advantage to store a large amount of data with easy access and handling. Biological databases store information about life sciences as a whole, mainly from scientific research. They contain information from many research areas, such as genomics, microarray gene expression, proteomics, phylogenetics, metabolomics, gene function, structure, localization and similarities of biological sequences (Bhatt et al., 2018).

In the context of this work is used three different biological databases, which are described in the following sections.

2.7.1 *GenBank*

GenBank¹ is a public database that contains information about nucleotide sequences, supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI). GenBank is part of the International Nucleotide Sequence Database Collaboration (INSDC), which comprises the DNA DataBank of Japan (DDBJ), European Nucleotide Archive (ENA), and GenBank at NCBI. The data are exchanged daily between these three organizations, which allows the centralization of the most recently available sequence data from all sources (Clark et al., 2016).

2.7.2 *Sequence Read Archive*

The Sequence Read Archive (SRA)² is a public database that stores NGS data established under the guidance of the INSDC. The mission of this database is to preserve sequencing data and to provide free, unrestricted and permanent access to the data. Currently, it supports a high range of sequencing platforms, such as Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLID System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT® (Leinonen et al., 2011).

2.7.3 *Mycology Antifungal Resistance Database*

The Mycology Antifungal Resistance Database (MARDY)³ is a MySQL relational database, that includes information about antifungal resistance mechanisms of human, animal and plants, and associated drugs. The mechanisms of antifungal resistance that are stored include amino acid substitution, tandem repeat sequences and genome ploidy with a record to the corresponding literature. The classes of drugs comprise polyenes, azoles, and echinocandins (Nash et al., 2018).

MARDY is the first database for the fungi kingdom with information about resistance mechanisms and their association with species, antifungal drugs, and genes.

¹ www.ncbi.nlm.nih.gov/genbank

² www.ncbi.nlm.nih.gov/sra

³ www.mardy.net/

NEXT-GENERATION SEQUENCING

This chapter presents an overview of **NGS**, its technologies, technical procedures and steps, and bioinformatics tools for data analysis. Lastly, it makes reference to the usage of **NGS** in clinical mycology and the available tools for the identification and detection of antifungal resistance in *Candida* species.

3.1 OVERVIEW OF NEXT-GENERATION SEQUENCING

In the 1970s, **Sanger et al. (1977)** and **Maxam and Gilbert (1977)** developed methods of sequencing **DNA** by chain termination (Sanger sequencing) and fragmentation techniques, respectively. The advent of these sequencing techniques represented the creation of the First-Generation Sequencing. Over 30 years later, the Sanger sequencing method has remained the most commonly used **DNA** sequencing due to the absence of toxic and radioisotopes chemicals. The creation of sequencing techniques allowed deciphering complete genes and genomes, which revolutionized biology (**Morozova and Marra, 2008; Thermes, 2014**).

The emergence of the Human Genome Project stimulated the development of new sequencing methods with reduced costs, which took the name of **NGS** technologies (**Collins et al., 2004**). These technologies have three improvements over Sanger sequencing. First, the preparation of **NGS** libraries in a cell-free system, in contrast with Sanger sequencing that required bacterial cloning **DNA** fragments. Second, thousands-to-many-millions of sequencing reactions are produced at the same time. Finally, the sequencing output is directly detected without the use of electrophoresis (**Thermes, 2014**).

NGS is performed by repeated cycles of polymerase-mediated nucleotide extensions or by iterative cycles of oligonucleotide ligation, i.e., **Sequencing By Synthesis (SBS)** or **Sequencing By Ligation (SBL)**. In **SBS** approaches, a polymerase is used with a fluorophore or a change in ionic concentration and the incorporation of a nucleotide into an elongation strand is identified. In **SBL** approaches, a probe sequence is bound to a fluorophore that hybridizes to a **DNA** fragment and to an adjacent oligonucleotide. The emission spectrum of the fluorophore indicates the identity of the base or bases complementary in a specific

position. In both approaches, DNA is clonally amplified on a solid surface (Goodwin et al., 2016).

In 2005, the first NGS technology was created by 454 Life Sciences (now Roche), and represented the emerging of the Second-Generation Sequencing. This technology was based on pyrosequencing method, that uses the SBS approach, and can generate about 200000 reads (approximately 20 Megabyte (MB)) of 110 Base-Pairs (bp) (Margulies et al., 2005). In 2006, the Solexa/Illumina sequencing platform was commercialized. In 2007, Applied Biosystems (now Life Technologies) was created, with the Sequencing by Oligo Ligation Detection (SOLiD), that also uses the SBL approach. Both Illumina and SOLiD sequencers generated much larger numbers of reads than the first technology, but they were only 35 bp long (Valouev et al., 2008). In 2010, Ion Torrent (now Life Technologies) released the Personal Genome Machine (PGM), developed by Jonathan Rothberg, the founder of 454. The difference between technologies is that the PGM uses semiconductor technology and does not rely on the optical detection of incorporated nucleotides using fluorescence and camera scanning. This technology uses the SBS approach for sequencing. This resulted in higher speed, lower cost, and smaller instrument size (Liu et al., 2014).

In 2010, the Third-Generation of Sequencing emerges through the creation of the PacBio RD instrument by Pacific Biosciences. This technology generated several thousands of up-to-several-kilobase-long reads. This method is based on the detection of natural DNA synthesis by a single DNA polymerase. Incorporation of phosphate-labeled nucleotides leads to base-specific fluorescence, which is detected in real time (Rank et al., 2009).

Nowadays, there are six next-generation technologies commercially available: Illumina, PacBio SMRT systems, Heliscope, Solid, PGM and 454 systems. Each platform has different protocols for DNA or RNA library preparation and different techniques to detect the signal and read of the DNA or RNA sequences (Ansoerge, 2010). Table 2 summarizes the main characteristics of each sequencing platform.

NGS can be used for many applications, such as Whole Genome Sequencing (WGS), whole-exome and targeted sequencing and RNA Sequencing (RNA-Seq). The WGS is the sequencing of the entire genome without using methods for sequence selection. The whole-exome and targeted sequencing is the sequencing of only exons or other selected regions. A system of capture or amplification is used to isolate or enrich only exons or target regions. This is done by designing probes or primers for the regions of interest. The RNA-Seq is a method of sequencing Complementary DNA (cDNA) derived from RNA. This approach can be used to sequence both coding and non-coding RNA (Goodwin et al., 2016).

Over the last years, the evolution of high-throughput sequencing platforms allowed the increase of quality of sequencing, reduced costs and the decrease of sequencing time. Therefore, NGS has become an important technology used in clinical laboratories (Thermes, 2014).

Company	Platform	Sequence by	Reads per run	Run time	Read length (bp)	Output per run
Roche	GS FIX Titanium XL+	Synthesis	1 million	23h	700	700 Mb
	GS Junior System	Synthesis	0.1 million	10h	400	400 Mb
	Ion Torrent	Synthesis	4 millions	4h	200-400	1.5-2 Gb
Life Technologies	Proton	Synthesis	60-80 millions	4h	125	8-10 Gb
	Abi/Solid	Ligation	2.7 billions	10 days	75+35	300 Gb
Illumina/Solexa	HiSeq2000/2500	Synthesis	3 billions	12 days	2 x 100	600 Gb
	MiSeq	Synthesis	25 millions	65h	2 x 300	15 Gb
	RSII	Synthesis	0.8 million	2 days	>10kb (50% of reads)	5 Gb
Helicos Biosciences	Heliscope	Synthesis	500 millions	10 days	~30	15 Gb

Table 2.: Main characteristics of the sequencing platforms (Ansorge, 2010).

3.2 TECHNICAL PROCEDURES

NGS allows sequencing of the whole or a specific part of the genome of several organisms in one sequence run. The principle of NGS sequencing is the massively parallel sequencing of clonally amplified or single DNA molecules, which are spatially separated in a flow cell (Head et al., 2014; Voelkerding et al., 2009).

Before sequencing, it is necessary to prepare the DNA or RNA library. The library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Then, it follows the next step – base calling. In the medium is added sequencing reagents, that include fluorescently labeled nucleotides. The flow cell is imaged and the emission wavelength and intensity are used to identify the base. Each base is recorded and stored in sequencing files which are used for data analysis (Hui, 2014). This entire process is exposed in Figure 1. The following sections detail the most relevant steps: library preparation and base calling.

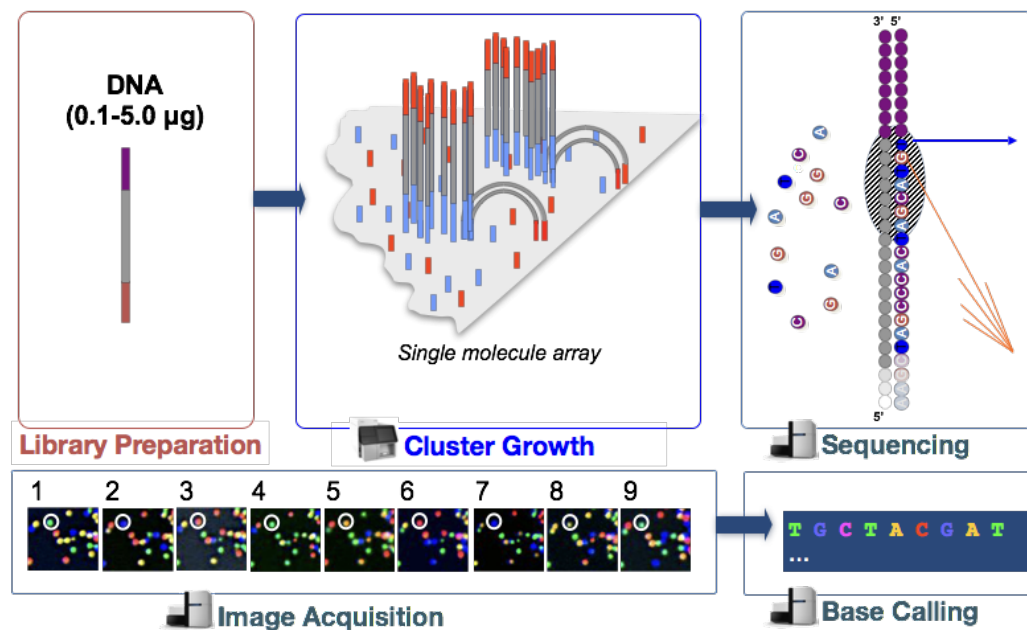


Figure 1.: The technical procedures of NGS – library preparation and base-calling. Adapted from Harvard Chan Bioinformatics Core (2019).

3.2.1 Library Preparation

A library is a collection of randomly sized DNA or RNA fragments representing the sample input. However, depending on the type of NGS applications, different steps of preparation are taken (Head et al., 2014).

This work addresses the library construction using DNA applied in WGS. The NGS library construction using DNA starts with the fragmentation of the target sequences in the desired length, through physical, enzymatic or chemical techniques. The 3' ends of the fragments are additionally adenylated with a single base, allowing hybridization to the 3' thymine overhang of sequencing adapters. The resulting DNA fragments are ligated to the adapters and amplified through PCR reactions. The adapters will enable the DNA to hybridize with the surface of the sequencing reaction chip. The collection of adapter-ligated fragments forms a library which must be validated before sequencing. The need for validation is due to the fact that NGS libraries may contain errors, which decrease the data quality, and this can change the data interpretation (Head et al., 2014).

To check for errors, it is necessary to evaluate the existence of bias, i.e., the systematic distortion of data due to the experimental design, and the library complexity that can be measured by the percentage of duplicate reads, which are present in the sequencing data. The main sources of bias are the PCR amplification reactions and the steps that involve enzymatic reactions. In PCR amplification, the Guanine-Cytosine (GC) content has considerable impact on the efficiency of PCR reactions. In enzymatic steps, bias can be reduced in purification steps by pooling barcoded samples before gel or bead purification (Head et al., 2014).

3.2.2 Base Calling

Base calling is a process in which the signal intensities of nucleotides incorporated are detected and converted into individual bases. Each NGS technology has its own technique for base calling, and in this work, we will address the techniques used by Illumina and Ion Torrent (Cliften, 2014; Ledergerber and Dessimoz, 2011).

On an Illumina instrument, the flow cell, that contains the incorporated nucleotides into the elongating DNA chain, is scanned for each of the four bases, i.e., the Adenine (A), Guanine (G), Cytosine (C) and Thymine (T), and the base with the highest intensity is determined to be the incorporated nucleotide (Cliften, 2014).

A single flow cell contains over one billion DNA clusters that are firmly packed into a small area. The first step is the template generation, i.e., the process of identifying the location of each cluster on a flow cell. Illumina uses image data from the first four cycles to identify the clusters. After nucleotide incorporation, two lasers are used to excite the

fluorophores, one that excites the **A** and **C** fluorophores and one that excites the **G** and **T** ones. The images from each of the four fluorophores are recorded (Cliften, 2014).

The second step is the alignment of each image of the flow cell to the previous template obtained, and then to the image extraction, i.e., the process of assigning an intensity value for each **DNA** clusters from an image. Finally, the intensity correction follows. Illumina applies a color filter matrix (adaptive color matrix), which corrects the relative intensity shifts of the four color channels over the course of the run or between different portions of the flow cell (Cliften, 2014).

On an Ion Torrent instrument, the signal of nucleotide incorporation is detected by the sensor at the bottom of the well, converted to voltage, and then digitized computationally off the chip. The signal of nucleotides is subjected to noise or variation during generation and detection. Thus, it is necessary to identify the background noise and subtract it from each incorporation event (Cliften, 2014).

The Ion Torrent instrument has a program denominated by Solver, that uses a generative model to create an artificial model of expected incorporation signals for each base. Each signal is compared with the models, and the most likely sequence is determined. Solver uses an iterative process to construct predicted signals for partial base sequences, and then measures their fit to the observed signals generated by the run (Cliften, 2014).

3.3 SEQUENCE DATA FORMATS

The output of **NGS** experiments consists of billions of short reads, which are stored in files with sizes that range from a few gigabytes to hundreds of gigabytes. To efficiently store, view, and manipulate the information of the **NGS** files, several file formats have been developed, including FASTQ, FASTA, SAM/BAM, GFF/GTF, BED, and VCF, typically used in the analysis of **NGS** data (Zhang, 2016). Only FASTQ and FASTA format are addressed in this section. Other file formats are described in Appendix A.

3.3.1 FASTQ

The FASTQ format stores simultaneously the nucleotide sequence and its corresponding Phred quality scores. Phred scores assign, using **American Standard Code for Information Interchange (ASCII)** characters, a score to each base call that indicates the quality or confidence of the attribution. The scores are in the 1-40 range and are logarithmically related to base calling error probabilities. The file extensions of the FASTQ are “.fq” or “.fastq” (Zhang, 2016).

In a FASTQ file, each sequence (short read of **NGS**) is defined by four lines of text. The first line starts with a “@” character and is followed by a sequence identifier and an optional

description. The second line is the raw sequence letters: *A*, *T*, *G*, *C*, and *N* (unknown). The third line begins with a “+” symbol and is optionally followed by the same sequence identifier and a description. The “+” serves as a marker indicating the end of the sequence. The fourth line keeps the quality values for the sequence and must contain the same number of symbols as letters in the sequence (Zhang, 2016). Figure 2 shows an example of a single sequence from the Illumina sequence system.

Due to the high number of short reads resulting from the NGS experiment, frequently the FASTQ files are compressed to GNU zip format (.gz file extension) to reduce file size (Zhang, 2016).

```
@HWI-ST193:542:C2H0GACXX:8:1101:4404:2179 1:Y:0:ACACGA
ATGCNTTTTATAATCAAAAGCGAAGACCTAGCAGGAGGTTAAAAACCTTT
+
<<<<#2<@5:9@44:@@?4(-8@(<9@<<658.==5=0<>???????)??
```

Figure 2.: Example of the FASTQ file. Adapted from Zhang (2016).

3.3.2 FASTA

The FASTA format has been the standard format for nucleotide sequence. In NGS, it is the standard format for reference genome sequences used by mapping/alignment. The FASTA file extensions are “.fa” or “.fasta” (Zhang, 2016).

A FASTA file starts with a header line followed by lines of sequence. The header line begins with a > sign for sequence identifier and may contain optional description information. Sequence lines consist of characters representing nucleotide bases in the sequence. Each nucleotide base is encoded as a single character (*A*, *T*, *G*, *C* or *N*, if undetermined) (Zhang, 2016). Figure 3 exhibits an example of the FASTA file.

```
>gi|568815581:c7687550-7668402 Homo sapiens chromosome
17, GRCh38 Primary Assembly
GATGGGATGGGGTPTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGT TTT
GAGCTTCTCAAAG TC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTG CGT
TCGGGCTGGGAGCGTG
```

Figure 3.: Example of the FASTA file format. Adapted from Zhang (2016).

3.4 DATA PREPROCESSING

The data produced by NGS assays are submitted to several steps before the mapping and variant analysis of the reads, to ensure the veracity and quality of results. The main steps of preprocessing include the quality check of FASTQ reads and the removal of the undesired sequences such as reads of low quality, rRNA and host sequences. There are several bioinformatics tools for the different steps of preprocessing, integrated in pipelines that allow easy workflows for data analysis. A brief overview of the main steps and most common tools used in each step of NGS data preprocessing is given in the next sections.

3.4.1 *Quality Analysis*

NGS experiments can be influenced by several factors that happen during the library preparation and sequencing process, which can affect the data veracity. These factors include platform specific error profiles, systematic variation in quality scores across the sequence read, biases in sequence generation driven by base composition, departure from optimal library fragment sizes, variation in the proportions of duplicate sequences introduced by PCR amplification bias, and contamination from known and unknown species other than the sequencing target (Endrullat et al., 2016; Trivedi et al., 2014).

Hence, the generated output from a sequencing run must be analyzed to control the quality of data. For assessing the overall quality of a sequencing run, there are several tools compatible with FASTQ format such as FASTQC¹, NGS QC Toolkit (Patel and Jain, 2012), QC-Chain (Zhou et al., 2013) and ChromaPipe (Otto et al., 2008). The FASTQC tool is the most used due to the quality report of important characteristics of NGS data, which includes boxplots, line plots, and a colored code indicating if there is any quality problem at each level. Other tools have similar features, such as the mapping rate of the expected target, levels of fragment or sequence duplication and estimates of the library insert sizes, but the analysis produced by FASTQC is fast, easy to use and interpret, which leads to it being the most used quality control tool (Andrews and al., 2010; Trivedi et al., 2014).

3.4.2 *Undesired Sequences Removal*

The generated output from a sequencing run can contain undesirable sequences, such as reads with reduced dimensions, reads with a low confidence level, reads from species that are not the focus of the study, sequences of RNA or artificial sequences from the experimental work. Trimming is the process where the sequences of less interest are removed (Fabbro et al., 2013).

¹ <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

There are several tools available for removing undesired sequences such as Trimmomatic (Bolger et al., 2014), SolexaQA (Cox et al., 2010) and FastX². Trimmomatic is the most popular toolbox for tailoring NGS data. An example of tool for removing human and other hosts derived sequences is BMTagger (Rotmistrovsky and Agarwala, 2011).

3.5 DATA ANALYSIS

During the data analysis, the sequence reads are aligned to a reference genome using bioinformatics tools. This process is designated by read mapping. After alignment, differences between the reference genome and the newly sequenced reads can be identified, through the detection of Single-Nucleotide Polymorphisms (SNPs). The following sections explain the steps and most common tools used during data analysis.

3.5.1 Read Mapping

NGS experiments can produce numerous sequence reads with biological meaning and context. Consequently, the sequences need to be mapped into a reference genome. This is achievable through sequence alignment, that makes possible the annotation of an unknown sequence through known sequences and still to detect differences between sequences (Cliften, 2014). One of the most known alignment algorithm is the Basic Local Alignment Search Tool (BLAST), which compares one or more unknown sequences to a large database of annotated sequences (Altschul et al., 1990).

Due to the large amount of raw data produced by NGS experiments, robust and fast algorithms are required for analysing NGS reads. The alignment algorithms are constructed in auxiliary data structures, called indices and can be divided into two groups: algorithms based on hash tables and algorithms based on the suffix/prefix trees. A hash table is a data structure that can map keys to values and uses a hash function to compute an index into an array, from which the desired value can be found. A suffix/prefix tree is a tree data structure, in which each node has a key. In the prefix tree, all the descendants of a node have a common prefix of the text associated with that node. Suffix trees contain all the suffixes of the given text as their keys and positions in the text as their value (Li and Homer, 2010).

Several algorithms that have been developed to map NGS data into a genome. Despite that fact, in this work, is considered the following algorithms: MAQ, Bowtie, BWA and Novoalign (Li and Homer, 2010).

² http://hannonlab.cshl.edu/fastx_toolkit

MAQ is categorized as an algorithm based on hash tables and uses the base quality scores to improve the accuracy of the alignment. This algorithm can simultaneously make alignments and variant detection (Cliften, 2014; Li and Homer, 2010; Li et al., 2008a).

Bowtie is an algorithm based on suffix trees and it is based on the Burrows-Wheeler transform³. It can be used for quick alignment of NGS reads, however it has limitations in the detection of variants (Cliften, 2014; Langmead et al., 2009; Li and Homer, 2010).

BWA, like Bowtie, is an algorithm-based on suffix trees and in the Burrow Wheeler transform. This algorithm can produce very accurate alignments and it is suitable for downstream variant analysis (Cliften, 2014; Li and Durbin, 2009; Li and Homer, 2010).

Novoalign⁴ is a commercial software package (Novocraft Technologies) based on suffix trees and Burrows-Wheeler transform. In addition to making the alignment and variant detection, the software also provides a number of useful features such as adapter and primer trimming and the alignment of bisulfite treated DNA (mapped sites of methylation) (Cliften, 2014; Li and Homer, 2010).

3.5.2 Detection of Single-Nucleotide Polymorphism

A SNP is a variation in a single nucleotide that occurs at a specific position in the genome and affects at least 1% in a random set of individuals in a population. If it affects less than 1% of the population, it is considered a mutation. The polymorphisms have a great importance because they represent an important part of the genetic variation between individuals in a population (Brookes, 1999).

The detection of SNPs in NGS data is divided into three phases: the data preparation, the Bayesian approach and the annotation (Magi et al., 2010). In the first step, the data are evaluated and filtered. In other words, the data are submitted for preprocessing, in which the paralogs or repeated sequences are discarded or considered (if there is a supporting evidence), and the quality values are reassigned (Magi et al., 2010). For the evaluation of data, there are several tools, where the most frequently used are SAMtools (Li and Durbin, 2009), GATK (Mckenna et al., 2010), SOAPsnp (Li et al., 2009a), and Sniper (Simola and Kim, 2011). For filtering, the main tools are SnpSift (Cingolani et al., 2012a) and VcfTools (Danecek et al., 2011).

The second step is the application of the Bayesian approach in the filtered data. This approach consists of computing the conditional likelihood of the nucleotides at each position by using the Bayes rule⁵. The tools that use a Bayesian approach are PolyBayes (Marth et al., 1999), SOAPsnp (Li et al., 2009b) and MAQ (Li et al., 2008b).

³ It is an algorithm used for rearrangement of the string in similar characters in the form of compact data structures.

⁴ <http://www.novocraft.com/>

⁵ Describes the probability of an event based on prior knowledge of conditions that might be related to the event.

The last step is the annotation of SNPs, that consists in giving a biological meaning to polymorphisms. For this process, the tool most used is SnpEff (Cingolani et al., 2012b). Other tools are ANNOVAR (Wang et al., 2010) and Vep (McClaren et al., 2010).

3.6 APPLICATION IN CLINICAL MYCOLOGY

In the last years, NGS has been applied in several clinical microbiology laboratories, as a complementary tool for conventional diagnostic testing. The capacity of NGS technology in sequencing has been increasing. Combining that fact with low costs of its usage, it is standing out as an important tool for microbiologists. The popularity of NGS assays is associated with fast, reproducible and accurate results (Deurenberg et al., 2017; Zoll et al., 2016).

In clinical microbiology, NGS is used for outbreak management, molecular case finding, characterization and surveillance of pathogens, rapid identification of microbial strains, taxonomy and metagenomics approaches on clinical samples (Deurenberg et al., 2017; Ko et al., 2014).

However, in the case of fungal species, the platforms of sequencing require a capacity of at least 10^{12} to 10^{14} nucleotides per sequencing run per sample. The application of WGS for the determination of mycobiomes for diagnostics in clinical samples is not possible for currently available sequencers like Illumina NextSeq and HiSeq. Hence, the information about mycobiomes is obtained by target sequencing of amplicons of the ITS region of the fungal ribosomal genes (Zoll et al., 2016).

Bittinger et al. (2014) analyzed fungal and bacterial species present in the human airway, through target sequencing of ribosomal RNA gene segments from fungi and bacteria. They compared fungal and bacterial communities from healthy subjects, HIV+ subjects, and lung transplant recipients. This work enabled the identification of fungal communities in the human respiratory tract and their interactions with bacterial communities in health and disease. Thus, it has been possible to trace a community pattern associated with pathogenic polymicrobial biofilms.

Besides the taxonomic identification of fungal species, another potentiality of NGS in mycology is the characterization of mechanisms of resistance, through the analysis of SNPs or other mutations in genes commonly involved in antifungal resistance (Zoll et al., 2016).

Sertour et al. (2015) investigated echinocandin and azole resistance in clinical *Candida* isolates using NGS approaches. For this, they analyzed six genes commonly involved in antifungal resistance (ERG11, ERG3, TAC1, CgPDR1, FKS1 and FKS2) through target sequencing of amplicons. As a result, they obtained a total of 391 SNPs in several genes associated with the antifungal resistance and new genetic alterations were detected. So, using NGS for detection of mechanisms of resistance in clinical strains allowed a directional

search, avoiding the usage of extensive molecular techniques, that are time-consuming and expensive comparing with the NGS assays.

In the work developed by Camps et al. (2012), four isogenic *Aspergillus fumigatus* isolates were collected, in which two isolates were azole-resistant due to prolonged azole treatment. The genomes of the resistant isolate and the wild-type *cyp51A* (related with lanosterol 14- α -demethylase) were sequenced using NGS to identify the resistance-conferring mutation. As a result, they identified several potential non-synonymous mutations in protein-coding regions. As several mutations had emerged in the isogenic isolates, sexual crossing experiments with a selection of the progeny on azole resistance phenotype were required to show that azole resistance was associated with a P88L amino acid substitution in the CCAAT-binding transcription factor complex subunit HapE. The HapE P88L mutation caused an increasing of expression in *CYP51A* gene. The WGS approach made possible to follow the genetic changes triggered in the *Aspergillus* genome.

NGS offers more tools for the study of genetic changes, aside from WGS or target sequencing of amplicons. An important tool is the analysis of epigenetic changes in the fungal genome. Gene expression is regulated by epigenetic mechanisms like cytosine methylation and hydroxylation. The study of the epigenetic changes can be a useful tool for the early identification of molecular mechanisms of resistance and the analysis of gene expression (Zoll et al., 2016).

Any diagnostic test in clinical laboratory requires analytical and clinical validation, as well as the monitoring and documentation of quality control for regulatory purposes. Thus, the sequencing methods used as diagnostic tools need the same requirements. Currently, NGS technologies used for testing clinical infectious diseases are being performed as laboratory developed tests, due to the absence of approval of any NGS test for clinical microbiology, by the US Food and Drug Administration (FDA). However, the FDA is finalizing the guidance for approving and validating NGS in clinical microbiology for microbial identification and the detection of antimicrobial resistance markers. The approval of NGS by FDA as a tool for diagnostic in clinical microbiology will aid the standardization of specimen handling, library preparation and sequencing. Also, in bioinformatics analysis, it will help in the data storage and interpretation, to ensure the accuracy and reproducibility of NGS-derived genotypic results. The standardization and quality assurance may be important for the implementation of NGS protocols used in research as a diagnostic test and turn NGS assays into a routine tool in clinical laboratory (Lefterova et al., 2015; Luh and Yen, 2018).

3.7 BIOINFORMATICS TOOLS FOR CLINICAL MYCOLOGY

Bioinformatics tools applied to clinical mycology are scarce, whereby, during the search, only tools for taxonomic identification of the fungi were found. However, bioinformatics

tools for detection of antifungal resistance were not discovered. The next sections explain available tools for taxonomic identification of fungi.

3.7.1 *ITSx*

ITSx is a Perl-based software, that extracts the components of the *ITS*₁, 5.8S and *ITS*₂ regions from Sanger and high-throughput sequencing datasets from fungi and nineteen other groups of eukaryotes (Bengtsson-palme et al., 2013).

The input of *ITSx* consists of sequences in FASTA format. The software examines the sequences in the default and reverse orientation. Each sequence is analysed for matches through Hidden Markov Models⁶. These models are applied in the recognition of the *ITS* regions. *ITSx* also allows the elimination of non-*ITS* sequences, and this is particularly useful for amplicon-based NGS datasets. The results obtained about *ITS* regions are provided in FASTA and CSV file formats (Bengtsson-palme et al., 2013).

3.7.2 *CLoVR-ITS*

CLoVR-ITS is a Standard Operating Procedure (SOP)⁷ implemented in Cloud Virtual Resource (CLoVR) framework. It implements an automated pipeline for *ITS* sequence analysis from metagenomics DNA isolates (White et al., 2013).

The *CLoVR-ITS* includes the preprocessing of the sequences, i.e., detection and removal of chimera sequences, quality control of the sequences, clustering of sequences into Operational Taxonomic Unit (OTU)⁸, and uses Hidden Markov Models for identifying *ITS* regions from GenBank sequence entries. *CLoVR-ITS* provides FASTA, FASTQC (quality report of sequencing) and PDF formats as results of the comparative analysis of hundreds of samples (White et al., 2013).

3.7.3 *FHiTINGS*

FHiTINGS is an open-source software, implemented using Python, that uses the output of a BLAST Nucleotide (BLASTn) search to identify, classify, and parse *ITS* DNA sequences obtained from NGS assays. This tool is appropriate for any sequencing platform and allows BLAST searches against the indicated *ITS* sequences (Dannemiller et al., 2014).

6 Probabilistic models that capture the statistical regularities of the sequences, through the multiple alignments or even in a single sequence.

7 Set of step-by-step instructions compiled by an organization to help workers carry out complex routine operations.

8 Refers to a cluster of organisms, grouped by DNA sequence similarity of a specific taxonomic marker gene.

FHiTINGS uses the [Lowest Common Ancestor \(LCA\)](#)⁹ method as a sequence identification method. Thus, it produces a single identification from [BLAST](#) output results and the taxonomy is assigned based on the Index Fungorum database¹⁰. FHiTINGS allows the simultaneous processing of many samples and the results are provided in a user-friendly way ([Dannemiller et al., 2014](#)).

3.7.4 PIPITS

PIPITS is a Python open-source software for automated processing of Illumina MiSeq sequences, allowing the analysis of fungal [ITS](#) sequences ([Gweon et al., 2015](#)).

PIPITS proceeds to the sequence preparation before the extraction of the [ITS](#) sequences and it includes the joining of the paired-end reads, a quality filter, and conversion of file formats. The extraction of the [ITS](#) regions is performed by the [ITSx](#) tool ([Bengtsson-palme et al., 2013](#)). The [ITS](#) sequences are taxonomically assigned with the [Ribosomal Database Project \(RDP\) Classifier](#)¹¹ ([Wang et al., 2007](#)) against the UNITE ¹² fungal [ITS](#) reference dataset. As result, the PIPITS returns [OTU](#) tables taxonomically annotated in a classical tabular and BIOM format ([Mcdonald et al., 2012](#)).

3.7.5 FindFungi

FindFungi is an open-source software, developed in Python, that can be used to identify fungi from available metagenomics datasets ([Donovan et al., 2018](#)).

In this tool, the low-quality reads are removed and the remaining reads are converted into FASTA format, which is analyzed afterwards by 32 implementations of [Kraken](#)¹³. The 32 [Kraken](#) predictions for each fungal read are consolidated and a consensus prediction is assigned. The best hit for each read is mapped to a pseudo-assembly of the relevant genome using [BLAST](#). After this, there is the implementation of statistical metrics such as Pearson's coefficient of skewness. Finally, the fungal predictions are written to PDF and CSV files ([Donovan et al., 2018](#)).

⁹ In computer science and graph theory, the lowest common ancestor of two nodes v and w in a tree or directed acyclic graph T is the lowest node that has both v and w as descendants.

¹⁰ <http://www.indexfungorum.org/>

¹¹ Allows the classification of the sequences from bacteria and fungi sequences based on the Ribosomal Database Project, using a naïve Bayesian classifier.

¹² <https://unite.ut.ee/index.php>

¹³ Sequence classification algorithm, where each k -mer of the sequence is mapped based on the lowest common ancestor. Each root-to-leaf path in the classification tree is scored by adding all weights in the path, and the maximal root-to-leaf path in the classification tree is the classification path. The leaf of this classification path is the classification used for the query sequence.

3.7.6 Comparison of Tools

After an extensive analysis of the main characteristics of the tools mentioned above, a summary of the differences and similarities between them was created, being presented in Table 3.

As a result, it was possible to choose the most adequate tool to integrate into the framework proposed. The choice was PIPITS due to the following characteristics:

1. Free access and available in the Bioconda package¹⁴;
2. Does preprocessing of data before the taxonomic identification, which allows the removal of noise of the sequencing, avoiding the errors in the classification of the species;
3. Uses ITSx to extract ITS regions and it is presented as the ideal extractor of ITS regions according to the specifications of the tool and the accuracy of the results (Gweon et al., 2015);
4. Uses the RDP classifier to assign the taxonomy of the species, that has the particularity of using a curated database with information of many fungi species;
5. Provides command-line manipulation, making it easy and intuitive to use.

However, this tool has one limitation: it only executes in the UNIX-based platform. This is considered irrelevant for PIPITS use in the solution, since there are ways of making the framework isolated from the underlying platform, e.g., by using containers¹⁵.

¹⁴ <https://bioconda.github.io/recipes/pipits/README.html>

¹⁵ A standard unit of software that contains all dependencies needed for the application to run successfully.

Table 3.: Analysis of the available tools for taxonomic identification.

Tools	ITSx	FHITINGS	FindFungi	PIPITS	CLoVR-ITS
Programming language	Perl	Python	Python	Python	Not specified
Operating system	UNIX-based platform	UNIX-based platform	UNIX-based platform	UNIX-based platform	Cloud Virtual Resource Framework
Type of system	Software	Software	Script	Script	Software
Interface with user	Command-line	Graphical User Interface	Command-line	Command-line	Graphical User Interface
User guide	Yes	No	No	Yes	Yes
Access	Free	Free	Free	Free	Subject to payment
Input	FASTA file	BLAST output	FASTQ file	FASTQ file	SFF file FASTQ file FASTA file Nucleotide BLAST database file Protein BLAST database file Metagenomics mapping file
Type of sequencing supported	Sanger NGS	NGS	NGS	NGS	NGS
Taxonomic identification	Hidden Markov models	Lowest Common Ancestor	Lowest Common Ancestor Kraken	ITSx software VSEARCH RDP Classifier	Hidden Markov models
Other functionalities	Removal of non-ITS sequences	-	-	Join paired-end reads Removal of chimera sequences Quality control of sequencing Conversion of file formats	Detection and removal of chimera sequences Quality control of sequencing
Dependencies	HMMER 3	Index Fungorum database BLAST	BLAST	VSEARCH RDP Classifier 2.10 ITSx BIOM format v.1.3 PEAR FASTX-toolkit	VirtualBox VMware Player DIAG Amazon EC2
Output	FASTA file TXT file CSV file	Taxonomic rank file	PDF file CSV file	FASTA file TXT file	FASTA file FASTQC file GenBank Flat File format Database file PDF file BER file SQL dump file

FRAMEWORK GENESIS

This chapter addresses the specification and constraints associated to the development of the solution, which includes the definition of the functional and non-functional requirements, the construction of the domain model, the design of the architecture of the framework, the definition of conceptual classes through the use cases realization, and finally, the description of the solution developed.

4.1 CONCEPTION

The conception phase deals with the gathering of requirements for the solution, associated functionalities, constraints and the domain of the problem. In other words, this phase infers upon what needs to be implemented.

4.1.1 *Requirements*

The requirements describe what the system should do, who is interacting with it, how it should react and behave in specific situations.

The functional requirements are described through the use cases, represented by use case diagrams, designed using [Unified Modeling Language \(UML\)](#). Using this approach makes the requirements easily understandable. On the other hand, there are other requirements, related to quality parameters of the solution, such as reliability, usability and storage – non-functional requirements – which are also relevant for the scope of this work.

The framework has two identified actors: the user and the system. The user can execute all functionalities of the system. The system executes all commands introduced by the user. Regarding their actions, [Figure 4](#) describes a set of actions, that the system can perform in collaboration with the actors. These are the identified use cases.

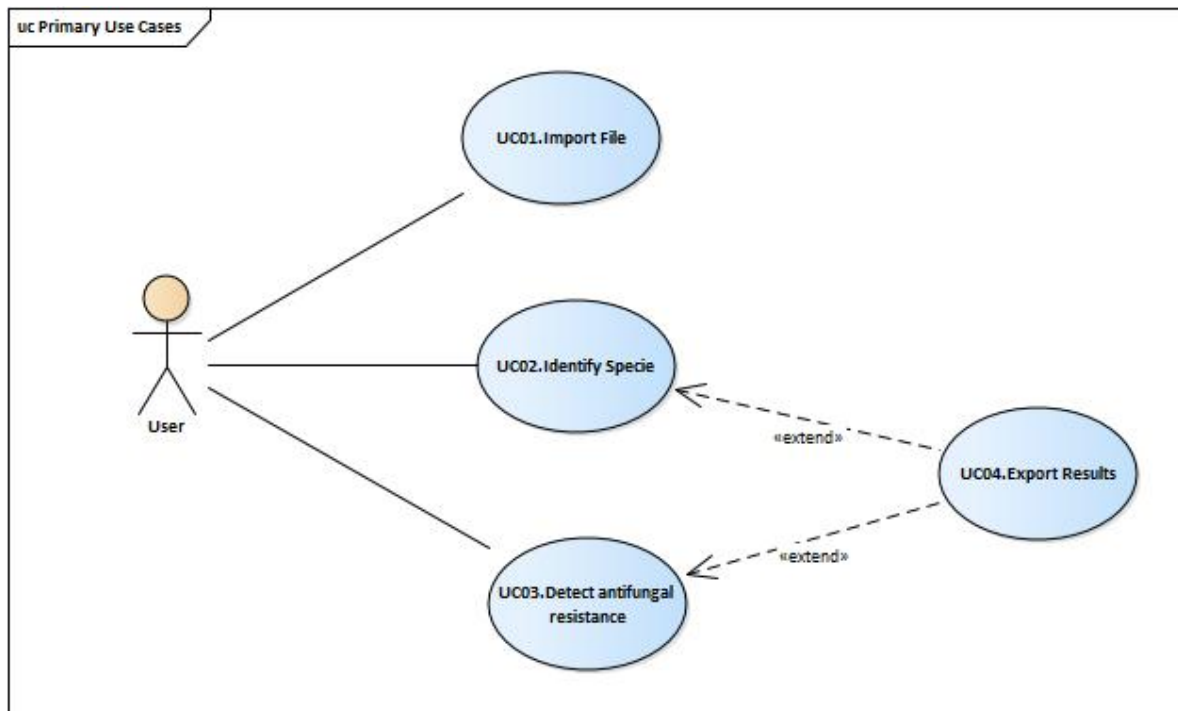


Figure 4.: Representation of the functional requirements using use case diagram. Designed in Enterprise Architecture.

UCo1 - Import Sequencing File

The user imports the file to the system. The system validates the file and verifies its format. The system reads the file and informs the user of the success of the operation.

UCo2 - Identify Specie

The user selects the functionality *Identify specie*. The system executes the functionality selected and informs the user of the success of the operation.

UCo3 - Detect Antifungal Resistance

The user selects the functionality *Detect antifungal resistance*. The system executes the functionality selected and informs the user of the success of the operation.

UCo4 - Export Results

The user can select the option *Export results* to obtain a file with the results related with the UC02 - file with the results of the specie identification - and UC03 - file with the results of the detection of antifungal resistance. The system executes the functionality selected and informs the user of the success of the operation.

Regarding the non-functional requirements, the ones identified and considered relevant to the solution are presented over the next sections.

Functionality

It is necessary to ensure the uniformity of the language of the system. In the framework proposed, it will be English.

Usability

Considering the diversity of users, it is imperative that the framework should have:

- Presentation of the system errors with succinct messages about the problem;
- *Help* button to clarify each functionality of the framework.

Interface

The interface with the user should be:

- Simple with smooth colors and enjoyable letter types;
- Easy access, intuitive and consistent during all execution;
- Guarantee the needs of the users;
- Show default options avoiding the user from introducing wrong choices.

Implementation

It should be considered the usage of design patterns to certify the:

- Development of a cohesive system with low coupling;
- Scalability of the system;
- Fast communication between different parts of the system.

4.1.2 *Domain Model*

The domain model is a visual representation of the conceptual classes or real objects of the domain of the system, represented using [UML](#) diagrams. It allows the identification of the main objects, their features, constraints and relationships. [Figure 5](#) shows the domain model of the solution.

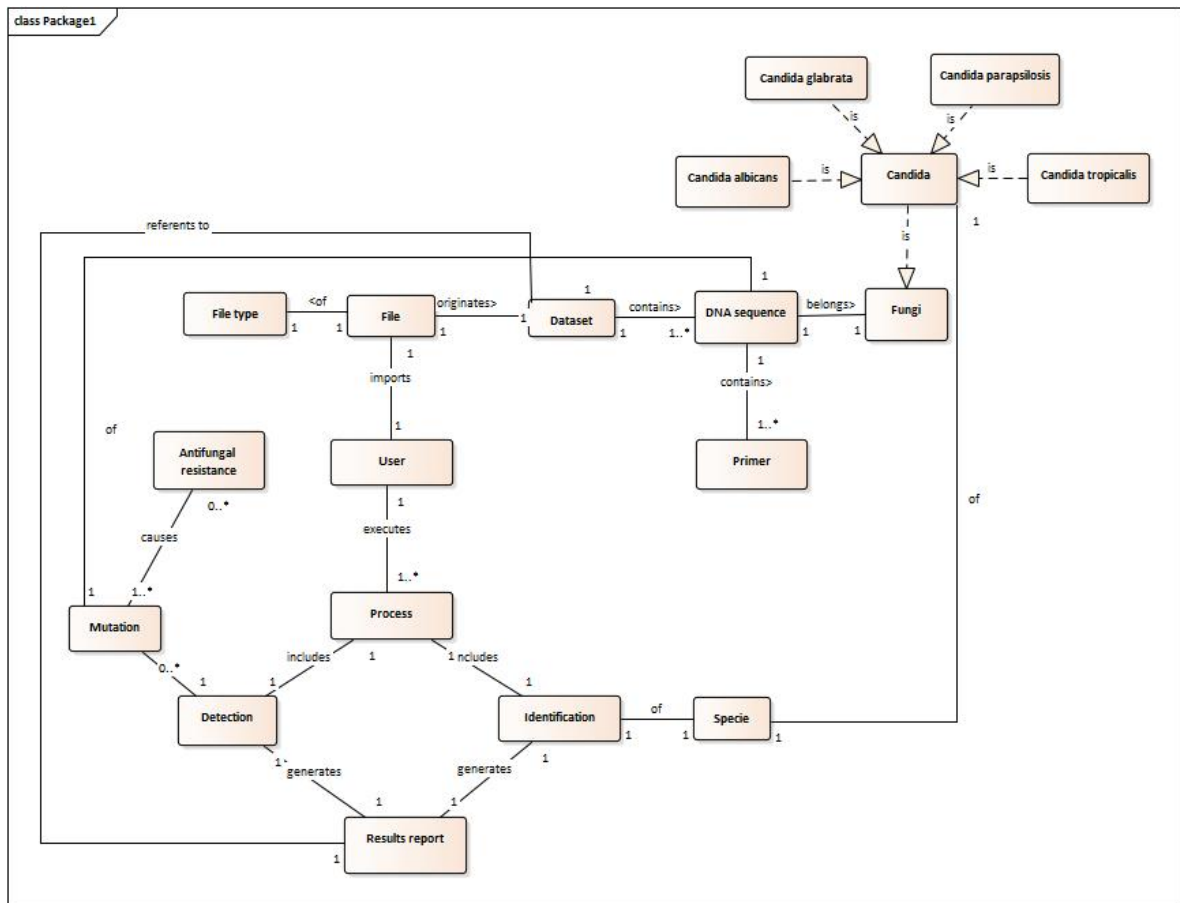


Figure 5.: Representation of the domain model. Designed in Enterprise Architecture.

The user imports a file into the system. The file has a specific format and originates a dataset which contains information about DNA sequences resulting from DNA sequencing. The DNA sequences contain primers resulting from the sequencing assays. The DNA sequences belong to species of the Fungi Kingdom, particularly to *Candida* species. The species of *Candida* are *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*.

By using this framework, the user can execute different processes. The process of the identification allows the identification of the specie belonging to *Candida* species. The process of detection allows to detect mutations present in the DNA sequences and their association with antifungal resistance. Finally, it is generated a report with the results of each process.

4.1.3 Architecture

In the design of the architecture, the organization of the system is established to satisfy the functional and non-functional requirements. The architecture is based on architectural

patterns, that are a reusable solution to a commonly occurring problem in software architectures, within a given context.

The architecture of the solution was designed using a layered architecture pattern, organizing the framework into layers, where each layer provides specific services. Figure 6 shows the logic view of the framework. The diagram is the mode of interaction of the different components. Each component represents a modular part of a system, with a specific responsibility and logic.

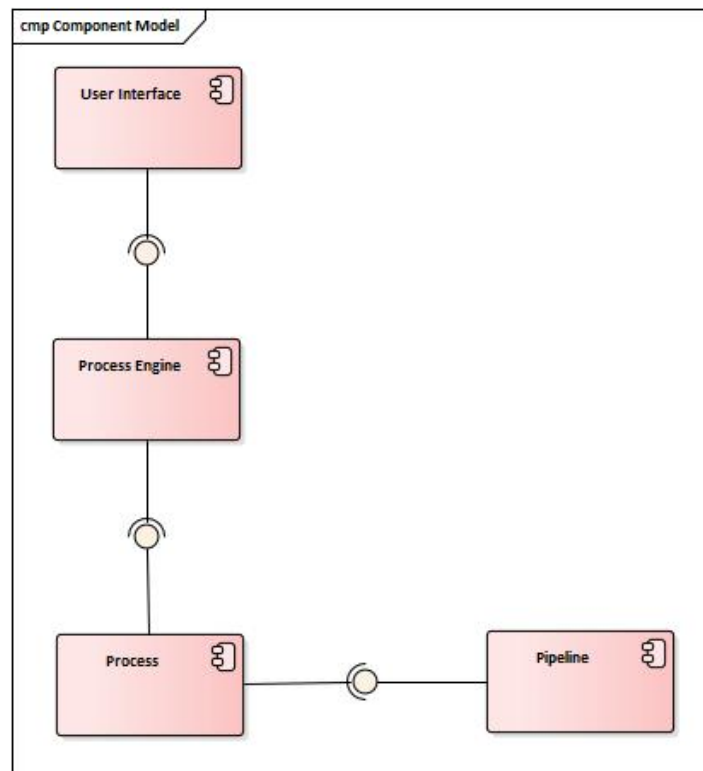


Figure 6.: Logic view. Designed in Enterprise Architecture.

So, it is possible to identify four components:

- **User interface** – displays and collects the data introduced by the user;
- **Process engine** – handles process management and distribution of the data to each process;
- **Process** – encapsulates the logic of each process, one for each functionality of the framework;
- **Pipeline** – represents the set of steps to perform a process.

4.2 DEVELOPMENT

After the analysis of requirements and identification of conceptual classes of the system, it is necessary to detail the implementation of the system, i.e., the definition of the classes/objects of the software and their methods, and description of the dynamic structure of the system, which shows the interaction between objects.

The concepts mentioned previously are described in the next sections, through the usage of sequence and class diagrams, designed in [UML](#).

4.2.1 Package Structure

Figure 7 shows the packages view of the framework. The diagram represents the structure of the solution at the level of packages and the interaction between them. Packages are used to group elements that are related in terms of functionality.

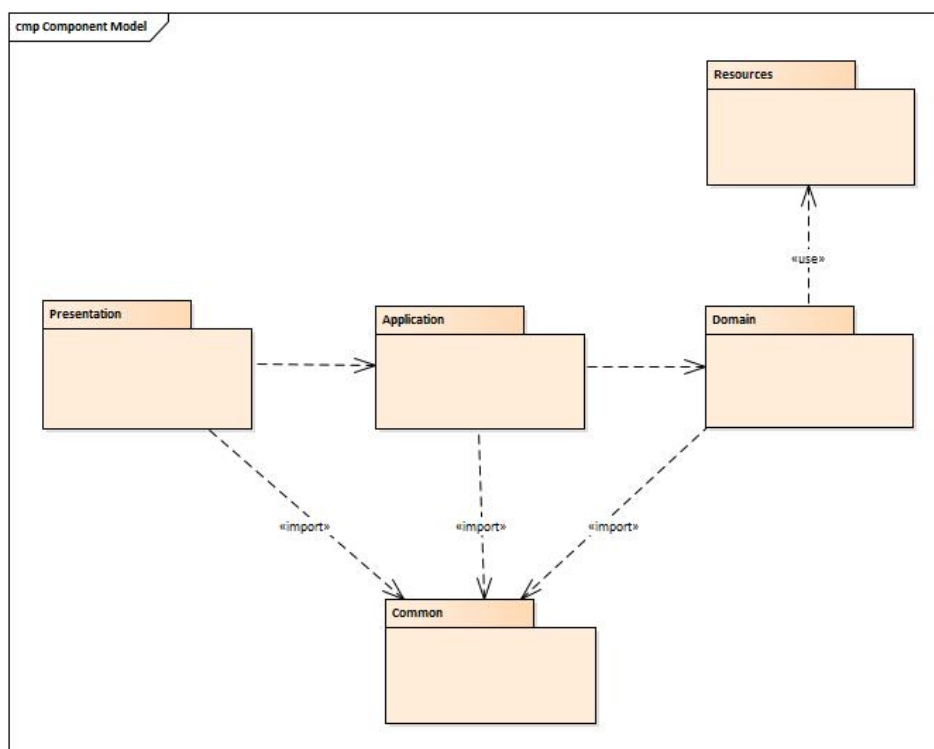


Figure 7.: Packages view of the framework. Designed in Enterprise Architecture.

So, the description of each package is the following:

- **Presentation** – includes modules with the user interface logic, responsible for interaction with the user;

- **Application** – responsible for the communication between the presentation and domain layers;
- **Domain** – represents all the classes with the domain logic. The interaction between them allows the execution of each functionality. This package uses the resources package to extract data;
- **Common** – includes modules that are common to all packages;
- **Resources** – behaves as a repository of data.

4.2.2 Application Start

Figure 8 shows the interaction between objects during the application start. The user initializes the system through the introduction of commands in the command line. The commands are interpreted by a specific module of Python - *argparse* - and converted into arguments that will be used in the next steps. At the same time, the system reads the configuration file using a specific module of Python - *configparser* - and the configurations are transmitted during the execution of the system.

Figure 9 shows the structure of the application start at the level of classes, represented in the form of a class diagram. The class *Application* has the logic of the initialization of the framework and prepares the execution environment. This class aggregates the class *Configuration*, that has the settings of the framework.

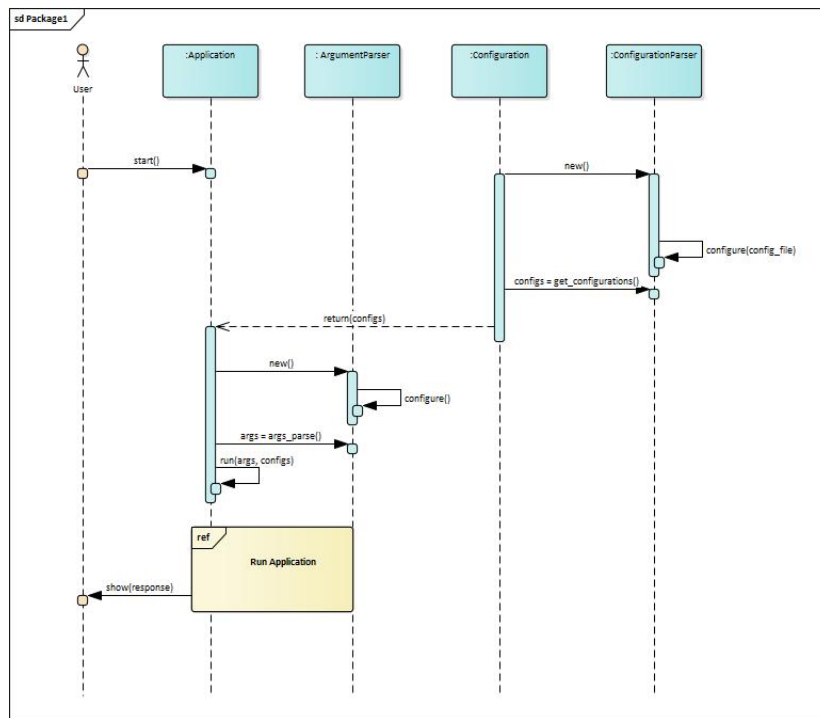


Figure 8.: Sequence diagram of the application start. Designed in Enterprise Architecture.

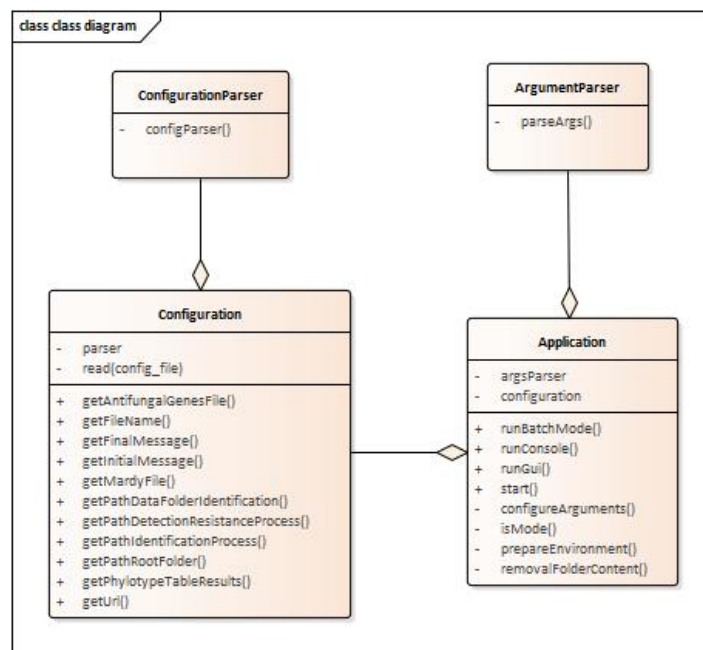


Figure 9.: Class diagram of the application start. Designed in Enterprise Architecture.

4.2.3 Application Run

Figure 10 shows the interaction between objects during the application run. The system provides two modes of interaction with the user: **Graphical User Interface (GUI)** and console view. The options selected by the user are transformed into a dictionary with the instructions to execute. The instructions are transmitted to the controller, which operates as a mediator between the user and the system.

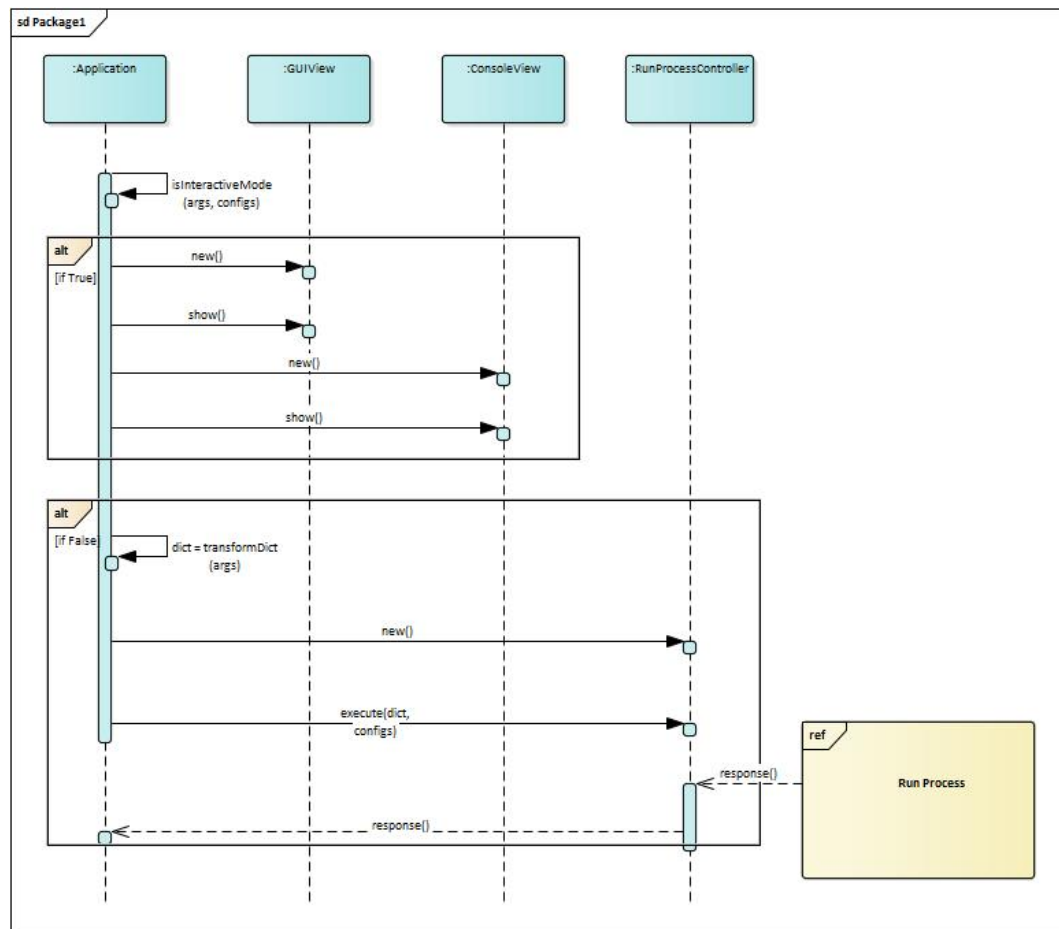


Figure 10.: Sequence diagram of the application run. Designed in Enterprise Architecture.

Figure 11 shows the class diagram of the application run. The class *Application* besides the responsibilities mentioned in the subsection **Application Start**, has the role of instantiating the class *RunProcessController*, *ConsoleView* and *GuiView*. The class *RunProcessController* represents the controllers of the framework. The class *ConsoleView* and *GuiView* have the behavior of user interface, represented by the interface *IUserInterface*. *PyQt5* module was used to implement the class *GuiView*.

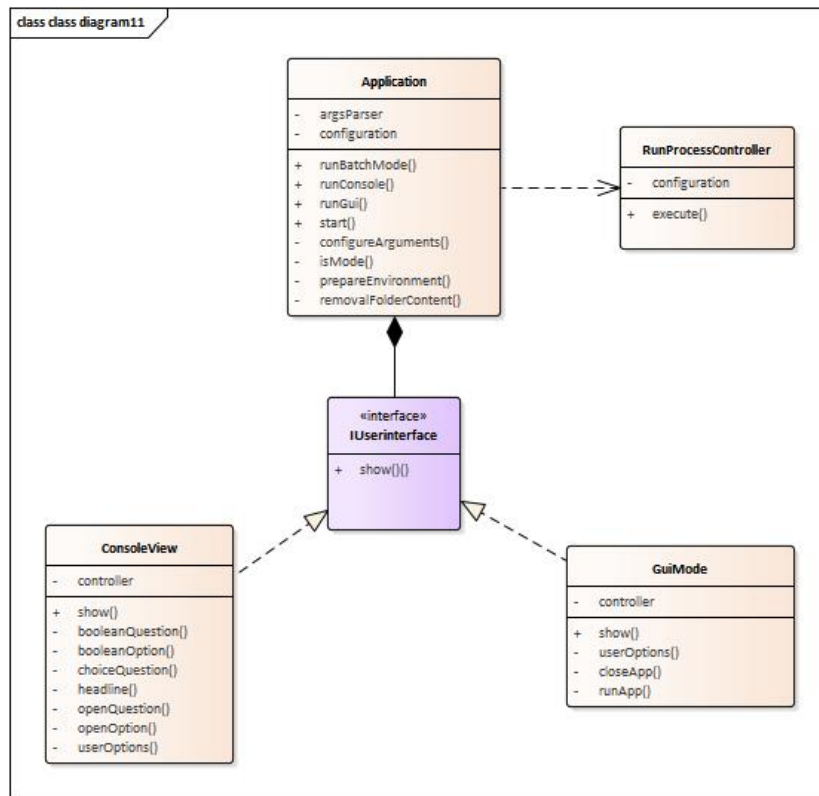


Figure 11.: Class diagram of the application run. Designed in Enterprise Architecture.

4.2.4 Process Run

Figure 12 shows the interaction between objects during the process run. The controller sends the instructions for the process, which can be the specie identification or antifungal resistance process. Each process adds the steps to execute according to the instructions. The response of each step is returned for the referring process. The process sends the response to the controller and this one sends it to the class *GuiView* or the class *ConsoleView*, according to the mode of interaction selected. One of these classes will show the response for the user.

Figure 13 shows the class diagram of the specie identification process. The class *RunProcessController* instantiates the class *IdentificationSpecieProcess*, and this class adds the steps that will allow the execution of the process. The steps are the class *FileImported* and the class *Taxonomy*. The class *FileImported* imports sequencing files for the framework. The class *Taxonomy* allows the integration of pipelines to identify species. On the other hand, the class *Taxonomy* aggregates the class *PipitsPipeline*, that integrates the PIPITS tool.

Figure 14 shows the class diagram of the antifungal resistance process. The class *RunProcessController* instantiates the class *DetectionResistanceProcess*, and this adds the pipeline that will allow the execution of the process. It was developed a new pipeline – class *Anti-*

fungusResistancePipeline – for the analysis and treatment of mutations present in the ERG11 and FKS1 genes in *Candida* species. This pipeline is composed by seven steps. The steps are implemented in the following classes:

- The class *FileRead* is responsible for reading files with the TXT, FASTA and CSV extension;
- The class *InformationExtraction* extracts the information from the [Mycology Antifungal Resistance Database](#);
- The class *SequenceWithoutPrimer* removes the primers from the DNA sequence;
- The class *SequenceTrimmed* trims the DNA sequences;
- The class *Translation* translates the DNA sequence to amino acid sequence;
- The class *Mutation* identifies the mutations present in the amino acid sequence;
- The class *AntifungalResistance* identifies if the species has antifungal resistance, through the evaluation of the mutations identified in the class *Mutation*.

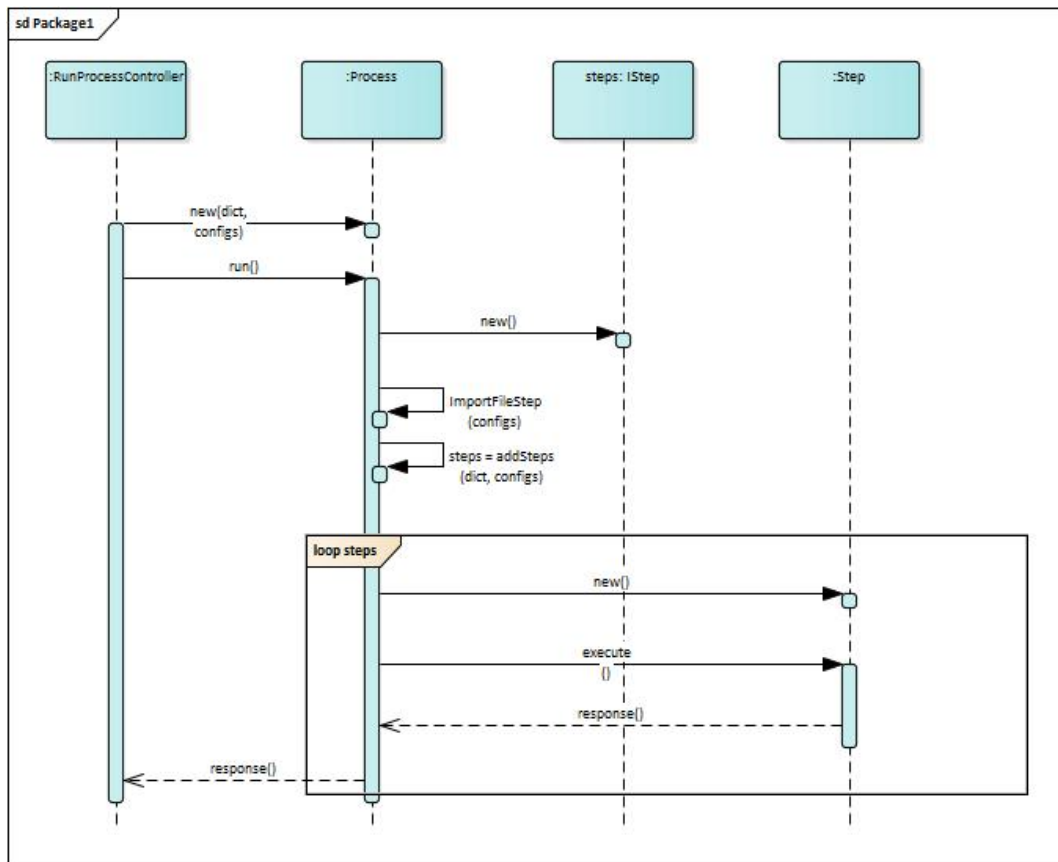


Figure 12.: Sequence diagram of the process run. Designed in Enterprise Architecture.

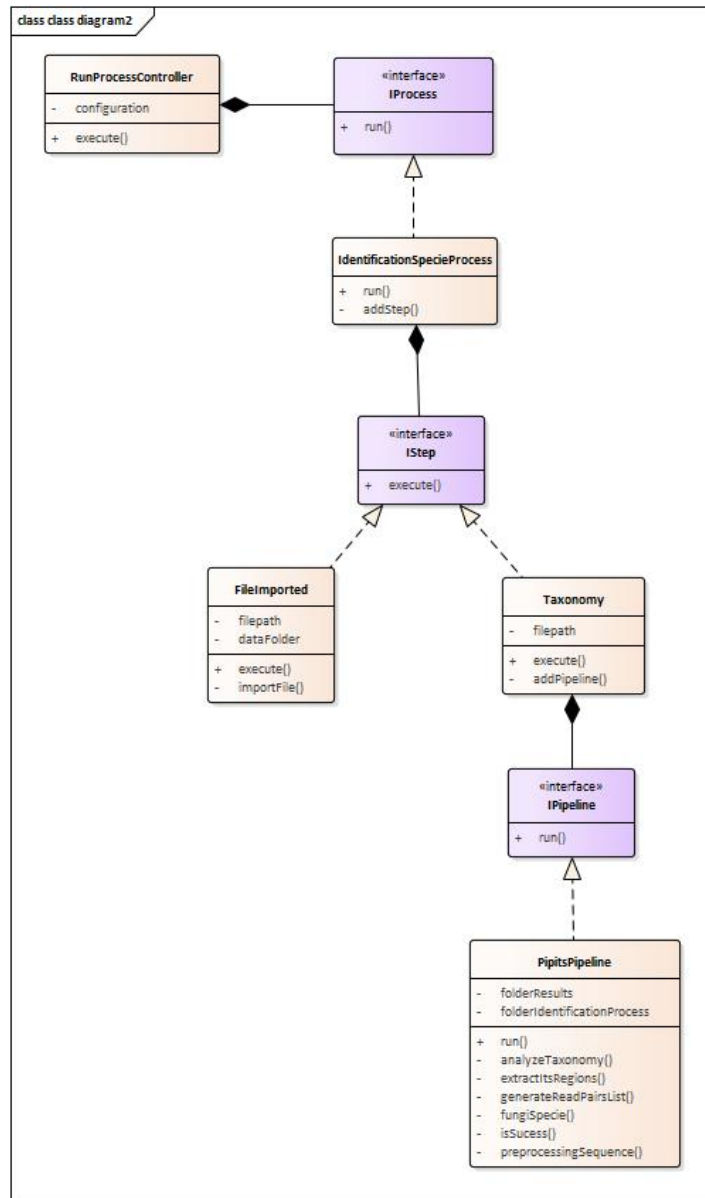


Figure 13.: Class diagram of the specie identification process. Designed in Enterprise Architecture.

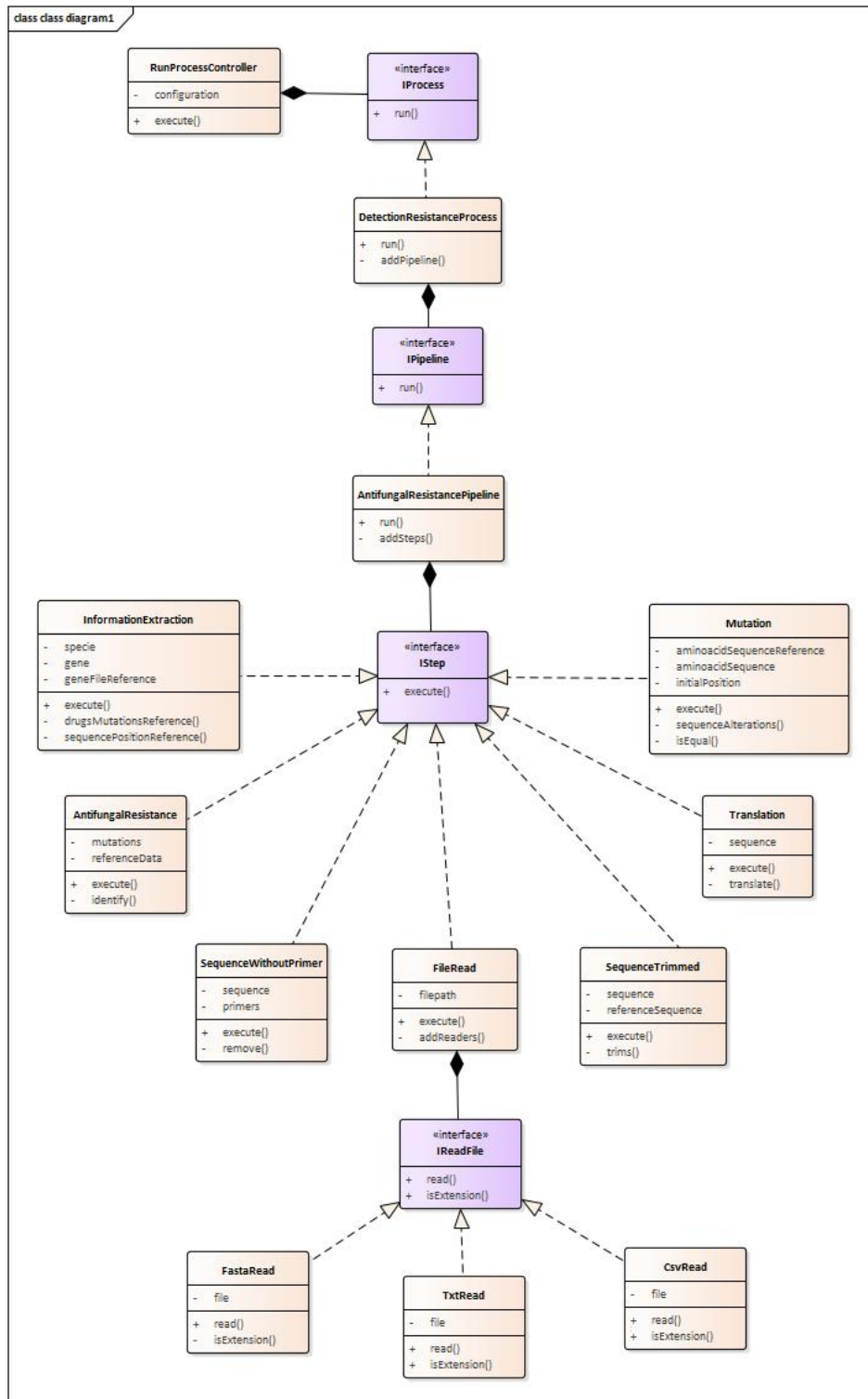


Figure 14.: Class diagram of the antifungal resistance detection process. Designed in Enterprise Architecture.

4.3 DEMONSTRATION

The framework - *Identification and Detection - Candida sp* - is a tool that allows the identification of fungi species, and the detection of the antifungal resistance in *Candida* species. It has two modes of interaction with the user: console and GUI. These modes are described in the next sections.

For the *Identify specie* functionality, the PIPITS tool can identify a high range of fungi species. This functionality only accepts sequencing files with FASTQ extension from the Illumina sequencing platform.

For the *Detect antifungal resistance* functionality, the available species are: *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis*, and the genes are ERG11 and FKS1. So, the solution presupposes that the user has knowledge about the species and the gene. This functionality accepts files with FASTA and TXT extensions. The results of each functionality are exported to the directory where the input data is kept.

After the execution of each functionality, the console and the GUI show the information about the success or failure of the functionalities selected. In the case of success, the framework displays the following messages:

- *Specie identification was executed with success. Please check the results in the directory of the input data.*
- *Antifungal resistance detection was executed with success. Please check the results in the directory of the input data.*

And in the case of failure, it displays the following messages:

- *It wasn't possible execute the specie identification.*
- *It wasn't possible execute the detection of antifungal resistance.*

Oliveira (2019) presents the source code of the framework resultant of the genesis process exposed in this chapter.

4.3.1 Console Mode

Figure 15 shows an example of interaction with the console. The console provides the options of each functionality, and the user chooses the options according his preferences.

In Figure 15, the user chooses *Identify specie* option through typing of *y*, and thus gives the directory of the files that he desires to identify. The *Detect antifungal resistance* option follows the same logic, but the user needs to select the specie and the gene, through typing of the corresponding number. Finally, he writes the forward and the reverse primers.

If the user does not want to execute some functionality, he needs to write *n* and the console does not exhibit the respective options.

Before the execution, the user is questioned about the continuation of the execution, through the question *Will you continue executing other pipelines?*, and the user writes *y* or *n*, according to his preference.

The console informs the user of success or failure of the functionalities selected through the display of the message in the command-line.

```
(base) paty@DESKTOP-NJL87FU:/mnt/c/Users/anapatricia/Documents/dissertation$ python -m framework -i console
Welcome to Identification and Detection - Framework!
#####
### Identification and Detection - Candida sp ###
#####
Identify specie [Y|n]: y
File directory: C:\Users\anapatricia\Documents\testing_specie_identification\data_from_pipits_test
Detect antifungal resistance? [Y|n]: y
File directory: C:\Users\anapatricia\Documents\testing_antifungal_resistance\test_ctropicalis_erg11.txt
1 - Candida albicans
2 - Candida glabrata
3 - Candida parapsilosis
4 - Candida tropicalis
Choose the specie: 4
1 - ERG11
2 - FKS1
Choose the gene: 1
Forward primer: AAAAAAT
Reverse primer: TTTTITA
Will you continue executing other pipelines? [y|N]: n
```

Figure 15.: An example of interaction in console mode of the framework.

4.3.2 Graphical User Interface Mode

Figure 16 shows an example of interaction with the GUI. The GUI displays the functionalities and options that are possible to choose from, and the user selects the options desired.

In Figure 16, the user selects the *Identify species* and *Detect resistance* options through a mouse click on the specific fields. The user gives the directory of the file through the dialog with the operating system. In the case of the *Detect resistance* option, the user needs to select the species and the gene. Each button of the mentioned fields gives a list of valid options. The *Specie* field is composed by the following options: *Candida albicans*, *Candida glabrata*, *Candida tropicalis* and *Candida parapsilosis*. The *Gene* field is composed by *ERG11* and *FKS1* options. Finally, the user writes the forward and the reverse primers.

If the user does not click on the *Identify species* or *Detect resistance* options, the framework does not execute this functionality.

Finally, the user has to click on *Run* button to execute the functionalities selected. The *Cancel* button suspends the execution and closes the framework.

The GUI informs the user of success or failure of the functionalities selected through the messages written in the *Results status* field.

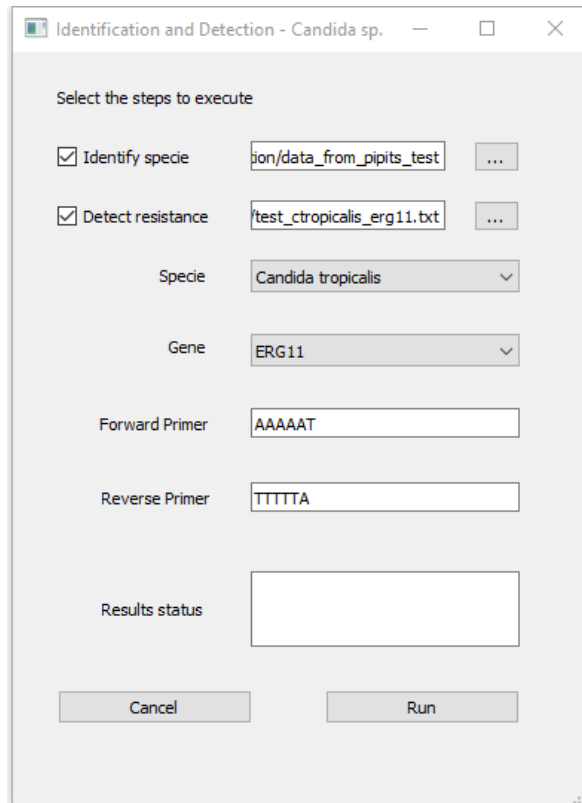


Figure 16.: An example of interaction in GUI mode of the framework.

FRAMEWORK TESTING

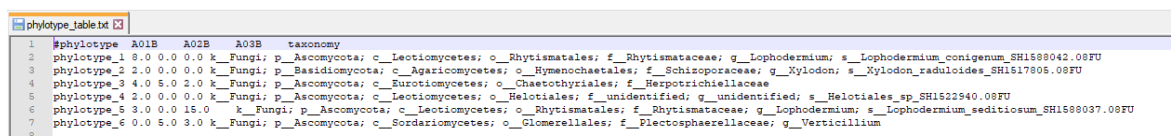
This chapter exposes the testing process of the framework. The main functionalities – *Identify specie* and *Detect antifungal resistance* –, intrinsically related to the major necessities of a solution, were tested and the results are provided in the next sections.

5.1 IDENTIFICATION OF SPECIE

During the search of datasets from NGS assays in the [Sequence Read Archive](#) database, it was very hard to find complete ones from *Candida* species. The datasets selected were tested in the [PIPITS](#) tool, but failed during the identification due to the absence of the ITS₁ and ITS₂ regions. Probably, all the datasets tested do not have the ITS₁ and ITS₂ regions sequenced. Thus, it was used the test datasets from the [PIPITS](#) tool to validate its incorporation in the framework.

Figure 17 shows the species identified from the test datasets. The output from the *Identify specie* functionality is a file in tabular form and in TXT extension. It has two main columns: *phylotype* and *taxonomy*; other columns refer to the names of datasets. The taxonomy of each phylotype is classified from the kingdom to the specie.

In the test datasets, it was identified six different phylotypes: *Lophodermium conigenum*, *Xylodon raduloides* and *Lophodermium seditiosum*; the remaining phylotypes were not possible to identify at the level of species, only at the level of genus: *Herpotrichiellaceae*, *Helotiales* and *Verticillium*.



```

1 #phylotype A01B A02B A03B taxonomy
2 phylotype_1 8.0 0.0 0.0 k_Fungi; p__Ascomycota; c__Leotiomycetes; o__Rhytismatales; f__Rhytismataceae; g__Lophodermium; s__Lophodermium_conigenum_SH1589042.08FU
3 phylotype_2 2.0 0.0 0.0 k_Fungi; p__Basidiomycota; c__Agaricomycetes; o__Hymenochaetales; f__Schizoporaceae; g__Xylodon; s__Xylodon_raduloides_SH1617805.08FU
4 phylotype_3 4.0 5.0 2.0 k_Fungi; p__Ascomycota; c__Eurotiomycetes; o__Chaetothyriales; f__Herpotrichiellaceae
5 phylotype_4 2.0 0.0 0.0 k_Fungi; p__Ascomycota; c__Leotiomycetes; o__Helotiales; f__unidentified; g__unidentified; s__Helotiales_sp_SH1522940.08FU
6 phylotype_5 3.0 0.0 15.0 k_Fungi; p__Ascomycota; c__Leotiomycetes; o__Rhytismatales; f__Rhytismataceae; g__Lophodermium; s__Lophodermium_seditiosum_SH1589037.08FU
7 phylotype_6 0.0 5.0 3.0 k_Fungi; p__Ascomycota; c__Sordariomycetes; o__Glomerellales; f__Plectosphaerellaceae; g__Verticillium
8

```

Figure 17.: An example of the species identification output.

In summary, the *Identify specie* functionality is capable of classifying fungi at different levels of taxonomy, and making species distinctions of the same genus, as it can be seen in Figure 17. But, due to the impossibility to test with data from *Candida* species, it was

not possible to validate the framework with these species. However, PIPITS uses UNITE database – mentioned in the section PIPITS – that stores information about ITS regions from many fungi species, and includes *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis*. So, it is probable that PIPITS can identify *Candida* species in datasets from NGS assays.

5.2 DETECTION OF ANTIFUNGAL RESISTANCE

To test the *Detect antifungal resistance* functionality, data was selected from GenBank, referent to *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis*, and ERG11 and FKS1 genes. All selected DNA sequences are related to antifungal resistance, and have the following accession numbers in NCBI: KR998018.1 (*C. tropicalis* - ERG11), KF211452.1 (*C. glabrata* - FKS1) and KM875725.1 (*C. albicans* - ERG11). To each DNA sequence, a mocking primer sequence was added to imitate a real context of DNA sequencing.

The output of this functionality is a CSV file with three columns: *Reference*, *Position* and *Substitutions*. The *Reference* column is the original amino acid sequence. The *Position* column is the position of the mutation in the amino acid sequence. The *Substitutions* column is the amino acid of the mutated amino acid sequence. After the mutations listing, it follows the information about the antifungal resistance, with the description of the drugs to which the specie is resistant.

If the framework does not identify mutations or antifungal resistance, the results file will contain a line with the following information: *No mutations identified* and *No antifungal resistance detected*, respectively.

Figure 18 shows snippets of the output for three *Candida* species and the corresponding description of the result. In summary, the *Detect antifungal resistance* functionality has the ability to verify if the *Candida* species are resistant or susceptible to antifungals, through the analysis of mutations in the ERG11 and FKS1 genes.

test_calbicans_erg11_results.csv	
1	Reference Position Substitutions
2	V104D
3	F105A
4	N106K
5	A107L
6	K108S
7	L109D
8	S110V
9	D111S
10	V112A
11	S113E
12	A114D
399	Antifungal Resistance
400	Voriconazole
401	Fluconazole
402	Itraconazole
403	

test_ctropicalis_erg11_results.csv	
1	Reference Position Substitutions
2	Q357L
3	K358P
4	P360V
5	L361N
6	V362N
7	N363T
8	N364I
9	T365K
10	I366E
11	K367T
12	E368L
164	Antifungal Resistance
165	No antifungal resistance detected.

(a) *C. albicans* – KM875725.1.

The framework identifies mutations, and these are associated with antifungal resistance, i.e., the specie is resistant to voriconazole, fluconazole and itraconazole.

(b) *C. tropicalis* – KR998018.1.

Mutations are identified, but these are not associated with antifungal resistance.

test_cglabrata_fks1_results.csv	
1	Reference Position Substitutions
2	No mutations identified.
3	
4	Antifungal Resistance
5	No antifungal resistance detected.

(c) *C. glabrata* – KF211452.1.

Mutations were not identified, and consequently, antifungal resistance was not detected.

Figure 18.: Snippet of outputs generated by the framework.

CONCLUSION

This chapter presents the final considerations about the framework developed, its contribution in the context of clinical mycology and the main constraints occurred during its development. And to conclude, it exposes future improvements and prospects.

6.1 FINAL CONSIDERATIONS

The framework developed is the first bioinformatic tool that identifies and detects antifungal resistance in *Candida* species, by using NGS data. It is a user-friendly tool, which does not require programming or bioinformatics skills.

The framework was developed in Python according to the principles of the software engineering and object-oriented paradigm. The solution aggregates two main functionalities: *Identify specie* and *Detect antifungal resistance*. The *Identify specie* functionality allows the identification of a high range of fungi species, and generates a file with the description of the taxonomy of each phylotype identified. On the other hand, the *Detect antifungal resistance* functionality detects the presence of antifungal resistance in FKS1 and ERG11 genes for *C. albicans*, *C. parapsilosis*, *C. glabrata* and *C. tropicalis*, and produces a file with the list of mutations and drugs which the specie is resistant.

The main constraint during the development was the limited access to sequencing data of the *Candida* species, that may have compromised the validation of the results. However, the functionalities and the results are promising, and in the future, this solution can be used as a substitute to conventional laboratory techniques, because the usage of the framework will avoid the cultivation of the species and the search of the mechanisms of resistance *in vivo*, which are time-consuming and expensive.

So, the improvement of the process of identifying species and detecting antifungal resistance in *Candida* species, through the employment of bioinformatics tools, such as the framework developed, will guide clinicians to choose the best therapy to apply, and consequently, reduce the number of deaths caused by IFIs.

6.2 FUTURE PROSPECTS

In terms of future improvements and suggestions, it would be interesting to extend the framework in order to support new genes related to antifungal resistance in *Candida* species, such as ERG3 and FKS2. For this, a new database would be created with information about mutations of the mentioned genes. The framework would use the new database, alongside with the current one ([Mycology Antifungal Resistance Database](#)), to access and extract information about mutations and antifungal resistance.

The expansion of the *Detect antifungal resistance* functionality for other pathogenic fungi species, such as *Aspergillus* species. Indeed, the species of the genus *Aspergillus* cause IFIs with a high mortality rate. So, the incorporation of *Aspergillus* species in the mentioned functionality will allow rapid detection of antifungal resistance, that will improve the clinical interventions and treatments.

And finally, the development of a web platform that allows the execution of the framework. Therefore, open access would be guaranteed and usability of the solution would be improved, e.g., by providing a smartphone application.

BIBLIOGRAPHY

- George J Alangaden. Nosocomial Fungal Infections : Epidemiology , Infection Control , and Prevention. *Infectious Disease Clinics of NA*, 25(1):201–225, 2011. ISSN 0891-5520. doi: 10.1016/j.idc.2010.11.003. URL <http://dx.doi.org/10.1016/j.idc.2010.11.003>.
- S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL <http://www.sciencedirect.com/science/article/pii/S0022283605803602>.
- S Andrews and Et al. Fastqc: a quality control tool for high throughput sequence, 2010.
- W.J. Ansorge. Next generation DNA sequencing techniques and applications. *New Biotechnology*, 27(April):S3, 2010. ISSN 18716784. doi: 10.1016/j.nbt.2010.01.291. URL <http://linkinghub.elsevier.com/retrieve/pii/S1871678410002931>.
- M.C. Arendrup, M. Cuenca-Estrella, C. Lass-Flörl, and W. Hope. EUCAST-AFST. EUCAST technical note on the EUCAST definitive document EDef 7.3: Method for the determination of broth dilution minimum Inhibitory concentrations of antifungal agents for yeasts EDef 7.3 (EUCAST-AFST). *EUCAST E.DEF 7.3*, 2015. ISSN 14713063.
- Johan Bengtsson-palme, Martin Ryberg, Martin Hartmann, Sara Branco, Zheng Wang, and et al. Godhe. Improved software detection and extraction of ITS₁ and ITS₂ from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Microbial Ecology*, pages 914–919, 2013. doi: 10.1111/2041-210X.12073.
- Vaibhav D Bhatt, Monika Patel, and Chaitanya G Joshi. An Insight of Biological Databases Used in Bioinformatics BT. In Gulshan Wadhwa, P Shanmughavel, Atul Kumar Singh, and Jayesh R Bellare, editors, *Current trends in Bioinformatics: An Insight*, pages 3–25. Springer Singapore, Singapore, 2018. ISBN 978-981-10-7483-7. doi: 10.1007/978-981-10-7483-7-1. URL https://doi.org/10.1007/978-981-10-7483-7_{_}1.
- Kyle Bittinger, Emily S Charlson, Elizabeth Loy, David J Shirley, Andrew R Haas, Alice Laughlin, Yanjie Yi, Gary D Wu, James D Lewis, Ian Frank, Edward Cantu, Joshua M Diamond, Jason D Christie, Ronald G Collman, and Frederic D Bushman. Improved characterization of medically relevant fungi in the human respiratory tract using Next-Generation Sequencing. *Genome Biol.*, pages 1–14, 2014.

- Meredith Blackwell. The fungi: 1, 2, 3 ... 5.1 million species? *American Journal of Botany*, 98 (3):426–438, 2011. ISSN 00029122. doi: 10.3732/ajb.1000298.
- Henry M Blumberg, William R Jarvis, J Michael Soucie, Jack E Edwards, Jan E Patterson, Michael A Pfaller, and et al. Rangel-frausto. Risk Factors for Candidal Bloodstream Infections in Surgical Intensive Care Unit Patients : The NEMIS Prospective Multicenter Study. *Clinical Infectious Diseases*, 33:177–186, 2001.
- Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu170.
- Małgorzata Bondaryk, Wiesław Kurzatkowski, and Monika Staniszevska. Antifungal agents commonly used in the superficial and mucosal candidiasis treatment: Mode of action and resistance development. *Postepy Dermatologii i Alergologii*, 30(5):293–301, 2013. ISSN 1642395X. doi: 10.5114/pdia.2013.38358.
- Anthony J Brookes. The essence of SNPs. *Gene*, 234:177–186, 1999.
- Simone M T Camps, Bas E Dutilh, Maiken C Arendrup, Antonius J M M Rijs, Eveline Snelders, Martijn A Huynen, Paul E Verweij, and Willem J G Melchers. Discovery of a hapE Mutation That Causes Azole Resistance in *Aspergillus fumigatus* through Whole Genome Sequencing and Sexual Crossing. *PLoS One*, 7(11), 2012. doi: 10.1371/journal.pone.0050034.
- Emilia Cantón, Javier Pemán, Macrina Sastre, Mónica Romero, and Ana Espinel-Ingroff. Killing kinetics of caspofungin, micafungin, and amphotericin B against *Candida guilliermondii*. *Antimicrobial Agents and Chemotherapy*, 50(8):2829–2832, 2006. ISSN 00664804. doi: 10.1128/AAC.00524-06.
- Mar Masiá Canuto and Félix Gutiérrez Rodero. Antifungal drug resistance to azoles and polyenes. *Lancet Infectious Diseases*, 2(9):550–563, 2002. ISSN 14733099. doi: 10.1016/S1473-3099(02)00371-7.
- M Castanheira, LN Woosley, DJ Diekema, SA Messer, RN Jones, and MA Pfaller. Low prevalence of FKS1 hot spot 1 mutations in a worldwide collection of *Candida* strains. *Antimicrobial Agents and Chemotherapy*, 54(6):2655–9, 2010.
- Sandro Centonze, Roberto Luzzati, Silvia Cavinato, Manuela Giangreco, Gianluca Gran, Maria L Deiana, and et al. Biolo. Peripheral and total parenteral nutrition as the strongest risk factors for nosocomial candidemia in elderly patients: a matched case – control study. *Mycoses*, pages 664–671, 2013. doi: 10.1111/myc.12090.

- Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, and Douglas M Ruden. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program , SnpSift. *Front Genet*, 3(March):1–9, 2012a. doi: 10.3389/fgene.2012.00035.
- Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, and et al. Lu. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Landes Bioscience*, 6(2):80–92, 2012b. doi: 10.4161/fly.19695.
- Karen Clark, Ilene Karsch-mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Research*, 44:67–72, 2016. doi: 10.1093/nar/gkv1276.
- Paul Cliften. *Base Calling, Read Mapping, and Coverage Analysis*. Elsevier Inc., 2014. ISBN 9780124051737. doi: 10.1016/B978-0-12-404748-8.00007-1. URL <http://dx.doi.org/10.1016/B978-0-12-404748-8.00007-1>.
- CLSI. Reference Method for Broth Dilution Antifungal Susceptibility Testing of Yeasts; Fourth Informational Supplement. CLSI document M27-S4. Wayne, PA: Clinical and Laboratory Standards Institute. *Clinical and Laboratory Standards Institute*, 2012.
- F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004. ISSN 00280836. doi: 10.1038/nature03001.
- Ercole Concia, Anna Maria Azzini, and Michela Conti. Epidemiology , Incidence and Risk Factors for Invasive Candidiasis in High-Risk Patients. *Drugs*, pages 5–14, 2009.
- S. Costa-De-Oliveira, C. Pina-Vaz, D. Mendonça, and A. Gonçalves Rodrigues. A first Portuguese epidemiological survey of fungaemia in a university hospital. *European Journal of Clinical Microbiology and Infectious Diseases*, 27(5):365–74, 2008.
- S. Costa-De-Oliveira, I. Marcos Miranda, RM. Silva, A. Pinto E Silva, R. Rocha, A. Amorim, A. Gonçalves Rodrigues, and C. Pina-Vaz. FKS2 mutations associated with decreased echinocandin susceptibility of *Candida glabrata* following anidulafungin therapy. *Antimicrobial Agents and Chemotherapy*, 55(3):1312–4, 2011.
- Alix Coste, Anna Selmecki, Anja Forche, Dorothee Diogo, Marie Elisabeth Bougnoux, Christophe D'Enfert, and et al. Berman. Genotypic evolution of azole resistance mechanisms in sequential *Candida albicans* isolates. *Eukaryotic Cell*, 6(10):1889–1904, 2007. ISSN 15359778. doi: 10.1128/EC.00151-07.
- Murray P Cox, Daniel A Peterson, and Patrick J Biggs. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11(1):

- 485, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-485. URL <https://doi.org/10.1186/1471-2105-11-485>.
- Maryam Dadar, Ruchi Tiwari, Kumaragurubaran Karthik, Sandip Chakraborty, Youcef Shalhali, and Kuldeep Dhama. Microbial Pathogenesis *Candida albicans* - Biology , molecular characterization , pathogenicity , and advances in diagnosis and control – An update. *Microbial Pathogenesis*, 117(February):128–138, 2018. ISSN 0882-4010. doi: 10.1016/j.micpath.2018.02.028. URL <https://doi.org/10.1016/j.micpath.2018.02.028>.
- Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, and et al. Handsaker. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011. doi: 10.1093/bioinformatics/btr330.
- Karen C Dannemiller, Darryl Reeves, Kyle Bibby, Naomichi Yamamoto, and Jordan Pecchia. Fungal High-throughput Taxonomic Identification tool for use with Next-Generation Sequencing (FHiTINGS). *J Basic Microbiol.*, pages 315–321, 2014. doi: 10.1002/jobm.201200507.
- Jigar V Desai, Vincent M Bruno, Shantanu Ganguly, Ronald J Stamper, Kaitlin F Mitchell, Norma Solis, Elizabeth M Hill, Wenjie Xu, Scott G Filler, David R Andes, Saranna Fanning, Frederick Lanni, and P Mitchell. Regulatory Role of Glycerol in *Candida albicans* Biofilm Formation. *MBio*, 4(10):1–10, 2013. ISSN 21612129. doi: 10.1128/mBio.00637-12. Invited.
- Ruud H Deurenberg, Erik Bathoorn, Monika A Chlebowicz, Natacha Couto, Mithila Ferdous, and et al. García-cobos. Application of Next Generation Sequencing in clinical microbiology and infection prevention. *Journal of Biotechnology*, 243:16–24, 2017. ISSN 0168-1656. doi: 10.1016/j.jbiotec.2016.12.022. URL <http://dx.doi.org/10.1016/j.jbiotec.2016.12.022>.
- Paul D Donovan, Gabriel Gonzalez, Desmond G Higgins, Geraldine Butler, and Kimihito Ito. Identification of fungi in shotgun metagenomics datasets. *PLOS One*, pages 1–16, 2018.
- Christoph Endrullat, Jörn Glökler, Philipp Franke, and Marcus Frohme. Standardization and quality management in Next-Generation Sequencing. *Applied & Translational Genomics*, 10:2–9, 2016. ISSN 2212-0661. doi: 10.1016/j.atg.2016.06.001. URL <http://dx.doi.org/10.1016/j.atg.2016.06.001>.
- D A Enoch, H A Ludlam, and N M Brown. Invasive fungal infections : a review of epidemiology and management options. *Journal of Medical Microbiology*, 55:809–818, 2006. doi: 10.1099/jmm.0.46548-0.

- Helga Erlendsdo, Gunnsteinn Haraldsson, Hong Guo, and Jianping Xu. Molecular Epidemiology of Candidemia : Evidence of Clusters of Smoldering Nosocomial Infections. *Clin Infect Dis*, 47:17–24, 1999. doi: 10.1086/589298.
- A. Espinel-Ingroff. Mechanisms of resistance to antifungal agents: yeasts and filamentous fungi. *Rev Iberoam Micol*, 25(2):101–6, 2008.
- Cristian Del Fabbro, Simone Scalabrin, Michele Morgante, and Federico M Giorgi. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS One.*, 8(12):1–13, 2013. doi: 10.1371/journal.pone.0085024.
- Saranna Fanning and Aaron P. Mitchell. Fungal biofilms. *PLoS pathogens*, 8(4):1–4, 2012. ISSN 15537374. doi: 10.1371/journal.ppat.1002585.
- Walter Florio, Arianna Tavanti, Emilia Ghelardi, and Antonella Lupetti. MALDI-TOF MS Applications to the Detection of Antifungal Resistance: State of the Art and Future Perspectives. *Frontiers in Microbiology*, 9(October):1–7, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.02577. URL <https://www.frontiersin.org/article/10.3389/fmicb.2018.02577/full>.
- Scott K Fridkin. Epidemiology of Nosocomial Fungal Infections. *Clin Microbiol Rev.*, 9(4): 499–511, 1996.
- Deborah D Garbee and Jennifer M Manning. Opportunistic Fungal Infections in Critical Care Units. *Crit Care Nurs Clin North Am.*, 2017. doi: 10.1016/j.cnc.2016.09.011.
- Gary Garber. An Overview of Fungi Infections. *Drugs*, 61:1–12, 2001.
- Guillermo Garcia-Effron, Santosh K. Katiyar, Steven Park, Thomas D. Edlind, and David S. Perlin. A naturally occurring proline-to-alanine amino acid change in Fks1p in *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* accounts for reduced echinocandin susceptibility. *Antimicrobial Agents and Chemotherapy*, 52(7):2305–2312, 2008. ISSN 00664804. doi: 10.1128/AAC.00262-08.
- Guillermo Garcia-Effron, Samuel Lee, Steven Park, John D. Cleary, and David S. Perlin. Effect of *Candida glabrata* FKS1 and FKS2 mutations on echinocandin sensitivity and kinetics of 1,3- β -D-glucan synthase: Implication for the existing susceptibility breakpoint. *Antimicrobial Agents and Chemotherapy*, 53(9):3690–3699, 2009. ISSN 00664804. doi: 10.1128/AAC.00443-09.
- Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: Ten years of Next-Generation Sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. ISSN 14710064. doi: 10.1038/nrg.2016.49. URL <http://dx.doi.org/10.1038/nrg.2016.49>.

- Nina T. Grossman, Tom M. Chiller, and Shawn R. Lockhart. Epidemiology of Echinocandin Resistance in *Candida*. *Current Fungal Infection Reports*, 8(4):243–248, 2014. ISSN 1936377X. doi: 10.1007/s12281-014-0209-7.
- Hyun S Gweon, Anna Oliver, Joanne Taylor, Tim Booth, Melanie Gibbs, Daniel S Read, Robert I Griffiths, and Karsten Schonrogge. PIPITS : an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods in Ecology and Evolution*, pages 973–980, 2015. doi: 10.1111/2041-210X.12399.
- Harvard Chan Bioinformatics Core. Understanding the Illumina sequencing technology, 2019. URL https://hbctraining.github.io/Intro-to-ChIPseq/lessons/02_QC_FASTQC.html.
- Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, and et al. Van Nieuwerburgh. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2):61–77, 2014. ISSN 07366205. doi: 10.2144/000114133.
- P Hui. Next Generation Sequencing: Chemistry, Technology and Applications. *Top Curr Chem*, 336:1–18, 2014. ISSN 23222271. doi: 10.1007/128.
- Zeina A Kanafani and John R Perfect. Resistance to antifungal agents: Mechanisms and clinical impact. *Chinese Journal of Infection and Chemotherapy*, 10(4):320, 2010. ISSN 10097708. doi: 10.1086/524071.
- Muthu K. Kathiravan, Amol B. Salake, Aparna S. Chothe, Prashik B. Dudhe, Rahul P. Watode, Maheshwar S. Mukta, and Sandeep Gadhwhe. The biology and chemistry of antifungal agents: A review. *Bioorganic and Medicinal Chemistry*, 20(19):5678–5698, 2012. ISSN 09680896. doi: 10.1016/j.bmc.2012.04.045. URL <http://dx.doi.org/10.1016/j.bmc.2012.04.045>.
- Santosh K. Katiyar, Ana Alastruey-Izquierdo, Kelley R. Healey, Michael E. Johnson, David S. Perlin, and Thomas D. Edlind. Fks1 and Fks2 are functionally redundant but differentially regulated in *Candida glabrata*: implications for echinocandin resistance. *Antimicrobial Agents and Chemotherapy*, 56(12):6304–6309, 2012. ISSN 00664804. doi: 10.1128/AAC.00813-12.
- SK Katiyar and TD Edlind. Identification and expression of multidrug resistance-related ABC transporter genes in *Candida krusei*. *Med Mycol*, 39(1):109–16, 2001.
- Claudio U Ko, Matthew J Ellington, and Sharon J Peacock. Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.*, 30(9):401–407, 2014. doi: 10.1016/j.tig.2014.07.003.

- Julia R Ko, Arturo Casadevall, and John Perfect. The Spectrum of Fungi That Infects Humans. *Cold Spring Harb Perspect Med.*, pages 1–22, 2015. doi: 10.1101/cshperspect.a019273.
- DP Kontoyiannis, KA Marr, BJ Park, BD Alexander, EJ Anaissie, and TJ Walsh. Prospective surveillance for invasive fungal infections in hematopoietic stem cell transplant recipients, 2001–2006: overview of the Transplant-Associated Infection Surveillance Network (TRANSNET) Database. *Clin Infect Dis*, 50(8):1091–1100, 2010.
- Frederic Lamoth, Shawn R. Lockhart, Elizabeth L. Berkow, and Thierry Calandra. Changes in the epidemiological landscape of invasive candidiasis. *Journal of Antimicrobial Chemotherapy*, 73:i4–i13, 2018. ISSN 14602091. doi: 10.1093/jac/dkx444.
- Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3), 2009. doi: 10.1186/gb-2009-10-3-r25.
- Christian Ledergerber and Christophe Dessimoz. Base-calling for Next-Generation Sequencing platforms. *Briefings in Bioinformatics*, 12(5):489–497, 2011. ISSN 14675463. doi: 10.1093/bib/bbq077.
- Martina I Lefterova, Carlos J Suarez, Niaz Banaei, and Benjamin A Pinsky. Next-Generation Sequencing for Infectious Disease Diagnosis and Management A Report of the Association for Molecular Pathology. *The Journal of Molecular Diagnostics*, 17(6):623–634, 2015. ISSN 1525-1578. doi: 10.1016/j.jmoldx.2015.07.004. URL <http://dx.doi.org/10.1016/j.jmoldx.2015.07.004>.
- Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1):2010–2012, 2011. ISSN 03051048. doi: 10.1093/nar/gkq1019.
- Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows – Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi: 10.1093/bioinformatics/btp324.
- Heng Li and Nils Homer. A survey of sequence alignment algorithms for Next-Generation Sequencing. *Brief Bioinform*, 11(5), 2010. doi: 10.1093/bib/bbq015.
- Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, pages 1851–1858, 2008a. doi: 10.1101/gr.078212.108.
- Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18:1851–1858, 2008b. doi: 10.1101/gr.078212.108.

- R Li, Y Li, X Fang, H Yang, and K Kristiansen. SNP detection for massively parallel whole-genome resequencing. *Genome research*, 19(6):1124–32, 2009a.
- Ruiqiang Li, Chang Yu, Yingrui Li, Tak Wah Lam, Siu Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009b. ISSN 13674803. doi: 10.1093/bioinformatics/btp336.
- T Lion. *Human Fungal Pathogen Identification*. Springer Protocols, 2017. ISBN 9781493965137.
- Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of Next-Generation Sequencing systems. *The Role of Bioinformatics in Agriculture*, 2012:1–25, 2014. ISSN 11107243. doi: 10.1201/b16568.
- Frank Luh and Yun Yen. FDA guidance for Next Generation Sequencing-based testing : balancing regulation and innovation in precision medicine. *Genomic Medicine*, 3(28):2–4, 2018. ISSN 2056-7944. doi: 10.1038/s41525-018-0067-2. URL <http://dx.doi.org/10.1038/s41525-018-0067-2>.
- S. MacPherson, B. Akache, S. Weber, X. Deken, M. Raymond, and B. Turcotte. *Candida albicans* Zinc Cluster Protein Upc2p Confers Resistance to Antifungal Drugs and Is an Activator of Ergosterol Biosynthetic Genes. *Mutagenesis*, 49(5):1745–1752, 2005. ISSN 1064-3745. doi: 10.1128/AAC.49.5.1745.
- Alberto Magi, Matteo Benelli, Alessia Gozzini, Francesca Girolami, Francesca Torricelli, and Maria Luisa Brandi. Bioinformatics for Next Generation Sequencing data. *Genes*, 1(2):294–307, 2010. ISSN 20734425. doi: 10.3390/genes1020294.
- Raman Manoharlal, Naseem Akhtar Gaur, Sneha Lata Panwar, Joachim Morschhäuser, and Rajendra Prasad. Transcriptional activation and increased mRNA stability contribute to overexpression of CDR1 in azole-resistant *Candida albicans*. *Antimicrobial Agents and Chemotherapy*, 52(4):1481–1492, 2008. ISSN 00664804. doi: 10.1128/AAC.01106-07.
- Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, and et al. Bader. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005. ISSN 00280836. doi: 10.1038/nature03959.
- P Marichal, J Gorrens, MC Coene, L Le Jeune, and H Vanden Bossche. Origin of differences in susceptibility of *Candida krusei* to azole antifungal agents. *Mycoses*, 38(3):111–7, 1995.
- Patrick Marichal, Luc Koymans, Staf Willemsens, Danny Bellens, Peter Verhasselt, Walter Luyten, Marcel Borgers, Frans C.S. Ramaekers, Frank C. Odds, and Hugo Vanden Bossche. Contribution of mutations in the cytochrome P450 14 α -demethylase (Erg11p, Cyp51p) to azole resistance in *Candida albicans*. *Microbiology*, 145(10):2701–2713, 1999. ISSN 13500872. doi: 10.1099/00221287-145-10-2701.

- Gabor T Marth, I Korf, MD Yandell, RT Yeh, G Zhijie, H Zakeri, NO Stitzel, L Hillier, P Kwok, and WR Gish. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, 23:452–456, 1999.
- Walter Maxam and Allan M. Gilbert. A new method for sequencing DNA. *Pnas*, 74(2):99–103, 1977. ISSN 0740-7378. doi: 10.1073/pnas.74.2.560. URL <http://www.ncbi.nlm.nih.gov/pubmed/1422074>.
- François L Mayer, Duncan Wilson, and Bernhard Hube. *Candida albicans* pathogenicity mechanisms. *Virulence*, 4(2):119–28, 2013. ISSN 2150-5608. doi: 10.4161/viru.22913. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3654610&tool=pmcentrez&rendertype=abstract>.
- Daniel Mcdonald, Jose C Clemente, Justin Kuczynski, Jai Ram Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J Gregory Caporaso. The Biological Observation Matrix (BIOM) format or : how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(7):1–6, 2012.
- Aaron Mckenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytzsky, and et al. Garimella. The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, pages 1297–1303, 2010. doi: 10.1101/gr.107524.110.20.
- William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070, 2010. doi: 10.1093/bioinformatics/btq330.
- Marisa H Miceli, José A Díaz, Samuel A Lee, and Non-albicans Candida. Emerging opportunistic yeast infections Emerging yeasts. *The Lancet Infectious Diseases*, 11(2):142–151, 2011. ISSN 1473-3099. doi: 10.1016/S1473-3099(10)70218-8. URL [http://dx.doi.org/10.1016/S1473-3099\(10\)70218-8](http://dx.doi.org/10.1016/S1473-3099(10)70218-8).
- Olena Morozova and Marco A. Marra. Applications of Next-Generation Sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, 2008. ISSN 08887543. doi: 10.1016/j.ygeno.2008.07.001. URL <http://dx.doi.org/10.1016/j.ygeno.2008.07.001>.
- CA Munro, S Selvaggini, I de Bruijn, L Walker, MD Lenardon, B Gerssen, S Milne, AJ Brown, and NAR Gow. The PKC, HOG and Ca²⁺ signalling pathways co-ordinately regulate chitin synthesis in *Candida albicans*. *Mol Microbiol*, 63:1399–1413, 2007.
- Hannah Muskett, Jason Shahin, Gavin Eyres, Sheila Harvey, Kathy Rowan, and David Harrison. Risk factors for invasive fungal disease in critically ill adult patients : a systematic review. *Crit Care*, 2011.

- Anthony Nash, Thomas Sewell, Rhys A. Farrer, Alireza Abdolrasouli, Jennifer M.G. Shelton, Matthew C. Fisher, and Johanna Rhodes. MARDy: Mycology Antifungal Resistance Database. *Bioinformatics*, 34(18):3233–3234, 2018. ISSN 13674811. doi: 10.1093/bioinformatics/bty321.
- Melyssa Negri, Mariana Henriques, David W Williams, and Joana Azeredo. *Candida glabrata*, *Candida parapsilosis*, *Candida tropicalis*: biology, epidemiology, pathogenicity and antifungal resistance. *FEMS Microbiology Reviews*, 36, 2012. doi: 10.1111/j.1574-6976.2011.00278.x.
- Patrícia Oliveira. Development of a framework for identification of *Candida* species and detection of antifungal resistance – Source Code, 2019. URL <https://github.com/paty-oliveira/dissertation>.
- T D Otto, E A Vasconcellos, L H F Gomes, and A S Moreira. ChromaPipe : a pipeline for analysis , quality control and management for a DNA sequencing facility. *Genet Mol Res*, 7(3):861–871, 2008.
- L Pagano, M Caira, A Candoni, and Et al. The Epidemiology Of Fungal Infections In Patients With Hematologic Malignancies: The SEIFEM-2004 Study. *Haematologica*, 91: 1068–75, 2006.
- Elisabeth Paramythiotou, Frantzeska Frantzeskaki, Aikaterini Flevari, Apostolos Armanidis, and George Dimopoulos. Invasive fungal infections in the ICU: How to approach, how to treat. *Molecules*, 19(1):1085–1119, 2014. ISSN 14203049. doi: 10.3390/molecules19011085.
- Ravi K Patel and Mukesh Jain. NGS QC Toolkit : A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, 7(2):e30619, 2012. doi: 10.1371/journal.pone.0030619.
- Javier Pemán, Emilia Cantón, and Ana Espinel-Ingroff. Antifungal drug resistance mechanisms. *Expert Review of Anti-Infective Therapy*, 7(4):453–460, 2009. ISSN 14787210. doi: 10.1586/ERI.09.18.
- Sofia Perea and Thomas F. Patterson. Antifungal Resistance in Pathogenic Fungi. *Clinical Infectious Diseases*, 35(9):1073–1080, 2002. ISSN 1058-4838. doi: 10.1086/344058. URL <https://academic.oup.com/cid/article-lookup/doi/10.1086/344058>.
- David S. Perlin. Echinocandin Resistance in *Candida*. *Clinical Infectious Diseases*, 61(Suppl 6): S612–S617, 2015. ISSN 15376591. doi: 10.1093/cid/civ791.
- M Pesti, M Sipiczki, and Y Pinter. Scanning electron microscopy characterisation of colonies of *Candida albicans* morphological mutants. *J Med Microbiol.*, 48:167–172, 1999.

- M. A. Pfaller, S. A. Messer, L. Boyken, S. Tendolkar, R. J. Hollis, and D. J. Diekema. Geographic variation in the susceptibilities of invasive isolates of *Candida glabrata* to seven systemically active antifungal agents: A global assessment from the ARTEMIS antifungal surveillance program conducted in 2001 and 2002. *Journal of Clinical Microbiology*, 42(7): 3142–3146, 2004. ISSN 00951137. doi: 10.1128/JCM.42.7.3142-3146.2004.
- M a Pfaller, D J Diekema, and D J Sheehan. Interpretive Breakpoints for Fluconazole and *Candida* a blueprint for the future of antifungal susceptibility. *Clinical Microbiology Reviews*, 19(2):435–447, 2006. ISSN 0893-8512. doi: 10.1128/CMR.19.2.435.
- MA Pfaller and DJ Diekema. Epidemiology of invasive candidiasis: a persistent public health problem. *Clinical Microbiology Reviews*, 20(1):133–163, 2007.
- MA Pfaller, SA Messer, and RJ Hollis. Strain delineation and antifungal susceptibilities of epidemiologically related and unrelated isolates of *Candida lusitanae*. *Diagnostic Microbiology and Infectious Disease*, 20(3):127–133, 1994.
- Michael A Pfaller. Nosocomial Candidiasis : Emerging Species , Reservoirs , and Modes of Transmission. *Clin Infect Dis.*, 22(suppl 2):89–94, 1996.
- C. Pina-Vaz, F. Sansonetty, AG. Rodrigues, S. Costa-De-Oliveira, C. Tavares, and J. Martinez-de Oliveira. Cytometric approach for a rapid evaluation of susceptibility of *Candida* strains to antifungals. *Clin Microbiol Infect*, 7(11):609–18, 2001.
- C. Pina-Vaz, S. Costa-De-Oliveira, AG. Rodrigues, and A. Espinel-Ingroff. Comparison of two probes for testing susceptibilities of pathogenic yeasts to voriconazole, itraconazole, and caspofungin by flow cytometry. *J Clin Microbiol*, 43(9):4674–9, 2005.
- Gordon Ramage, Stephen P Saville, and Derek P Thomas. *Candida* biofilms: an update. *American Society for Microbiology*, 4(4):633–638, 2005. ISSN 1535-9778 (Print). doi: 10.1128/EC.4.4.633.
- R. Ramani and V. Chaturvedi. Flow cytometry antifungal susceptibility testing of pathogenic yeasts other than *Candida albicans* and comparison with the NCCLS broth microdilution test. *Antimicrobial Agents and Chemotherapy*, 44(10):2752–2758, 2000. ISSN 00664804. doi: 10.1128/AAC.44.10.2752-2758.2000.
- David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, and et al. Chaudhuri. Real-time DNA sequencing from single polymerase molecules. *Science*, 323 (5910):133–138, 2009. ISSN 00368075. doi: 10.1126/science.1162986.
- John H Rex, Michael A Pfaller, Thomas J Walsh, Vishnu Chaturvedi, A N A Espinel-ingroff, and et al. Ghannoum. Antifungal Susceptibility Testing : Practical Aspects and Current Challenges. *Clinical Microbiology Reviews*, 14(4):643–658, 2001. doi: 10.1128/CMR.14.4.643.

- Malcolm D Richardson. Changing patterns and trends in systemic fungal infections. *J Antimicrob Chemother*, pages 5–11, 2005. doi: 10.1093/jac/dki218.
- K Rotmistrovsky and R Agarwala. Bmtagger: Best match tagger for removing human reads from metagenomics datasets, 2011.
- F. Sanger, S. Nicklen, and AR Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, 74(12):5463–5467, 1977. ISSN 09237534. doi: 10.1093/annonc/mdp039.
- Dominique Sanglard and Frank C. Odds. Resistance of *Candida* species to antifungal agents: Molecular mechanisms and clinical consequences. *Lancet Infectious Diseases*, 2(2):73–85, 2002. ISSN 14733099. doi: 10.1016/S1473-3099(02)00181-0.
- Dominique Sanglard, Leah E Cowen, Dominique Sanglard, Susan J Howard, P David Rogers, and David S Perlin. Mechanisms of Antifungal Drug Resistance . Mechanisms of Antifungal Drug Resistance. *Cold Spring Harb Perspect Med*, 5(July 2015), 2014. ISSN 2157-1422. doi: 10.1101/cshperspect.a019752.
- Maurizio Sanguinetti and Brunella Posteraro. Susceptibility Testing of Fungi to Antifungal Drugs. *Journal of Fungi*, 4(3):110, 2018. ISSN 2309-608X. doi: 10.3390/jof4030110. URL <http://www.mdpi.com/2309-608X/4/3/110>.
- J. C O Sardi, L. Scorzoni, T. Bernardi, A. M. Fusco-Almeida, and M. J S Mendes Giannini. *Candida* species: Current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *Journal of Medical Microbiology*, 62:10–24, 2013. ISSN 00222615. doi: 10.1099/jmm.0.045054-0.
- Natacha Sertour, Marie-elisabeth Bougnoux, Eric Dannaoui, Sylvie Larrat, Christophe Hennequin, and Muriel Cornet. Next-generation sequencing offers new insights into the resistance of *Candida* spp. to echinocandins and azoles. *J Antimicrob Chemother*, 70(9): 2556–2565, 2015. doi: 10.1093/jac/dkv139.
- Sónia Silva, Melyssa Negri, Mariana Henriques, Rosário Oliveira, David W. Williams, and Joana Azeredo. *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*: Biology, epidemiology, pathogenicity and antifungal resistance. *FEMS Microbiology Reviews*, 36(2): 288–305, 2012. ISSN 01686445. doi: 10.1111/j.1574-6976.2011.00278.x.
- DF Simola and J Kim. Sniper: improved SNP discovery by multiply mapping deep sequenced reads. *Genome Biology and Evolution*, 12(6):R55, 2011.
- You Bum Song, Moo Kyu Suh, Gyoung Yim Ha, and Heesoo Kim. Antifungal susceptibility testing with etest for *Candida* species isolated from patients with oral candidiasis. *Annals of Dermatology*, 27(6):715–720, 2015. ISSN 20053894. doi: 10.5021/ad.2015.27.6.715.

- Claude Thermes. Ten years of Next-Generation Sequencing technology. *Trends in genetics : TIG*, 30(9):418–426, 2014. ISSN 01689525. doi: 10.1016/j.tig.2014.07.001.
- Athanasios Tragiannidis, Christos Tsoulas, Kornelius Kerl, and Andreas H Groll. Invasive candidiasis : update on current pharmacotherapy options and future perspectives. *Expert Opin Pharmacother.*, pages 1–14, 2013.
- Urmi H Trivedi, Timothée Cézard, Stephen Bridgett, Anna Montazam, Jenna Nichols, Mark Blaxter, Karim Gharbi, Simon Andrews, and The Babraham. Quality control of Next-Generation Sequencing data without a reference. *Front Genet.*, 5(May):1–13, 2014. doi: 10.3389/fgene.2014.00111.
- V M Urban, S Silva, L N Dovigo, J H Jorge, and N H Campanha. Identification of *Candida* species in the clinical laboratory : a review of conventional , commercial , and molecular techniques. *Oral Diseases*, 30:329–344, 2014. doi: 10.1111/odi.12123.
- A Valouev, J Ichikawa, T Tonthat, J Stuart, S Ranade, H Peckham, K Zeng, JA Malek, and et al. Costa. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18:1051–1063, 2008. ISSN 10889051. doi: 10.1101/gr.076463.108.3.
- J Vincent, J Rello, J Marshall, and et Al. International study of the prevalence and outcomes of infection in intensive care units. *JAMA*, 302(21):2323–2329, dec 2009. ISSN 0098-7484. URL <http://dx.doi.org/10.1001/jama.2009.1754>.
- Karl V. Voelkerding, Shale A. Dames, and Jacob D. Durtschi. Next-Generation Sequencing:from basic research to diagnostics. *Clinical Chemistry*, 55(4):641–658, 2009. ISSN 00099147. doi: 10.1373/clinchem.2008.112789.
- LA Walker, DM Maccallum, G Bertram, NA Gow, FC Odds, and AJ Brown. Genome-wide analysis of *Candida albicans* gene expression patterns during infection of the mammalian kidney. *Fungal Genet Biol*, 46(2):210–9, 2009.
- Louise A. Walker, Neil A R Gow, and Carol A. Munro. Elevated chitin content reduces the susceptibility of *Candida* species to caspofungin. *Antimicrobial Agents and Chemotherapy*, 57(1):146–154, 2013. ISSN 00664804. doi: 10.1128/AAC.01486-12.
- Louise A. Walker, Keunsook K. Lee, Carol A. Munro, and Neil A R Gow. Caspofungin treatment of *Aspergillus fumigatus* results in ChsG-dependent upregulation of chitin synthesis and the formation of chitin-rich microcolonies. *Antimicrobial Agents and Chemotherapy*, 59(10):5932–5941, 2015. ISSN 10986596. doi: 10.1128/AAC.00862-15.

- Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):1–7, 2010. doi: 10.1093/nar/gkq603.
- Qiong Wang, George M Garrity, James M Tiedje, James R Cole, and Wang E T Al. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007. doi: 10.1128/AEM.00062-07.
- Martin Welker. Proteomics for routine identification of microorganisms. *Proteomics*, 11(15): 3143–3153, 2011. ISSN 16159853. doi: 10.1002/pmic.201100049.
- Sarah G. Whaley, Elizabeth L. Berkow, Jeffrey M. Rybak, Andrew T. Nishimoto, Katherine S. Barker, and P. David Rogers. Azole antifungal resistance in *Candida albicans* and emerging non-*albicans* *Candida* Species. *Frontiers in Microbiology*, 7(JAN):1–12, 2017. ISSN 1664302X. doi: 10.3389/fmicb.2016.02173.
- James Robert White, Cynthia Maddox, Owen White, Samuel V Angiuoli, and W Florian Fricke. CloVR-ITS : Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. *BMC Bioinformatics*, pages 1–11, 2013.
- Nathan P. Wiederhold. Echinocandin Resistance in *Candida* Species: a Review of Recent Developments. *Current Infectious Disease Reports*, 18(12), 2016. ISSN 15343146. doi: 10.1007/s11908-016-0549-2. URL <http://dx.doi.org/10.1007/s11908-016-0549-2>.
- Nathan P. Wiederhold. Antifungal resistance: current trends and future strategies to combat. *Infection and Drug Resistance*, 10:249–259, 2017. ISSN 11786973. doi: 10.2147/IDR.S124918.
- N Yapar. Epidemiology and risk factors for invasive candidiasis. *Ther Clin Risk Manag*, 10: 95–105, 2014.
- Hongen Zhang. Overview of Sequence Data Formats. In Ewy Mathé and Sean Davis, editors, *Statistical Genomics: Methods and Protocols*, volume 1418, chapter 1, pages 3–17. Methods Mol Biol, New York, springer s edition, 2016. ISBN 9781493935789. doi: 10.1007/978-1-4939-3578-9.
- Qian Zhou, Xiaoquan Su, Anhui Wang, Jian Xu, and Kang Ning. QC-Chain : Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS One*, 8(4), 2013. doi: 10.1371/journal.pone.0060234.
- Jan Zoll, Eveline Snelders, Paul E Verweij, and Willem J G Melchers. Next-Generation Sequencing in the Mycology Lab. *Current Fungal Infection Reports*, pages 37–42, 2016.

ISSN 1936-3761. doi: 10.1007/s12281-016-0253-6. URL <http://dx.doi.org/10.1007/s12281-016-0253-6>.



SEQUENCING DATA FORMATS

SAM/BAM

SAM/BAM (Sequence Alignment Map/Binary Alignment Map) files have an important role in NGS due to their structure, which allows the data representation obtained from the mapping/alignment of sequence reads. SAM is a file that stores the alignments against reference sequences and has the file extension “.sam”. BAM is the binary version of a SAM file and has the file extension “.bam” (Zhang, 2016). In Figure 19, an example of a SAM file (Zhang, 2016) is provided.

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:47
ref 516 ref 1 0 14M2D31M * 0 0 AGCATGTTAGATAAGATAGCTGTGCTAGTAGGCAGTCAGCGCCAT *
r001 99 ref 7 30 14M1D3M = 39 41 TTAGATAAAGGATACTG *
* 768 ref 8 30 1M * 0 0 * * CT:Z:.;Warning;Note=Ref wrong?
r002 0 ref 9 30 3S6M1D5M * 0 0 AAAAGATAAGGATA * PT:Z:1;4;+;homopolymer
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 18 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 31 30 6H5M * 0 0 TAGGC * NM:i:0
r001 147 ref 39 30 9M = 7 -41 CAGCGGCAT * NM:i:1
```

Figure 19.: Example of SAM file format with the header and alignment sections. Adapted from Zhang (2016).

BAM/SAM files can be divided in two sections: header and alignment. The header section is optional and if present, can have multiple lines and each line must start with a “@” symbol. The information can be divided into four types: @HD, @SQ, @RG and @PG. An optional comment line starting with @CO may exist at the end of the header (Zhang, 2016).

If the header section exists in a BAM/SAM file, the first line must start with @HD followed by required and optional field(s) including:

- VN: format version. This field is required.
- SO: sorting order of alignments. Valid values are: unknown, unsorted, queryname, and coordinate.

- GO: group order. Valid values are: none, query, or reference.

@SQ line serves as reference sequence dictionary. A BAM/SAM file has multiple @SQ lines and the order of these lines defines the alignment sorting order. @SQ line has following fields:

- SN: reference sequence name.
- LN: reference sequence length. This is a required field.
- AS: genome assembly identifier.
- M5: checksum of the sequence in the uppercase.
- SP: species.
- UR: URI of the sequence. This value may start with one of the standard protocols, e.g., http: or ftp:.

Read group information is listed in @RG line(s). An @RG line has also required and optional fields:

- ID: read group identifier. This is a required field.
- CN: name of sequencing center production the read.
- DS: description.
- DT: the data the run was produced, using ISO8601 data or date/time format.
- FO: flow order.
- KS: the array of nucleotide bases that correspond to the key sequence of each read.
- LB: library.
- PG: programs used for processing the read group.
- PI: predicted median insert size.
- PL: platform/technology used to produce the reads.
- PM: platform model.
- PU: platform unit. It must be a unique identifier.
- SM: samples name.

@PG line(s) describe the program used to generate BAM/SAM file with following fields:

- ID: program record identifier. Each @PG line must have a unique ID.
- PN: program name.
- CL: command line.
- PP: previous @PG-ID. It must match another @PG header's ID tag.
- DS: description of the program.
- VN: program version.

In the alignment section, the SAM/BAM file contains sequences with genomic position and other descriptive information. Each single sequence and its associated information are

presented as one-line text and each line consists of multiple tab-delimited text fields (Zhang, 2016). For each line, there are eleven required fields:

- QNAME: query name (sequence identifier in FASTQ file).
- FLAG: bitwise flag of two bytes length to indicate the read property (mapped or unmapped, passed or not passed in quality controls).
- RNAME: reference sequence name.
- POS: the leftmost position of the first matching base in reference sequence.
- MAPQ: mapping quality.
- CIGAR: character string indication the match status of bases in the short read.
- RNEXT: reference name of next aligned read.
- PNEXT: position of the primary alignment of the next read.
- TLEN: signed observed template length.
- SEQ: sequence of the short read.
- QUAL: quality scores for each base in the short read.

GFF/GTF

The GFF/GTF (General Feature Format/General Transfer Format) file formats are used as data formats to provide genomic annotation information for NGS data analysis. The GFF/GTF file extensions are “.gff” and “.gtf”, respectively. Each feature is represented with one-line text and each line has nine fields. Data fields must be tab-separated, and each field must contain a value. Empty fields should be denoted with a “.” (Zhang, 2016). The nine fields are:

- SEQNAME: the name of the sequence. Normally, it is the identifier of the sequence in the FASTA format file.
- SOURCE: name of the program that generated this feature.
- FEATURE: feature type name, e.g., gene, exon, transcript.
- START: start position of the feature, with sequence numbering starting at 1.
- END: end position of the feature, with sequence numbering starting at 1.
- SCORE: floating point value.
- STRAND: defined as “+” (forward) or “-” (reverse).
- FRAME: “0” indicates that the specified region is in frame, i.e., that its first base corresponds to the first base of a codon. “1” indicates that there is one extra base, i.e., that the second base of region corresponds to the first base of a codon. “2” indicates that the third base of the region is the first base of a codon.
- ATTRIBUTE: a semicolon-separated list of tag-value pairs, providing additional information about each feature.

GFF/GTF files may contain comment lines at the beginning which must start with "#". Figure 20 represents an example of the GFF/GTF file format (Zhang, 2016).

```

#!genome-build GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.15
#!genebuild-last-updated 2014-08
1 havana gene 11869 14409 . + . gene_id "ENSG00000223972"
1 havana exon 11869 14409 . + . gene_id "ENSG00000223972"
1 havana exon 11869 12227 . + . gene_id "ENSG00000223972"
1 havana exon 12613 12721 . + . gene_id "ENSG00000223972"

```

Figure 20.: Example of the GFF/GTF file format. Adapted from Zhang (2016).

BED

The BED (Browser Extensible Data) format is another file format to store the features of genome annotations. The BED format is a tab-delimited text file and each line represents one feature. The BED format has the file extension ".bed". A basic BED file has only three required fields in each line, i.e., *chrom* to define the chromosome name, *chromStart* and *chromEnd* to define the start and end position of the feature on the chromosome (Zhang, 2016). The optional fields are:

- NAME: defines the name of the BED line.
- SCORE: a score between 0 and 1000.
- STRAND: defined as "+" (forward) or "-" (reverse).
- THICKSTART: the starting position.
- THICKEND: the ending position.
- ITEMRGB: if "On", the RGB value will determine the display color of the data contained in the BED line.
- BLOCKCOUNT: the number of blocks (exons) in the BED line.
- BLOCKSIZES: a comma-separated list of the block sizes. The number of items in this list should correspond to BLOCKCOUNT.
- BLOCKSTARTS: a comma-separated list of blocks starts. All the BLOCKSTART positions should be calculated relative to CHROMSTART. The number of items in this list should correspond to BLOCKCOUNT.

A BED file can also contain track line definition to configure the display further, e.g., by grouping features into separate tracks. These lines should be placed at the beginning of the list of features that they are to affect. They start with the word "track" followed by space-

separated key=value pairs (Zhang, 2016). Figure 21 represents an example of the BED file format (Zhang, 2016) with a track line.

```
track name="ItemRGBDemo" description="Item RGB demonstration" itemRgb="On"
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
```

Figure 21.: Example of the BED file format with the track line. Adapted from Zhang (2016).

VCF

The VCF (Variant Call Format) format stores the most prevalent types of genomic sequences variation, such as SNPs and small INDELS (insertions/deletions), enriched by annotations. The VCF format has the file extension “.vcf”. A VCF file contains header and data lines. The header lines start with “##” and have a field in the format of “ID=value”. The first header line is always the VCF format version followed by lines starting with ##INFO=, ##FILTER=, and ##FORMAT, which define the name, length, value types, and description of each item of each data line. Data lines in VCF files are tab-delimited text lines and each line contains nine fixed fields followed by one or more sample column(s) (Zhang, 2016). The nine fields are:

- CHROM: chromosome name.
- POS: the leftmost position of the variant in the sequence.
- ID: variant identifier such as SNP id.
- REF: reference base for SNP or sequence for INDEL.
- ALT: alternate base or bases.
- QUAL: Phred-scaled quality score.
- FILTER: PASS for passed all filters or a semicolon-separated list of code for filters that fail.
- INFO: additional information.
- FORMAT: colon separated key and value.
- Sample(s): one or more sample columns may exist in a VCF file.

Figure 22 represents an example of the VCF file format (Zhang, 2016).

```
##fileformat=VCFv4.2
##fileDate=20151002
##source=callMomV0.2
##reference=gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
##contig=<ID=MT,length=16569,assembly=b37>
##INFO=<ID=VT,Number=.,Type=String,Description="Alternate allele type. S=SNP, M=MNP, I=Indel">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate allele counts, comma delimited when multiple">
##FILTER=<ID=fa,Description="Genotypes called from fasta file">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG000096 HG000097 HG000099
MT 10 . T C 100 fa VT=S;AC=3 GT 0 0 0
MT 16 . A T 100 fa VT=S;AC=3 GT 0 0 0
MT 26 . C T 100 fa VT=S;AC=3 GT 0 0 0
MT 35 . G A 100 fa VT=S;AC=2 GT 0 0 0
MT 40 . TC CT 100 fa VT=M;AC=1 GT 0 0 0
```

Figure 22.: Example of the VCF file format. Adapted from Zhang (2016).

