TESI DOCTORAL UPF / 2020

Bioinformatic approaches for integration and analysis of fungal omics data oriented to knowledge discovery and diagnosis

Ahmed Ibrahem Hafez Khafaga

Thesis supervisors

Dr./ Carlos Llorens
Biotechvana S.L.

Dr./ Toni Gabladón
Comparative Genomics Group,
Barcelona Supercomputing Centre (BSC-CNS)

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

BIOTECHVANA

upf. Universitat Pompeu Fabra Barcelona

Acknowledgements

"All praises and thanks to Allah, who supplied me with courage, guidance, and love to finish this journey"

I cannot forget the assistance of several individuals whose contributions are gratefully acknowledged, without their guidance, this thesis could not appear in its current format. I would like to thank my supervisors, Carlos and Toni, for their patience and invaluable advices and support.

I would like to thank all my friends and collages how had helped me along the way, however close or far they are.

Many Thanks to all OPATHians for a great time and a lot of fun, great collaborations, great project and so many birthday cakes. And Thanks Toni for this wonderful ITN.

And Many Thanks to Biotechvana family for their constant support and the lovely environment, and sure, thanks for all the coffee. Carlos, thank you for all your help and support along the way, and your great hospitality.

And at the beginning, I really apricate and thank Toni and Carlos for giving me the opportunity to be part of this great ITN.

Finally, I am so grateful to my family, my parents, and my lovely wife Aya for their love, support, and encouragement, and to my lovely sons; Hamza and Hazem, who make me smile and laugh.

Thanks, Aya for everything and now it's your turn.

"O My Lord! Advance me in knowledge", Qur'an [20:114]

This work is an industrial thesis in the framework of a collaboration between the academy and the industry, supported by the European Union’s Horizon 2020 Research and Innovation programme under the Marie Sklodowska-Curie grant agreement (No. 642095) of the OPATHY consortium.

Abstract

Infections caused by pathogenic yeasts are becoming increasingly prevalent, infecting billions of people every year. Nowadays, these infections are poorly understood, difficult to diagnose, and are becoming increasingly frequent and severe. Pressing problems are the emergence of resistance to antifungal drugs, the persistence in the host, and the lack of fast and efficient diagnostic tools to direct to the appropriate treatment. The aim of this thesis has been to develop a series of bioinformatic resources (tools and methodologies) to support the OPATHY consortium in the analysis of data provided by next generation sequencing, proteomics, or other omics technologies in the field of study and diagnosis of yeast infections (and by extension to researcher working on omics, in general). In particular, we have explored and designed distinct computational techniques to identify novel biomarker candidates of resistance traits, to predict DNA/RNA sequences' features, and to optimize sequencing strategies for host-pathogen transcriptome sequencing studies (Dual RNA-seq). We have also designed and developed an efficient bioinformatic solution composed of a server-side component constituted by distinct pipelines for VariantSeq, Denovoseq and RNAseq analyses as well as another component constituted by distinct GUI-based software to let the user to access, manage and run the pipelines using friendly-to-use and secure interfaces. We have also designed and developed SeqEditor a software for sequence analysis and primers design that can be used to design primer for species identification and detection in PCR diagnosis. Finally, we have developed CandidaMine an integrated data warehouse for fungal omics data and for data analysis and knowledge discovery.

Resum

Aquesta tesi presenta una sèrie de recursos bioinformàtics desenvolupats per a donar suport en l'anàlisi de dades de NGS i altres òmics en el camp d'estudi i diagnòstic d'infeccions fúngiques. Hem dissenyat tècniques de computació per identificar nous biomarcadors i determinar potencial trets de resistència, pronosticant les característiques de les seqüències d'ADN/ARN, i planejant estratègies optimitzades de seqüenciació per als estudis de hoste-patogen transcriptomes (Dual RNA-seq). Hem dissenyat i desenvolupat tambe una solució bioinformàtica composta per un component de costat de servidor (constituït per diferents pipelines per a fer anàlisi VariantSeq, Denovoseq i RNAseq) i un altre component constituït per eines software basades en interfícies gràfiques (GUIs de l'acrònim en anglès) per permetre a l'usuari accedir, gestionar i executar els pipelines mitjançant interfícies amistoses. També hem desenvolupat i validat un software per a l'anàlisi de seqüències i el disseny dels primers (SeqEditor) orientat a la identificació i detecció d'espècies en el diagnòstic de la PCR. Finalment, hem desenvolupat una base de dades integrant dades omiques de fongs patògens que hem anomenat CandidaMine.

Overview

This thesis was developed in the framework of the EU-funded international training network OPATHY (**O**mics of **PATH**ogenics **Y**easts www.opathy.org), an inter-disciplinary research network aiming to develop novel solutions involving high throughput technologies (omics) to study, treat, and diagnose yeast pathogens. The overall aim of the present thesis was to fill in existing gaps in exploiting these technologies. One of the central goals was to develop a bioinformatic infrastructure including an integrative database, scripting and pipelines tools for pipeline-workflow automatization, friendly-to-use interfaces for efficient processing, analysis, and storage of omics fungal data. This includes pipeline tools for data analysis and knowledge discovery based on the data interrogation through the development of state of the art computational tools for automatization of primer design, analysis of collected data as well as techniques for identification of novel biomarker candidates that could be utilized for diagnostic purposes.

This thesis has been divided into ten chapters, which I briefly introduce here:

- **Chapter 1 (“*Introduction”*)** presents an overall overview to different perspectives and recent trends in yeast infections diagnostics and NGS techniques. It also emphasizes the current needs to develop new diagnostic techniques targeting fungal pathogens and highlight challenges that lie Ahead. Finally, it discusses the need to develop new bioinformatics solutions that exploit recent advances in the fields of genomics and proteomics that can aid in discovering new biomarkers and developing new diagnostics methods.

- **Chapter 2 (“*Objectives”*)** presents the main objectives of the present thesis.

- **Chapter 3 (“*SeqEditor: an application for primer design and sequence analysis with or without GTF/GFF files”*)** presents the research and work performed for designing and creating SeqEditor a cross-platform standalone desktop application for the analysis of nucleotide and protein sequences, including a set of tools for the search and design of PCR primers, including singleplex, multiplex and target-specific primers. The tool is showcased with a real case study involving the design and experimental validation of primers for detecting the presence of *Candida* pathogens by PCR.

- **Chapter 4 (“*CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies”*)** focuses on the research and work done in the co-development of CrossMapper a tool to estimate cross-mapping rates and optimizing experimental design in multi-species sequencing studies *e.g.* dual RNA-seq. CrossMapper tool is a pipeline assessing, prior to sequencing, the potential rates of multi-mapping and erroneous mapping for various combinations of sequencing parameters and any number of reference sequences.

- **Chapter 5 *(“Recurrent neural networks for classification and regression problems in DNA and RNA sequences with rnnXna”*)** presents the research done to develop a computation technique based on recurrent neural network models implemented to enhance structure profiling of RNA secondary structures.

- **Chapter 6** ***("Applications and pipeline infrastructure of the GPRO suite for RNASeq, DeNovoSeq and VariantSeq analysis")*** presents the approaches done in developing a set of friendly-to-use interfaces to manage, run and control bioinformatics pipelines and. RNASeq provides a set of tools to perform differential expression and functional enrichment analysis. DeNovoSeq provides a set of tools to perform sequence de novo assembly and annotation. VariantSeq provides a set of tools to calling and annotation of (SNPs and INDELs) variants.

- **Chapter 7 (*"Biomarkers of caspofungin resistance in C. albicans isolates: a proteomic approach")*** presents a proteomic approach on LC-MS/MS protein data to find and identify candidate biomarkers of echinocandin resistance *Candida albicans* isolates using differential expression and network analysis.

- **Chapter 8 ("*CandidaMine, an integrative omics database for Candida yeast pathogens")*** presents the integrative research work by the candidate in order to build CandidaMine an integrative database warehouse dedicated to Candida Species omics data. CandidaMine collects and integrates multiple data sets for different database resources to enable advanced mining features for the development of the integrative analysis tools. The chapter also presents a general overview of CandidaMine and its main features and how to use them to mine and perform an integrative analysis.

- **Chapter 9** ***("Summarizing Discussion")*** presents and overall discussion of the results and **Chapter 10** ***("Conclusions")*** presents overall conclusions obtained in this thesis.

Finally, **Appendix A** provides a list of studies in which I have participated during my PhD and the intellectual property[1] registration of the GPRO project. **Appendix B** provides online resources of manuals and tutorials for all the software tools developed in this work.

[1] Intellectual property is the formal action for patenting software in Spain

Table of Contents

1 Introduction

With the advent of the post-genomic era, high throughput technologies have revolutionized the way in which omics studies *e.g.* genomics, transcriptomics, proteomics or metabolomics are performed (van Dijk *et al.* 2014, Goodwin *et al.* 2016). Advances in next generation sequencing (NGS) and proteomics have triggered a bioinformatics revolution involving the full spectrum of Technology Readiness Levels (TRL) to develop tools able to manage and mine biologically meaningful information from omics data (*i.e.* genomes, transcriptomes, proteomes, etc.). Such advances have also triggered a new generation of techniques and applications for clinical diagnostics (Gu *et al.* 2019, Lefterova *et al.* 2015). Despite most attention is still oriented to applications such as cancer prognosis (Luthra *et al.* 2015) or diagnosis of hereditary disorders (Di Resta *et al.* 2018), the interest toward infectious disease diagnosis and monitoring has recently increased. Although most NGS applications in clinical microbiology focus on bacteria and viruses, fungi are equally amenable to NGS. Thus, NGS holds enormous promise of offering novel diagnostics tools for yeast infections (Consortium Opathy and Gabaldón 2019). However, NGS still poses many challenges that need to be addressed before its widespread adoption in the clinics. This includes reducing cost and turnaround time, standardizing protocols and bioinformatics pipelines, pipeline automation, improving reference databases, homogenizing, or integrating heterogeneous data sources, and establishing state of the art quality control measures.

This chapter begins with a brief introduction to different perspectives and recent trends in diagnostics of yeast infections, emphasizing the current needs to develop new diagnostic techniques targeting fungal pathogens. Then new trends and techniques in NGS are introduced, highlighting general and fungal specific challenges, and how these can be applied to diagnostics. Finally, it discusses the need to develop new bioinformatics solutions that exploit recent advances in the fields of genomics and proteomics that can aid in discovering new biomarkers and developing new diagnostics methods.

1.1 Diagnosis of yeast infections: The Old, the New, and the Upcoming

Opportunistic yeast pathogens cause a wide range of infections, from superficial and mucosal infections to disseminated and bloodstream infections, which can often be fatal (Kullberg and Arendrup 2016). The incidence of these pathogens has increased in recent years, becoming a major source of life-threatening infections, especially in immunocompromised and hospitalized or critically ill patients. Among pathogenic yeasts, *Candida* spp. are the most common cause of life-threatening invasive infections (Brown *et al.* 2012). The mortality rates associated with invasive candidiasis remain high at 40% even with antifungal therapy, due to the increasing resistance to antifungals of such pathogens, as well as the newly emerging novel pathogenic species (Papon *et al.* 2013, Gabaldón *et al.* 2016, Ksiezopolska and Gabaldón 2018).

Although the most common cause of candidiasis is *Candida albicans*, the emergence of non-albicans Candida species such as *Candida glabrata*, *Candida dubliniensis*, *Candida parapsilosis*, and *Candida tropicalis* has increased over the past decades (da Matta *et al.* 2017). Another example *is Candida auris* which has recently been recognized as a globally emerging multidrug-resistant species (Geddes-McAlister and Shapiro 2019, Sekyere and Asante 2018). In addition, new virulent pathogens have emerged due to hybridization events among pathogenic and non-pathogenic lineages (Mixão and

Gabaldón 2018). Virulence and antifungal resistance may vary between species (Schmalreck *et al.* 2014), and even between strains of the same species (Farmakiotis and Kontoyiannis 2017), hence it is very important to provide a resolution at the species level for effective therapy.

Classical clinical diagnosis of infectious diseases requires a physician formulating a differential diagnosis and then ordering a series of tests to identify the causative agent. Such tests are currently mostly based on microscopy, selective culture, and/or biochemical approaches (Ellepola and Morrison 2005, Cuenca-Estrella *et al.* 2012, Arendrup *et al.* 2014). Most of these methods require isolation and cultivation of the infective agent from clinical specimens, a process that takes up to 48 hours for most common pathogenic yeasts and can require more time for some samples or species. In addition, the identification procedures require specific expertise, may provide ambiguous results, and are generally time-consuming, delaying an effective diagnosis. The spectrum of conventional testing for pathogens in clinical samples ranges from the identification of microorganisms growing in a selective culture, for example, by biochemical phenotype testing or matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) (Mellmann *et al.* 2009, Vlek *et al.* 2014, Hou *et al.* 2019), the detection of organism-specific biomarkers for example antigen testing by latex agglutination or antibody testing by enzyme-linked immunosorbent assay (ELISA) (Binnicker *et al.* 2012, Hartl *et al.* 2018) or molecular based testing by Polymerase chain reaction (PCR) (Mullis *et al.* 1986) for single pathogen or multiplexed PCR testing for multiple pathogens detection.

Recently, there is a growing interest in the development of alternative methods, based on direct detection of diagnostic molecules. These approaches collectively referred to as molecular diagnostics have the potential to be directly applied to clinical specimens and include proteomics-based methods and the detection of specific DNA sequences. For example, PCR enables the selective amplification of a targeted segment of DNA, generating millions of copies of that sequence (amplicon) within a few hours. The potential diagnostic use of this technique is obvious, as it allows the selective detection of minute amounts of the target DNA by using specific oligonucleotides. Diagnosis can be based merely on the presence of the amplicon (if it is unique for the targeted species), its particular size, or its specific sequence, which can be determined by sequencing or by hybridization to a specific probe. The combination of specific PCR designs with subsequent analysis has led to a plethora of alternative PCR-based approaches that are increasingly used in the diagnosis of yeast infections. In addition, specific patterns in the DNA of infectious microorganisms can be detected without the need of selective amplification by PCR. For instance, by means of direct hybridization with specific probes or by recognizing patterns in the length of fragments produced by enzymatic digestion of the DNA by specific endonucleases. Target-specific primers are central for the identification of species in many microbiological processes or for determining antimicrobial susceptibility as well as load infection. In fungal research for example, target-specific primers are frequently used for the identification of a yeast pathogen that is responsible for a given infection.

Other available methods and those currently under development differ in the need for cultivation of the infectious agent, the ability to directly use a clinical sample, sensitivity and accuracy, cost, time and expertise requirements, as well as in the range of species that can be identified. In addition, some emerging methods hold the promise of being able to readily diagnose both the species and drug resistance profile of the infecting agent. A common drawback of methods based on the detection of DNA is that the detection of the DNA may not necessarily correlate with the presence of actively infecting cells (*i.e.* if the species can also be a commensal), or even with the presence of living cells (DNA from dead cells can also be detected). For this reason, several recent approaches are based on the detection of RNA from actively transcribed genes, which are a better proxy for active cells, and which may also reveal signatures that distinguish invasive from commensal behaviours. The field of diagnosis of yeast infections has advanced significantly in the last decade and is currently experiencing a revolution with the usage of the new (and the not so new) high throughput omics technologies (Consortium Opathy and Gabaldón 2019), as we will explore in the next section.

1.2 Trends in Next generation Sequencing

Nowadays, a variety of NGS platforms are available with ongoing developments to increase throughput, read length, and accuracy. Current technologies range from “sequencing-by-synthesis”, implemented by Illumina, which produces short, accurate reads at high throughput, to nanopore-based sequencing, which produces long reads (up to 1 Mb), but with less throughput and accuracy (Goodwin *et al.* 2016). The increased sequencing capacity combined with reduction of cost and the development of novel approaches has facilitated the implementation of NGS outside the research lab, progressively entering the agronomic, forensic and clinical fields (Berkman *et al.* 2012, Lecuit and Eloit 2014, Børsting and Morling 2015, Hynes *et al.* 2017). More recently, NGS is increasingly being used in clinical microbiology laboratories (Deurenberg *et al.* 2017), as its precision outperforms most of the traditional diagnostic approaches (Turabelidze *et al.* 2013). Furthermore, it can provide very high sensitivity for the diagnostics of rare pathogens, as shown with the identification of *Leptospira* as the causative agent of encephalitis after the failure of 38 different diagnostic tests (Turabelidze *et al.* 2013, Wilson *et al.* 2014). NGS applications are still in its early stage when it comes to clinical mycology compared to bacteriology or virology in which such applications have become routine diagnostic procedures. However, we can see many NGS applications such as whole genome sequencing (WGS), metagenomics, single-nucleotide polymorphisms (SNPs), amplicon sequencing of ribosomal internal transcribed spacer (ITS) and RNA sequencing (RNA-seq) being applied in mycology research. Such applications are likely to be soon translated and adopted as a routine diagnostic procedure in the clinics.

The genetic diversity between different fungal isolates can be studied using microsatellite analysis, however SNPs analysis could be a more accurate marker for the evaluation of recombination and genetic relationships (Araujo 2014). Amplicon sequencing of the ribosomal ITS region and WGS of fungal isolates can be used as the most discriminative approach in genetic research of various fungal pathogens like *Candida* species and as a tool in taxonomic identification (Araujo 2014, Dannemiller *et al.* 2014). As an example; WGS is used for the parallel detection of several resistance markers for *C. glabrata*, which prove to be a good alternative to several PCR/DNA sequencing reactions (Biswas *et al.* 2017). In addition, some pathogenic yeast clades frequently form hybrids (Pryszcz *et al.* 2014, Hagen *et al.* 2015, Schröder *et al.* 2016), leading to the origin of new lineages that can be neglected or misidentified with conventional diagnostic tools. For instance, NGS allowed the identification of different parental and hybrid lineages in *C. orthopsilosis* species (Pryszcz *et al.* 2014, Schröder *et al.* 2016), which with conventional tools were always considered the same agent of infection. This is of particular concern, because these hybrid lineages present high genomic plasticity that may allow unpredictable adaptations to new environments or conditions (Mixão and Gabaldón 2018).

Apart from WGS as a tool for the identification of genomic characteristics of the pathogen, NGS techniques offer more tools for the study of genetic changes due to environmental influences. Gene expression can be studied more directly by transcriptome analysis, which may be explored to identify the presence of transcripts that, for example, are only expressed during invasive growth. In addition, global analysis of gene expression at the RNA level will greatly contribute to the investigation of interaction between the host and the pathogen, and thus will further our understanding of pathogenesis (Hovhannisyan and Gabaldón 2019). With RNA-seq, we can obtain tens of thousands of transcripts from one sample, in an unbiased manner (Van Keuren-Jensen *et al.* 2014). This technology has been used to profile the transcriptome of the host, the pathogen or even both by means of dual RNA-seq during infection (Westermann *et al.* 2012). Transcripts originated from the host and the pathogen can be distinguished by aligning to the corresponding reference genomes (De Cremer *et al.* 2013, Enguita *et al.* 2016), allowing the determination of RNA markers involved in the infection from both sides. While transcriptome analysis using RNA-seq technology can provide reproducible and reliable data, some biases and errors can be introduced during the experimental procedures and bioinformatics analysis as will be explained in the next section, which make it difficult to compare results from different samples (Van Keuren-Jensen *et al.* 2014). This is an important aspect for successful clinical applications.

The study of the microbiota has gained momentum by the introduction of metagenomics (*i.e.* analysis of the collective genetic material present in a microbial community). Deep sequencing of clinical samples or microorganisms isolated from different body sites can provide insights of possible correlations between the microbiome composition, presence of potential pathogens and pathogenic states (Ji and Nielsen 2015), furthermore it provided an important tool to assess the emergent properties of the diverse microbial communities (Nguyen *et al.* 2015, Zoll *et al.* 2016). Although fungi contribute approximately 0.1–1.0% of the total microbiome, it contributes a major role to the equilibrium between all microbiome communities (Qin *et al.* 2010, Botschuijver *et al.* 2017, Mar Rodríguez *et al.* 2016). As reported by Zoll *et al.*, changes in the mycobiome in immunocompromised patients can lead to opportunistic fungal infections (Badiee and Hashemizadeh 2014, Zoll *et al.* 2016). However, as mentioned before, the detection of an organism DNA by NGS does not necessarily correlate with the presence of actively infecting cells or that it is the causative agent of disease, especially in the case of opportunistic fungal pathogens as they can present in a commensal state. Nevertheless, integrating such information with the surrounding microbiome can provide a way to distinguish invasive from commensal behaviours. For instance, Bittinger *et al.* demonstrated that relative abundances of bacteria and fungi in the lung are different during fungal colonization and fungal infection, which can be exploited for clinical diagnosis (Bittinger *et al.* 2014). Metagenomics approaches have been used for monitoring the presence of pathogens and for pathogen detection (Cuomo 2017). However, a culture and PCR free metagenomics for a direct detection of human pathogens in clinical samples has to date only been applied to bacteria, in addition to it has still to overcome some major challenges associated with the low pathogen presence in a clinical sample, and the improvement of the DNA quality, sequencing library preparation methodology and computational bioinformatics pipelines that prevents it from widespread use (Gu *et al.* 2019). In addition, the fact that fungi represents a small portion of total microbiome, may decrease their detectability by currently available sequencers like Illumina NextSeq and HiSeq (Zoll *et al.* 2016) and impose a fungal specific limitation to the applicability of such technology to the determination of mycobiomes for clinical diagnostics. However; it is possible to use targeted sequencing of amplicons of the ITS region of the fungal ribosomal genes to study the fungal component of a complex microbial ecosystem (Huseyin *et al.* 2017). For example; McTaggart et al, has developed an NGS-based method based on ITS1 amplicon sequencing and a custom computational pipeline to detect a broad range of fungi in bronchoalveolar lavage (BAL) specimens and applied it to the analysis of the fungal microbiome of the lung during fungal infection, demonstrating its potential for distinguishing fungal infection from colonization (McTaggart *et al.* 2019).

The introduction of a newer generation of sequencing machines such as MinION opens new and promising possibilities to have direct and real time approaches for clinical fungal diagnostics. MinION based on Oxford Nanopore Technologies is a portable and a palm-sized sequencing device capable of generating ultra-long sequence reads in real-time, with low initial start-up costs and ease of use. For instance a recent study has presented the usage of MinION-based NGS sequencing to confirm the diagnosis of an invasive fungal infection by detecting *Pneumocystis jirovecii* directly from BAL and sputum specimens (Irinyi *et al.* 2020).

1.3 Bioinformatic tools and Databases

Advances in NGS and other high-throughput omics technologies have triggered a revolution in the field of bioinformatics, driven by the need to develop tools to manage and mine the ever-increasing amounts of biological information. These tools no longer consist of a single piece of software coupled with a knowledge database, but rather take the form of a complex workflow, commonly referred to as a “pipeline”, where different tools are sequentially or simultaneously used, including those consulting existing databases.

1.3.1 Pipelines and workflows

The implementation of bioinformatics pipelines in routine application of fungal omics is still a challenge for many clinical and medical research centres. Yet, there is a long path to go from the successful proof of concept of novel pipelines to its readiness for routine clinical use. Ideally, diagnostic tools must be cheap, fast, sensitive, accurate, reproducible, and easy to use. In the context of host-pathogen interactions, pipelines have been designed to analyse NGS data from various pathogens, including fungi, in both clinical and environmental settings. Genomes and transcriptomes of fungal pathogens, as well as metagenomes from complex microbial communities can be reconstructed from raw sequencing reads with graph-based algorithms that assemble NGS reads into contigs, scaffolds or chromosomes (Miller *et al.* 2010). Sequence features, like genes, are then predicted and annotated, which can be used to infer the phenotypic potential of the studied pathogen(s). Once an annotated genome is available, it can be used as a reference sequence to map newly obtained sequencing reads from genomes or transcriptomes belonging to strains from the same or related species. This process is based on read alignment or mapping algorithms (Li and Homer 2010).

The most common re-sequencing data analysis pipelines are mainly oriented to genotyping and comparative transcriptomic analyses by means of RNA-seq. The former is designed to call and annotate mutations at the genome-wide level or via gene panels and capture systems (*e.g.* the entire set of known resistance genes or resistome) to elucidate potential markers – mainly SNPs and INDELs (insertions and deletions) – related to the evolution, population dynamics, transmissibility and infectivity of fungal pathogens (Wilkening *et al.* 2013, Cornish and Guda 2015). In RNA-seq analysis for gene expression profiling, there are three main types of pipelines: mRNA-seq, small RNA-seq and dual RNA-seq. The first two are designed to quantify and compare the expression patterns of the mRNAs and small RNAs expressed by fungal pathogens under a particular condition. The dual RNA-seq pipeline deals with data obtained by simultaneous sequencing of host and pathogen RNA, and thus focuses on the crosstalk between pathogens and their host at different stages of the infection (Enguita *et al.* 2016, Wang *et al.* 2016, Hovhannisyan and Gabaldón 2019). Other mapping pipelines in fungal research focus on processing reads sequenced via 18S rRNA/ITS amplicons. These pipelines are widely used to obtain diversity landscapes of fungal communities to investigate any plausible interrelation among the abundance distribution of the species inhabiting the fungal community and the characteristics of the sampled environment (White *et al.* 2013, Cole *et al.* 2014).

From a bioinformatics perspective such analysis involves many different intermediate steps interacting with each other to go from raw sequencing reads to biological insights. For example, RNA-seq involves a pre-processing step to check quality of raw reads for any factor that can affect or introduce bias in downstream analysis, filter raw reads, mapping those reads to a reference genome, postprocess, filter and quantify the mapped reads, then performing differential expression tests. Each different step may require running more than one tool to perform the required task, in addition alternative tools could exist for each individual task. Running such bioinformatics analysis requires managing and installing third-party software, writing pipeline scripts, and normally using a command-line interface (CLI) instead of a graphical user interface (GUI). For some users, this may be a complex procedure as it requires knowledge of a shell usage (*e.g.* UNIX shell) and a scripting language such as bash to write and run bioinformatics analysis. CLI versus GUI is normally a trade-off between control over simplicity. Hence most bioinformaticians prefer the flexibility of CLI usage for them to tweak it as needed depending on the required analysis. On the contrary non-skilled bioinformaticians such as clinical or research laboratory personnel prefer to use interactive and easy to use GUI software with a fast learning curve that can boost their productivity. Automation is an important aspect for the bioinformatics pipelines in diagnostic applications. Automation allows such pipelines to be operated with small or minimal intervention from users, however automated pipelines should have a high fault tolerance to ensure that they will be able to recover in case of technical errors that affect their executions. In addition, automated pipelines should be able to detect possible soft or hidden data processing errors that, if unnoticed, could have a negative impact on downstream analyses and results. Also, it is worth noting that the purpose of NGS analyses performed in the clinical laboratory for patient care may differ from those in a research setting, even though the sequencing methods may be the same. As such, a clinical-oriented pipeline must follow detailed clinical laboratory standards, including guidelines for data storage to ensure its accuracy and precision in terms of reproducibility and repeatability of clinical tests , as well as the protection of personal clinical data (Rehm *et al.* 2013).

1.3.2 Databases

Over the last decades, the increasingly frequent recourse to -omics and bioinformatics to generate valuable knowledge from the high-throughput data has resulted in a wide repertory of resources that have improved and updated previously existent reference databases. The most popular databases for NGS data annotation are: the non-redundant (nr) and the Reference Sequence (RefSeq) database of the National Center for Biotechnology Information (Sayers *et al.* 2020), the UniProt Knowledgebase (Schneider *et al.* 2009) and Ensembl (Kersey *et al.* 2016) . For coding sequences, the gene identifiers or accessions provided by the aforementioned databases are correlated with other vocabularies like the Gene Ontology (GO) vocabulary (Gene Ontology Consortium 2015) or the system of Enzyme Commission (EC) numbers (Bairoch 2000) (http://enzyme.expasy.org), both facilitating a better understanding of the functional role of the annotated gene. On another level other databases maintain and curate protein and genetic interactions such as STRING database (Szklarczyk *et al.* 2019) and BioGRID (Oughtred *et al.* 2019). From a research perspective, archiving experimental data is an important key to the progress of reproducible science. For this reason, the sequence read archive (SRA) (Leinonen *et al.* 2011) was established as a public repository for next-generation sequence data as a part of the International Nucleotide Sequence Database Collaboration (INSDC). However, from a diagnostic perspective, clinical data generated by NGS will be subject to data privacy and protection.

In addition, the progression of fungal omics and bioinformatics has also promoted the emergence of a wide variety of databases and repositories specifically dedicated to fungi, including yeast pathogens.

- *Saccharomyces* Genome Database (SGD) (Cherry *et al.* 2012) collects community resources about the budding yeast

Saccharomyces cerevisiae aiming to integrate curated information and experimental results.

- *Candida* Genome Database (CGD) (Skrzypek *et al.* 2017) is modelled after SGD, CGD provides access to genomic and proteomic data and manually curated functional information about genes and proteins of the fungal pathogens *C. albicans, C. glabrata, C. parapsilosis, C. dubliniensis* and other *Candida* species along with web-based tools for accessing, analysing and exploring these data. CGD aims to achieve a real time curation of the literature and connect literature annotations to the latest version of the genomic sequences and their annotations.

- Candida Gene Order Browser (CGOB) (Fitzpatrick *et al.* 2010, Maguire *et al.* 2013) and Yeast Gene Order Browser (YGOB) (Byrne and Wolfe 2005) provide a manually curated homologous genes datasets in *Candida* species and other yeast species, respectively.

- EnsemblFungi is more general and contains many genomes for a wide range of fungal organisms and is a subset of Ensembl genomes (Kersey *et al.* 2016).

- MycoCosm (Grigoriev *et al.* 2014) integrates fungal genomics data and analytical tools to perform comparative genomics.

- ISHAM ITS barcode database for human and animal pathogenic fungi (Irinyi *et al.* 2015, Meyer *et al.* 2019) was established and maintained by ISHAM working group for Barcoding of Medical Fungi in 2015 as an effort to standardize sequence-based identification of fungi using ribosomal ITS1/2 region. The ribosomal ITS1/2 region was proposed as the primary fungal DNA barcode (Schoch *et al.* 2012), However; using ITS1/2 region in human pathogenic fungi was only able to identify approximately 75% of all fungal species accurately at the species level (Irinyi *et al.* 2016), therefore the translocation elongation factor 1 alpha (TEF1α) gene was proposed as a secondary fungal DNA barcode (Stielow *et al.* 2015). ISHAM Barcoding database maintains quality controlled reference sequences for both ITS1/2–and TEF1α targets of human and animal pathogenic fungi. Using both targets for fungi enables the identification of the majority of human and animal pathogenic fungi (Hoang *et al.* 2019).

There are other databases that are specialized in maintaining information on host-pathogen interactions to identify genes and proteins involved in host-pathogen interaction pathways, which could be used as possible targets for new drug discovery.

- FungiDB (Basenko *et al.* 2018) includes whole genome sequences and annotations, experimental and environmental isolate sequence data, comparative genomics, analysis of gene expression, as well as supplemental bioinformatics analyses and a web interface for data mining.

- The Pathogen-Host Interaction database (PHI-base) (Urban *et al.* 2017) is devoted to store and maintain effectively the vast and growing number of proven genes that have a role in pathogenicity. PHI-base includes a wide variety of hosts and pathogens, and *C. albicans* and *C. glabrata* are among them. PHI-base highly depends on domain experts to manually curate new gene entries provided by strong experimental evidence (gene disruption experiments) and literature references.

- The Host-Pathogen Interaction Database (HPIDB) (Kumar and Nanduri 2010, Ammari *et al.* 2016) stores and maintains protein-protein interactions from diverse mammalian and plant hosts infected by fungi, bacteria and other pathogens.

In addition, according to the International Code of Nomenclature for algae, fungi and plants (ICN) (Turland *et al.* 2018), it is a mandatory requirement to register fungal names for valid publications. MycoBank (Robert *et al.* 2013) is a database to document mycological nomenclatural novelties (new names and combinations) and its associated data. MycoBank registration system represents a coordination channel between different databases, such as the Index Fungorum and the Fungal Names (http://www.indexfungorum.org/), and it eases the registration process for the scientific community. StrainInfo (Dawyndt *et al.* 2005, Verslyppe *et al.* 2014) is an open platform designed to contain all known information about a particular microorganism at the strain level, giving a unique passport to every strain, with a number that traces back to the same isolate, thus providing a uniform overview for all known equivalent strain numbers. The CBS strain database (https://wi.knaw.nl/page/Collection) and the American Type Culture Collection (ATCC) (https://www.lgcstandards-atcc.org/) are another two examples of such repositories.

Most of the mentioned databases are isolated, with minimal or no interaction between them, thus making knowledge extraction from them is a complex task if multiple sources are involved. Normally mining such databases to form new insights requires a complex process including extracting raw information, data processing or wrangling and data integrating, then providing such insight in an easy to digest format *i.e.* interactive reports or visualizations. Federated databases and knowledge integration will certainly constitute the cornerstone of future automated diagnostic applications. Therefore, it is of capital importance to ensure their correct maintenance and to minimize the amount of introduced errors with the use of both manual curation and automated detection of potential errors (Stavrou *et al.* 2018).

1.3.3 Integration Approaches

In the emerging field of personalized and precision medicine, bioinformatics and integrative biology have gained increased attention for their promise to develop more efficient tools for integration and analysis of multi-omics data of complex diseases, including fungal infections and associated clinical data. Integrative analysis of multi-omics data is motivated by the basic idea that to fully understand any biological system, the underlying molecular mechanisms should be considered in the analysis (Kristensen *et al.* 2014, Rotroff and Motsinger-Reif 2016) Fungal infections are characterized by an interaction between the fungal pathogen and host cells. The integration and analysis of genomic, proteomic, metabolomic data, as well as other meta-data from the patient medical record represents an excellent opportunity to model infection, to make clinically-relevant predictions, and to discover biomarkers and target-specific drugs (Durmuş *et al.* 2016, Larsen *et al.* 2015, Culibrk *et al.* 2016). Among all the different methodologies that could be used for personalized and precise diagnosis based on multi-omics, the most promising are Bayesian networks (BNs), decision trees, artificial neural networks, nature-inspired and evolutionary algorithms (Larrañaga *et al.* 2013, Bersanelli *et al.* 2016). Of course, a domain expert (*e.g.* biologist or medical doctor) should be involved in the knowledge discovery process to interpret and validate the final result (Holzinger 2016, Holzinger and Jurisica 2014, Obermeyer and Emanuel 2016). Some existing tools are

BNOmics (Gogoshin *et al.* 2017), a reconstruction and modelling framework able to reverse engineer and model networks, and PARADIGM (Vaske *et al.* 2010), which applies BNs to identify patient-specific pathway activities by means of probabilistic inference (Larrañaga *et al.* 2013).

Overall, bioinformatics data analysis and knowledge database development are quickly evolving fields in fungal research. Given the continuously dropping costs of NGS and other -omics methods, data accumulation, storage, further analysis, and interpretation will evidently become more widespread in routine mycology labs. Moreover, the integration of multi-omics data holds a great potential towards understanding host-fungus interactions, potentially allowing to translate this knowledge into clinical practice, including diagnosis and prognosis. However, today there are still several important challenges, both technical and fundamental, that will need to be overcome.

2 Objectives

The main objectives of this thesis were to:

- Design, develop, and test in a proof-of-concept application, a friendly GUI software for genomic and proteomic sequences analysis, and a set of tools to increase researcher's productivity in designing and optimizing primers for PCR experiments.

- Design and develop computational methods to i) estimate cross-mapping rates to optimize experimental design in multi-species sequencing experiments and ii) predict the structural features of RNA molecules.

- Design and develop a set of GUI software and a bioinformatic infrastructure for managing, running, and automating bioinformatic workflows dedicated to fungal omics analysis.

- Identify candidate biomarkers of echinocandin resistance in *Candida albicans.*

- Design and develop an integrative database warehouse dedicated to *Candida* Species to host and integrate multiple data sets from different database resources with a set of tools to apply integrative analysis.

3 SeqEditor: an application for primer design and sequence analysis with or without GTF/GFF files

Hafez, Ahmed, Ricardo Futami, Amir Arastehfar, Farnaz Daneshnia, Ana Miguel, Francisco J. Roig, Beatriz Soriano, Jaume Perez-Sánchez, Teun Boekhout, Toni Gabaldón, and Carlos Llorens. 2020. "SeqEditor: an application for primer design and sequence analysis with or without GTF/GFF files." *Bioinformatics*. doi: 10.1093/bioinformatics/btaa903.

3.1 Abstract

3.1.1 Summary

SeqEditor is a cross-platform desktop application for the analysis of nucleotide and protein sequences. It is managed through a Graphical User Interface (GIU) and can work either as a graphical sequence browser or as a fasta task manager for multi-fasta files. SeqEditor has been optimized for the management of large sequences, such as contigs, scaffolds or even chromosomes, and includes a GTF/GFF viewer to visualize and manage annotation files. In turn, this allows for content mining from reference genomes and transcriptomes with similar efficiency to that of command line tools. SeqEditor also incorporates a set of tools for singleplex and multiplex PCR primer design and pooling that uses a newly optimized and validated search strategy for target and species-specific primers. All these features make SeqEditor a flexible application that can be used to analyse complex sequences, design primers in PCR assays oriented for diagnosis, and/or manage, edit, and personalize reference sequence data sets.

3.1.2 Availability and implementation

SeqEditor was developed in Java using Eclipse Rich Client Platform and is publicly available at: https://gpro.biotechvana.com/download/SeqEditor.

The user manual and tutorials are available online at: https://gpro.biotechvana.com/tool/seqeditor/manual.

3.2 Introduction

In the modern post-genomic era, analysis of DNA/RNA and proteins at the molecular sequence level is essential when characterizing gene features, regulatory elements and functional patterns of the vast amount of genomic and transcriptomic sequences assembled from *de novo* sequencing projects. Sequence analysis is also of significant importance in clinical medicine where PCR serves as an invaluable tool for diagnosis of infectious diseases. This has prompted the development of new and ever more effective tools for designing species-specific primers for PCR assays oriented to identify viruses like the SARs-CoV2 coronavirus (Attwood *et al.* 2020) fungal pathogens (Arastehfar *et al.* 2019), and/or multi-resistant bacteria (Strommenger *et al.* 2003). Most currently available tools for sequence analysis, including BuddySuite (Bond *et al.* 2017) and FAST (Lawrence *et al.* 2015) or primer design, including Primer3 (Untergasser *et al.* 2012), use a Command Line Interface (CLI) because of its efficiency and functional versatility. Using CLI, however, requires users to understand the basics of command line coding and this excludes researchers that do not have the bioinformatic training to manage CLIs. In addition, while the lack of graphical requirements for CLIs allow for efficient performance of various tasks, they do not allow graphical visualization of sequences. This often limits the practicality of many modern sequence analysis tools for investigating sequence patterns that require some degree of human abstraction for their identification. Because of this, among other reasons, there is a need to develop sequence analysis tools that use Graphical User Interfaces (GUIs). Unlike CLIs, GUIs do not require any knowledge of coding and are more accessible for people regardless of their computer literacy skills. Yet despite these advantages, sequence analysis tools using GUIs are still limited in their capacity to work with large and/or multiple sequences. As a result, they are often less efficient than CLI tools when managing complex datasets like genomes and transcriptomes. To address these limitations, we developed SeqEditor, a GUI based cross-platform desktop application for sequence analysis of the GPRO suite (Futami *et al.* 2011).

3.3 Overview

SeqEditor is a desktop Java application implemented with the following components; i) an upgraded version of TIME editor, the former graphical sequence editor of the GPRO suite (Munoz-Pomer *et al.* 2011) that is suited for the analysis of large nucleotide and protein sequences; ii) a fasta task manager that allows users to work with reference genomic, transcriptomic and proteomic data sets in fasta format; iii) a set of tools for primer design and pooling that allows for the design of specific primers used in singleplex and/ or multiplex PCR assays; iv) a GTF/GFF viewer for the management of GTF and GFF files associated to reference genomes. Details on the SeqEditor layout (including directories and menus) are provided in the supplementary data accompanying this article.

3.4 The Sequence Browser

The sequence browser is a graphical screen (formerly called TIME editor) that has been upgraded and optimized to manage very large sequences such as contigs, scaffolds, and chromosomes. The Sequence browser performs the same tasks as its predecessor (TIME editor), including sequence editing, searching and filtering for Open Reading Frames (ORFs) and motifs, translation of nucleotides to proteins using either the universal or a user-defined code, and changing the geometry and the orientation of the sequences (See section S1.2 in the supplementary data). The browser screen is also interactive, allowing for the visualization of any analyses as well as any edits to the browsed sequence. While only one sequence can be viewed per browser screen, as in the former TIME editor, it is now possible to use additional screens to analyse multiple sequences separately. Thus, when opening a multi-fasta file (with multiple sequences) a summary view of the sequences included in the file will be opened at the bottom of SeqEditor. The summary view is dynamic, so users only need to click on the sequence name to open it in a new screen. The summary view presents additional tools for the sorting of the sequences in the summarized fasta file or for the editing of sequence names.

3.5 Fasta task manager

SeqEditor implements a fasta task manager that allows for several fasta files with multiple sequences to be processed and analysed simultaneously, allowing for users to manage genome, transcriptome, and proteome files with similar efficiency to that of CLI tools. Some of these tasks are the same as those already included in the sequence browser (*e.g.* change sequence geometry and orientation, find ORFs and motifs, or translate ORFs to proteins), but while the browser will only process one sequence at the time, the fasta file manager can perform these tasks simultaneously on multiple files. For example, this could entail extracting the ORFs of several different bacterial genomes simultaneously. The fasta task manager of SeqEditor also allows for the filtering, sorting, removal, masking, and splitting of one or more fasta file sequences using search terms or matching criteria that are provided by the user, such as sequence names (exact or partial), sequence length, percentage of sequence indeterminations. Finally, users can simultaneously infer the size of all sequences contained in the fasta files and obtain a full set of metrics (*e.g.* number of sequences, size of the largest and shortest sequence, N50, L50, See section S1.3 Supplementary data for further details). The fasta task manager interfaces of SeqEditor are accessible from the main menu.

3.6 Set of Tools for PCR primer design and pooling

For primer design and pooling in singleplex and multiplex PCR experiments, SeqEditor provides three tools: *SinglePlexPCR, MultiPlexPCR*, and *PrimerPooler*. These tools are adapted from two CLI tools: *Primer3* (Untergasser *et al.* 2012) and *PrimerPooler* (Brown *et al.* 2017). To improve on these tools, *SinglePlexPCR* searches for suitable primer candidates within singleplex PCR experiments by applying *Primer3* search algorithm to search one or more sequences. For multiplex experiments, *MultiPlexPCR* executes an optimized search process based on two new algorithms that are specific to SeqEditor (Algorithm 3.1 and Algorithm 3.2 are described in Section 3.9.4 of the Supplementary Materials). The search builds a complex index over potential multiplex primer sets (using Algorithm 3.1) and then applies a greedy search strategy using the index (applying Algorithm 3.2). *PrimerPooler* allows users to input a list of primers and divide them into different pools, optimizing the multiplex PCRs primer search. *SinglePlexPCR*, *MultiPlexPCR*, and *PrimerPooler* are accessible from the browser´s main menu and will display the results in an interactive summary interface. They also allow users to manage and export the results in various file formats including csv, fasta or GTF/GFF.

3.7 GTF/GFF Viewer

SeqEditor implements a GTF/GFF viewer that allows for the mining of tasks on reference genomes or transcriptomes using the annotations provided in GTF, GFF, or bed files. The viewer is a dynamic grid of rows and columns that displays the contents of the GTF/GFF file and that provides options to search, filter, and/or extract the contents (*e.g*. chromosomes, genes, exons) from the genome or transcriptome assembly using specified search and/or matching criteria. The GTF/GFF viewer also allows editing tasks on the GTF/GFF file to manually correct or curate annotations, or for personalizing it. For example, the user can manually adjust the coordinates of a predicted exon to set the correct coordinates according to the knowledge of the user for that feature. The tasks and options provided by the GTF/GFF viewer from the main menu as detailed in section S1.5 of the supplementary data.

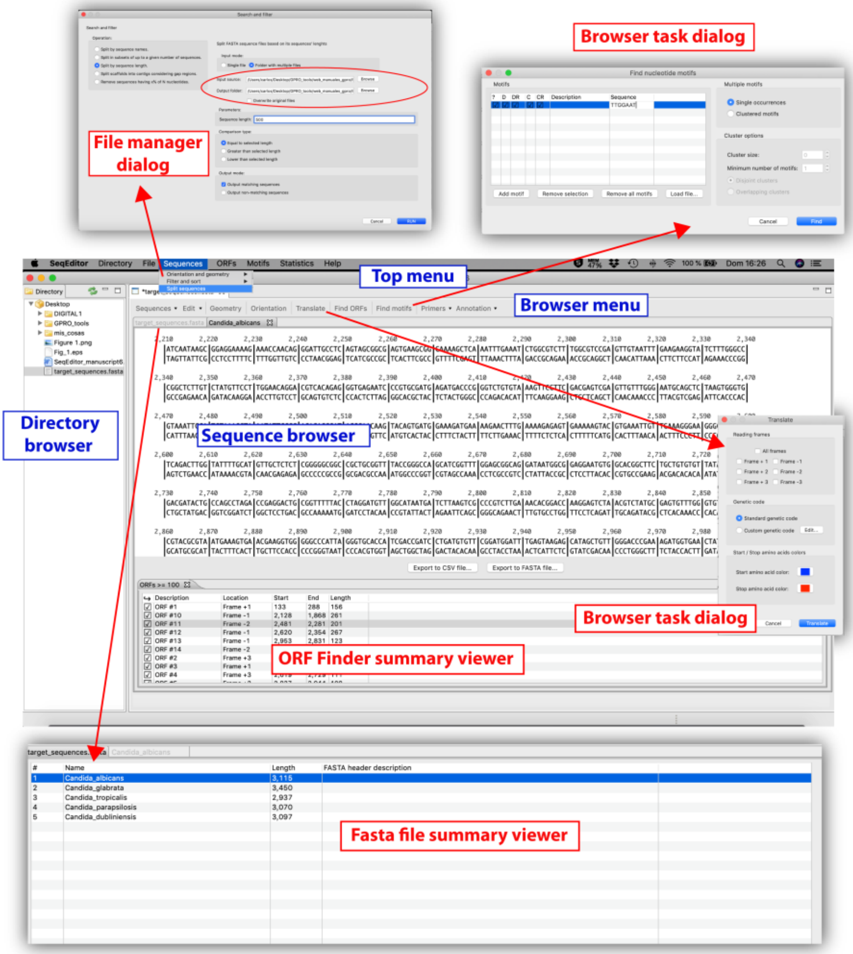
3.8 Performance

We compared the features and performance of SeqEditor respect to its predecessor (the TIME editor) as well as to other state-of-the-art tools such as GeneRunner (http://www.generunner.net), Geneious (https://www.geneious.com/), and Sequencher (https://www.genecodes.com). The comparison is provided in section 3.9.6 of Supplementary Data. Since the applications vary considerably in their functionality, we focused our comparison on the features and the operability criteria of each tool. Overall, SeqEditor showed a clear advantage over the other tools due to its graphical efficiency when handling very large sequences and its ability to multitask on multiple fasta files. Specific features unique to SeqEditor, such as its implementation of the GTF/GFF viewer or the new search strategy applied to Primer3 to find target and species-specific primers, make this a practical tool when working with reference genomes and transcriptomes or for designing multiplex assays oriented to identify specific pathogens via PCR. While it should be noted that GeneRunner, Geneious, and Sequencher contain functions that are not yet implemented in SeqEditor (*e.g.* graphical tools to read sanger chromatograms or for plotting restriction sites, motifs and ORFs as well as other tools for comparative analyses), we are committed to implementing these additional tools in further updates of SeqEditor. Finally, another non-trivial aspect to highlight of SeqEditor is that while Geneious and Sequencher are commercial applications distributed under a pay-for-use license, SeqEditor is freely available for download like GeneRunner. SeqEditor thus provides a cost-free, straightforward, and effective application for sequence analysis and primer design.

3.9 Supplementary Materials

3.9.1 An overview of SeqEditor

The layout of SeqEditor is organized into four different compartments: the “directory browser”, the “top menu”, the “sequence browser”, and the “browser menu” (highlighted in blue in Figure 3.1). When clicking on these sections, additional dialogs will often open to allow the user to run analyses and display summaries for each executed task (Figure 3.1, red boxes). The top menu provides access to the sequence browser and to other interfaces that execute file manager tasks. This sequence browser can also be called by right-clicking on an input sequence file in the directory browser, which can be pre-set as any folder in their PC. The sequence browser menu organizes the tasks for sequence analysis and provides two functions that are new to the current version of SeqEditor. These are the PCR primer design and the annotation file (GTF/GFF) viewer.



ORFs >= 100

↪	Description	Location	Start	End	Length
☑	ORF #1	Frame +1	133	288	156
☑	ORF #10	Frame -1	2,128	1,868	261
☑	ORF #11	Frame -2	2,481	2,281	201
☑	ORF #12	Frame -1	2,620	2,354	267
☑	ORF #13	Frame -1	2,953	2,831	123
☑	ORF #14	Frame -2			
☑	ORF #2	Frame +3			
☑	ORF #3	Frame +1			
☑	ORF #4	Frame +3			

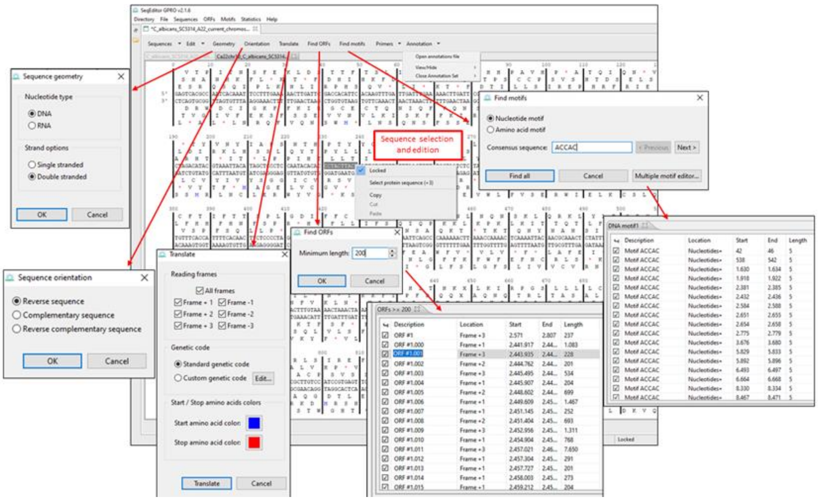
target_sequences.fasta Candida_albicans

#	Name	Length	FASTA header description
1	Candida_albicans	3,115	
2	Candida_glabrata	3,450	
3	Candida_tropicalis	2,937	
4	Candida_parapsilosis	3,070	
5	Candida_dubliniensis	3,097	

Figure 3.1 Main features of the SeqEditor’s layout.

3.9.2 Graphic visualization and analysis of single sequences

SeqEditor allows for the visualization, editing, and analysis of individual sequences via the sequence browser. The menu of the browser provides various options for editing, translation, and searching within the visualized sequence. Specifically, by first selecting the desired translation frame and genetic code, sequences can be edited, their geometry and orientation changed, or translated into proteins. Nucleotide or protein motifs of interest, as well as ORFs, can be searched for within the visualized sequences. The sequence browser menu also gives direct access to the primer design tools and the GTF/GFF viewer, which will be discussed in more detail in sections 3.9.4 and 3.9.5. Figure 3.2 shows the general layout of the sequence browser and the secondary interfaces that can be accessed through it.


SeqEditor GPRO v2.1.6
Directory File Sequences ORFs Motifs Statistics Help
*C_albicans_SC5314_A22_current_chromos...
Sequences Edit Geometry Orientation Translate Find ORFs Find motifs Primers Annotation
Open annotations file
View/Hide
Close Annotation Set
Ca22chr1A_C_albicans_SC5314...
Sequence selection and edition
Locked
Select protein sequence (+3)
Copy
Cut
Paste
OK
Cancel
Translate
Edit...
Find all
< Previous
Next >
Multiple motif editor...
Locked

Sequence geometry

Nucleotide type: ● DNA ○ RNA

Strand options: ○ Single stranded ● Double stranded

Sequence orientation

● Reverse sequence
○ Complementary sequence
○ Reverse complementary sequence

Translate

Reading frames: ☑ All frames; ☑ Frame + 1 ☑ Frame -1; ☑ Frame + 2 ☑ Frame -2; ☑ Frame + 3 ☑ Frame -3

Genetic code: ● Standard genetic code ○ Custom genetic code

Start / Stop amino acids colors: Start amino acid color: 🟦 Stop amino acid color: 🟥

Find ORFs

Minimum length: 200

Find motifs

● Nucleotide motif
○ Amino acid motif

Consensus sequence: ACCAC

ORFs >= 200

	Description	Location	Start	End	Length
☑	ORF #1	Frame +3	2.571	2.807	237
☑	ORF #1.000	Frame +1	2.441.917	2.44...	1.083
☑	ORF #1.001	Frame +3	2.443.935	2.44...	228
☑	ORF #1.002	Frame +2	2.444.762	2.44...	201
☑	ORF #1.003	Frame +3	2.445.495	2.44...	534
☑	ORF #1.004	Frame +1	2.445.907	2.44...	204
☑	ORF #1.005	Frame +2	2.448.602	2.44...	699
☑	ORF #1.006	Frame +1	2.449.609	2.45...	1.467
☑	ORF #1.007	Frame +1	2.451.145	2.45...	252
☑	ORF #1.008	Frame +2	2.451.404	2.45...	693
☑	ORF #1.009	Frame +3	2.452.956	2.45...	1.311
☑	ORF #1.010	Frame +1	2.454.904	2.45...	768
☑	ORF #1.011	Frame +3	2.457.021	2.46...	7.650
☑	ORF #1.012	Frame +1	2.457.304	2.45...	291
☑	ORF #1.013	Frame +1	2.457.727	2.45...	201
☑	ORF #1.014	Frame +1	2.458.003	2.45...	273
☑	ORF #1.015	Frame +1	2.459.212	2.45...	204

DNA motif1

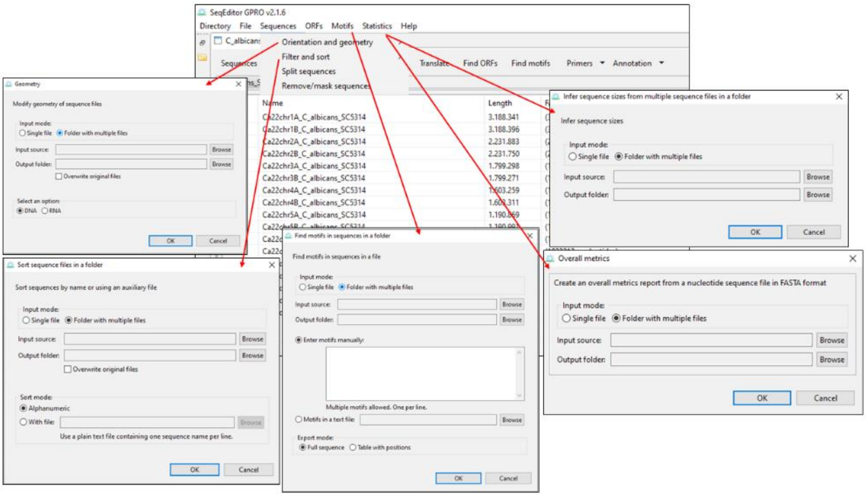
	Description	Location	Start	End	Length
☑	Motif ACCAC	Nucleotides+	42	46	5
☑	Motif ACCAC	Nucleotides+	538	542	5
☑	Motif ACCAC	Nucleotides+	1.630	1.634	5
☑	Motif ACCAC	Nucleotides+	1.918	1.922	5
☑	Motif ACCAC	Nucleotides+	2.381	2.385	5
☑	Motif ACCAC	Nucleotides+	2.432	2.436	5
☑	Motif ACCAC	Nucleotides+	2.584	2.588	5
☑	Motif ACCAC	Nucleotides+	2.651	2.655	5
☑	Motif ACCAC	Nucleotides+	2.654	2.658	5
☑	Motif ACCAC	Nucleotides+	2.775	2.779	5
☑	Motif ACCAC	Nucleotides+	3.676	3.680	5
☑	Motif ACCAC	Nucleotides+	5.829	5.833	5
☑	Motif ACCAC	Nucleotides+	5.892	5.896	5
☑	Motif ACCAC	Nucleotides+	6.493	6.497	5
☑	Motif ACCAC	Nucleotides+	6.664	6.668	5
☑	Motif ACCAC	Nucleotides+	8.330	8.334	5
☑	Motif ACCAC	Nucleotides+	8.467	8.471	5

Figure 3.2 General layout of the sequence browser and its sequence analysis tools.

As shown in the figure above, by selecting an individual sequence in the fasta file, SeqEditor will display the sequence in the central panel. By right clicking on sections of the sequence, these can be cut or copied for editing (see grey highlighting in the centre of the panel in Figure 3.2). Sequence editing can be manually locked or allowed to facilitate sequence editing. By clicking on the different options provided in the sequence browser menu (Figure 3.2 top), secondary windows will open providing editing, translation, and analysis criteria for each tool (depicted in red arrows in Figure 3.2). By opening the ‘Find ORFs’ and ‘Find motifs’ dialogs and selecting either the minimum ORF length or the desired motif, SeqEditor will open a third interface at the bottom of the browser with the search results (bottom, right in Figure 3.2).

3.9.3 Processing and analysing one or more multi-fasta files

The sequence browser allows for the visualization and analysis of sequences via an interactive GUI. In addition, SeqEditor has a fasta task manager accessible via the top menu to allow for multiple fasta files to be processed and analysed simultaneously without the need to open these files separately. This feature of SeqEditor allows for the entire genomes, transcriptome, and proteome files to be managed, edited, and analysed at the same processing speed as CLI tools. As shown in Figure 3.3, the task manager can be operated via the top menu and it provides various options for editing, sorting, and metric analysis of the fasta files by uploading them directly into the corresponding dialog box. Four examples of the different interfaces that are accessible from the top menu are shown, namely the ‘Change geometry orientation in fasta files’, the ‘Sort sequence files in a folder’ (both accessible through the ‘Sequences’ tab); the ‘Find motifs in a folder’ dialog (accessible through the ‘Motifs’ tab) and the two options from statistical analysis of the files that will open from the ‘Statistics’ tab (Inferring sequence sizes and overall metrics). By selecting the desired input source and mode (referring to the fasta files to analyse) as well as the desired output folder, SeqEditor will execute the selected task and placing the output in the desired file format.



Name	Length
Ca22chr1A_C_albicans_SC5314	3.188.341
Ca22chr1B_C_albicans_SC5314	3.188.396
Ca22chr2A_C_albicans_SC5314	2.231.883
Ca22chr2B_C_albicans_SC5314	2.231.750
Ca22chr3A_C_albicans_SC5314	1.799.298
Ca22chr3B_C_albicans_SC5314	1.799.271
Ca22chr4A_C_albicans_SC5314	1.603.259
Ca22chr4B_C_albicans_SC5314	1.603.311
Ca22chr5A_C_albicans_SC5314	1.190.869

Geometry

Modify geometry of sequence files

Input mode: ○ Single file ● Folder with multiple files

Input source:

Output folder:

□ Overwrite original files

Select an option: ● DNA ○ RNA

Sort sequence files in a folder

Sort sequences by name or using an auxiliary file

Input mode: ○ Single file ● Folder with multiple files

Input source:

Output folder:

□ Overwrite original files

Sort mode:
● Alphanumeric
○ With file:

Use a plain text file containing one sequence name per line.

Find motifs in sequences in a folder

Find motifs in sequences in a file

Input mode: ○ Single file ● Folder with multiple files

Input source:

Output folder:

● Enter motifs manually:

Multiple motifs allowed. One per line.

○ Motifs in a text file:

Export mode: ● Full sequence ○ Table with positions

Infer sequence sizes from multiple sequence files in a folder

Infer sequence sizes

Input mode: ○ Single file ● Folder with multiple files

Input source:

Output folder:

Overall metrics

Create an overall metrics report from a nucleotide sequence file in FASTA format

Input mode: ○ Single file ● Folder with multiple files

Input source:

Output folder:

Figure 3.3 Top menu for the fasta task manager

In Summary, through the top menu, sequences contained in the fasta files can be filtered and sorted, their orientation and geometry can be changed, or they can be masked or removed from the file. Likewise, ORFs and other nucleotide and protein motifs can be searched for and then be exported as a different file. Lastly, statistics on the sequences in the file can be provided using the 'Statistics' button. In this manner, several metrics can be extracted from the sequences contained in the fasta files and exported in an excel file. Such metrics include sequence length, N50 and L50, and the shortest and longest sequence found.

3.9.4 Implementation of search algorithms in SingleplexPCR and MultiplexPCR

SeqEditor provides a set of three tools for PCR primer design and pooling called: *SinglePlexPCR, MultiPlexPCR,* and *PrimerPooler*. These tools are based on two CLI tools - Primer3 (Untergasser *et al.* 2012) and PrimerPooler (Brown *et al.* 2017) but also incorporate a newly optimized search process for multiplex and target-specific primer design based on two algorithms (Algorithm 3.1 and Algorithm

3.2, described below). Following is a discussion of the search algorithms used by *SinglePlexPCR* and *MultiPlexPCR*:

- ***SinglePlexPCR*** uses an optimized search based on the Primer3 search algorithm. In the first step of this search, a list is populated of forward and reverse primer candidates. Next, the tool eliminates any candidates that do not comply with the initial design parameters, such as primer length, CG content, or melting temperature (Tm). Following this step S*inglePlexPCR* selects candidate primer pairs by producing virtual PCR products that satisfy the input design parameters. More complex computational evaluations, such as the detection of potential formation of primer-dimer and hairpin structures, are then performed. Lastly, *SinglePlexPCR* stores and sorts all suitable primer pairs based on a penalty score and the search ends as soon as a predefined number of accepted PCR primer pairs are found. This algorithm provides an exhaustive search of all possible primer combinations and guarantees that all suitable pairs of primers will be identified if they exist. In addition, users can continue interacting with the primer search to find more results without the need to restart the search from the beginning.

- ***MultiPlexPCR*** uses an efficient greedy strategy and an optimized search process to find primers utilizing complex indexes. Since finding multiplex primers is a complex process requiring excessive computational resources, the search strategy of *MultiPlexPCR* for multiplex search differs from that for *SinglePlexPCR* searches. This is primarily because the number of possible primer combinations grows exponentially with the size of the input and so the storage and checking of primers products that will not yield a suitable multiplex set is a waste of computational resources. Thus, to ensure that only valid potential multiplex primer sets are stored and validated, *MultiplexPCR* uses a complex index to access and store the potential set of candidate primers based on an approximation of the quality of the primer sets using two index keys: PCR product criteria (product length or Tm for conventional or Real time PCR, respectively). The index provides an optimized complex data structure to store and evaluate only those primers that populate the list of suitable multiplex primer sets. This significantly reduces computational time and memory. Nonetheless, it is worth noting that the process of building the index still consumes a lot of memory, particularly when analysing large sequences and large numbers of targets. Thus, a more restricted design parameter set, such as primer length range and PCR product or other design criteria, is recommended as less candidate primers will be stored in the index and this can significantly reduce index memory usage. To achieve this, *MultiplexPCR* is powered by a new optimized search process that starts from building a complex index over potential multiplex primer sets (outlined in Algorithm 3.1) that subsequently perform a brute force specific search using a greedy strategy that uses the complex potential multiplex sets index (outlined in Algorithm 3.2). In this way, *MultiplexPCR* first finds a list of forward and reverse candidate primers in a similar fashion to the *SinglePlexPCR* search but also includes additional criteria to ensure that designed primers are specific to their target and do not bind elsewhere. Should the primers bind to multiple targets, they are then marked as specific shared primers. As a result, a shared forward primer that perfectly binds to two different targets can be used in combination with two other specific reverse primers yielding two distinguishable PCR products for each target sequence.

Algorithm 3.1 Building Potential Multiplex Set Index.

```
Input : seqList : a list of n target sequences
        pArgs   : Design Parameters
Output: pIndex  : potential Multiplex Set Index
Result: Potential Multiplex Set Index

/* Potential Multiplex Set Index provide a fast and optimized complex    */
/* data structure to store and evaluate only primers that contribute to  */
/* acceptable multiplex primer set.                                      */
foreach targetSeq in seqList do
    /* Basic Primer3 search                                              */
    Populate forward/reverse primer list for targetSeq;
    Sort primers lists;
end
foreach primer in forward/reverse list do
    Check primer specificity;
    /*Specific primers bind only to any target sequences and do          */
    /* not bind to any sequences in not target library if provided       */

    Categorize/groups primers that bind to multiple target;
    /* Primers could bind to multiple targets, this way it can be        */
    /* used to minimize the number of primers as long any                */
    /* combination produces distinguishable PCR products.                */
end
pIndex = Initialize an empty potential Multiplex Set Index;
foreach possible pcrProduct formed by forward/reverse pairs do
    /* Check acceptable PCR product criteria (Length, Tm) */
    if pcrProduct is acceptable product then
        Insert pcrProduct into pIndex ;
    end
    /* pIndex is built with two keys :                                   */
    /* PCR product criteria and multiplex set score                      */
    /* Set score is an average score of all primers pairs                */
    /* Score is approximated as it is calculated before evaluating primers */
    /* to avoid heavy computational if not needed                        */
end
return pIndex;
```

Algorithm 3.2 Specific/Multiplex Primers search.

```
Input : seqList : a list of n target sequences
        pArgs   : Design Parameters
        pIndex  : potential Multiplex Set Index (Returned by Algorithm 3.1)
Output: List of candidate multiplex primers set
Result: Candidate Specific multiplex primers set

mSets = Initialize an empty list for candidate multiplex set ;
while TRUE do
    potentialSet = Pull a potential multiplex set from pIndex with the best score;
      foreach pcrProduct in potentialSet do
        Evaluate Forward/Reverse primers in pcrProduct;
        if any does not satisfy any criteria in pArgs then
            Ignore potentialSet;
            Update pIndex;
            /* Index update will invalidate any set contains any        */
            /* of the invalidated primers                               */
            continue;
        end
     end
    Evaluate primers;
    /* Evaluate all primers across different tagert to check for        */
    /* any primer dimer formations                                      */
    if potentialSet satisfy all criteria provided by pArgs then
        Add potentialSet to mSets;
    if mSets have enough solutions or max number of iterations reached then
        break;
    end
    if pIndex has not more candidates then
        /* Search exhausted no more possible solutions                  */
        break;
    end
end
return  mSets;
```

3.9.5 Using GTF or GFF files to mine sequences for specific annotated contents

The GTF/GFF viewer allows for the mining of assemblies and the extraction of information, such as exons, promoters, or gene families. The GTF/GFF viewer is accessible by double-clicking on a GTF or

GFF file placed in the directory browser or through the tab “Annotation” in the browser menu. Once the GTF or the GFF file has been uploaded, the user will see the visualized data as an annotated grid of rows and columns. The different tasks provided by the viewer can be executed via mouse (as detailed in Figure 3.4) or via menu (as shown in Figure 3.5).

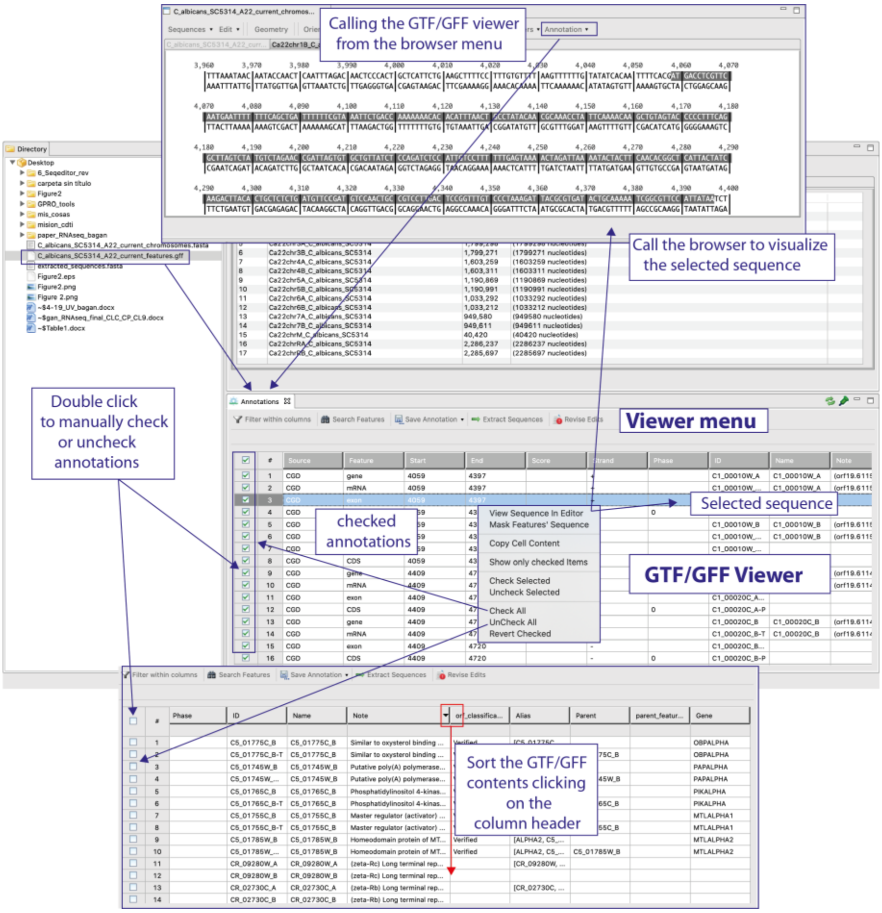
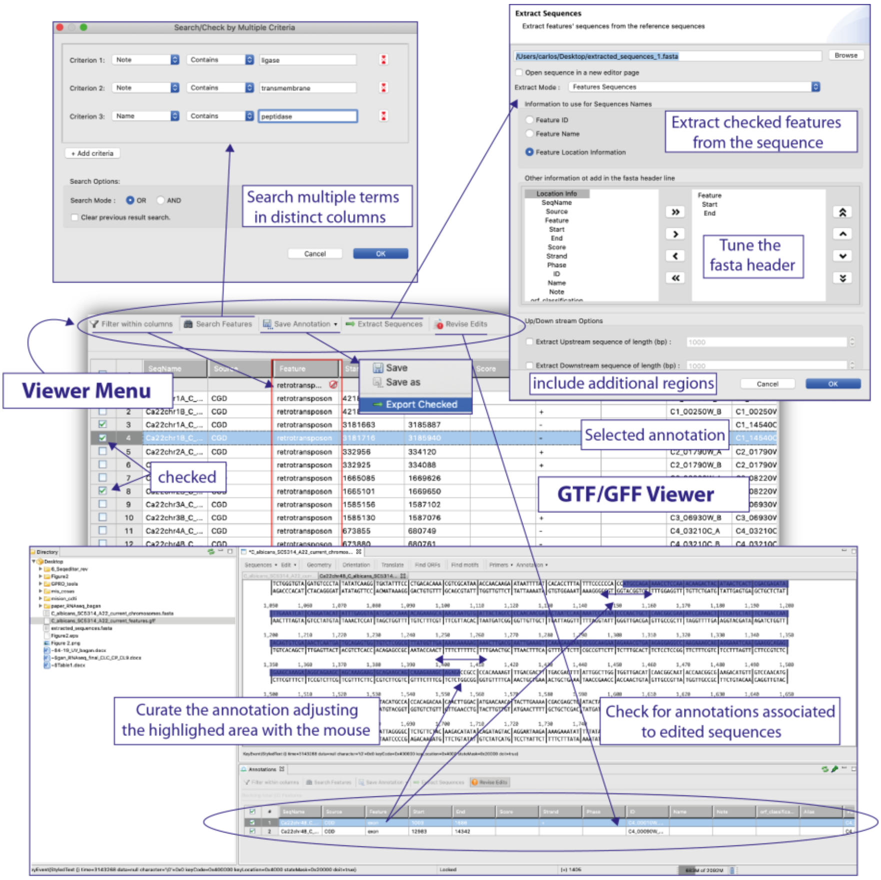


Figure 3.4 GTF/GFF viewer and the different options for mouse-dependent tasks.

As shown in Figure 3.4, annotations can be manually selected or deselected by clicking twice to check/uncheck rows or by right-clicking anywhere to call a context menu to provide distinct options for the visualization of specific features in the browser. By clicking one or two times on the column headers of the viewer, users can sort annotation file contents. Users can also select the texts of rows and from the column cells shown in the viewer using the mouse.



Search/Check by Multiple Criteria

Criterion 1: Note Contains ligase

Criterion 2: Note Contains transmembrane

Criterion 3: Name Contains peptidase

Search Options:

Search Mode : ● OR ○ AND

□ Clear previous result search.

Extract Sequences

Extract features' sequences from the reference sequences

/Users/carlos/Desktop/extracted_sequences_1.fasta

□ Open sequence in a new editor page

Extract Mode : Features Sequences

Information to use for Sequences Names

○ Feature ID

○ Feature Name

● Feature Location Information

Other information ot add in the fasta header line

Up/Down stream Options

□ Extract Upstream sequence of length (bp) : 1000

□ Extract Downstream sequence of length (bp) : 1000

		SeqName	Source	Feature	Start	End	Score	Strand			
				retrotransp...							
		1A_C_...	CGD	retrotransposon	421						
□	2	Ca22chr1B_C_...	CGD	retrotransposon	421			+		C1_00250W_B	C1_00250V
☑	3	Ca22chr1A_C_...	CGD	retrotransposon	3181663	3185887		-			C1_14540C
☑	4	Ca22chr1B_C_...	CGD	retrotransposon	3181716	3185940		-			C1_14540C
□	5	Ca22chr2A_C_...	CGD	retrotransposon	332956	334120		+		C2_01790W_A	C2_01790V
□	6			retrotransposon	332925	334088		+		C2_01790W_B	C2_01790V
□	7			retrotransposon	1665085	1669626					08220V
☑	8			retrotransposon	1665101	1669650					08220V
□	9	Ca22chr3A_C_...	CGD	retrotransposon	1585156	1587102					06930V
□	10	Ca22chr3B_C_...	CGD	retrotransposon	1585130	1587076		+		C3_06930W_B	C3_06930V
□	11	Ca22chr4A_C_...	CGD	retrotransposon	673855	680749		-		C4_03210C_A	C4_03210C
□	12	Ca22chr4B_C	CGD	retrotransposon	673880	680761		-		C4_03210C_B	C4_03210C

Figure 3.5 Task options provided by the menu of the GTF/GFF viewer.

As shown in Figure 3.5 , the tab “Filter within columns” enabled for a key word to be used to filter annotations in a particular column, showing only those that match the word. The tab “Search features” gives access to a context dialog allowing specification of one or more key words to search and check a subset of annotations matching these criteria (the options “or” and “and” can be used to improve the search). “Save Annotations” enables the saving of any edit or change in the GTF/GFF or to export only the checked annotations in a new GTF or GFF file. “Extract Sequences” calls another context dialog that permits the extraction of sequence features indicated as checked in the viewer. The dialog offers additional exporting options to name the fasta headers of exported sequences or for exporting the sequences with upstream and downstream nucleotide extensions of a user-defined size. Finally, “Revise edits” allows editing of the GTF/GFF file to correct or curate the annotation of any sequence if it has been previously edited with the browser. For example, if a user opens a sequence file and the associated GTF or GFF with the sequence browser and the GTF/GFF viewer, the user can edit the sequence in the browser. To update the GTF/GFF file according to this change the user only needs to click on the tab “Revise edits”. In doing so, the viewer detects and shows in the GTF/GFF viewer the annotations of those sequences that have been edited in the browser. Then, when clicking on the row of the edited sequence, the browser is called again, and the region affected by the edit is highlighted in it. Finally, the user can use the mouse to manually adjust the highlight of the edited region (for example an exon) by dragging the highlight until reaching the correct coordinate of that feature. After this action, the coordinates of that feature are corrected in the GTF/GFF viewer according to the final highlight stated in the sequence browser. Then the user only needs to save the new GTF or GFF file using the options provided by the “Save Annotation” tab.

3.9.6 Comparing SeqEditor with other tools.

We compared SeqEditor respect to the current state-of-the-art sequence analysis tool including the former TIME editor and three other GUI-based tools (GeneRunner, Sequencher and Geneious). It is worth clarifying that SeqEditor is only directly comparable with Sequencher and Geneious respecting the sequence editor utilities of these two applications albeit Sequencher and Geneious implement both a wide variety of tools for quality processing and analysis of next generation sequencing data and comparative analysis (including tools to create multiple alignments and inferring phylogenies) not offered by SeqEditor, which does not offer these tools because it has been designed to be a sequence editor while most of these functionalities will be provided in the context of three other applications of the GPRO suite providing access to distinct pipelines for analysis of RNASeq, DeNovoSeq and VariantSeq data. The features and properties of these applications will be discussed in a forthcoming publication we are preparing for introducing these three applications. Table 3.1 below summarizes the main features of SeqEditor, TIME, GeneRunner, Sequencher and Geneious.

Table 3.1 SeqEditor versus other applications.

	Features	Operability
SeqEditor	Sequence analysis of DNA, RNA and proteins • Sequence Brower/Editor • Fasta task manager • GTF/GFF viewer • Singleplex primer design • Multiplex primer design • Multiplex primer pooling	• Manual Available • Free to use • Runs on Windows, MacOS and Linux • Graphically browse scaffolds of up to 300 mega-bases. • Simultaneously processes multiple fasta files like reference databases, genomes, transcriptomes • Can use GTF/GFF files to mine contents, edit and/or personalize reference datasets.

	Features	Operability
TIME	Sequence analysis of DNA, RNA and proteins • Sequence Brower/Editor	• Manual Available • Free to use • Runs on Windows, MaCOS (Catalina not supported) and Linux • Graphically browse scaffolds of up to 25 mega-bases.
Sequencher	Sequence analysis of DNA, RNA and proteins and comparative analysis mainly based on Sanger and NGS data • Sequence Browser/Editor • Tools for quality analysis and pre-processing of sanger and NGS data • De novo Assembler • Mapper of resequencing data • Tools for creating and managing multiple alignments including reconstruction of consensus sequences • Graphical tasks to plot restriction and ORF maps • Other tools for statistical and downstream analysis	• Manual Available • Commercial application available through a pay for use license. • Runs on Windows, MaCOS (Catalina not ye supported) and Linux • Graphically browse sequences of up to 128 mega-bases • Performs de novo assemblies • Processes resequencing data in RNA-Seq and Variant-Seq experiments • Protocols for calling and annotation of SNP/indel variants • Protocols for differential expression analysis • Protocols for de novo assembly and • Protocols for comparative analysis and annotation

	Features	Operability
Gene Runner	DNA and RNA sequence analysis - Sequence Browser/Editor - Graphical tools to plot restriction and ORF maps - Tools for Site-directed mutagenesis, - Singleplex primer design.	- No manual available - Former commercial application, now freely available. - Runs on Windows operative systems only. - Graphically browse sequences of up to 32 mega-bases
Geneious	Sequence analysis of DNA, RNA and proteins and comparative analysis mainly based on Sanger and NGS data - Sequence Browser/Editor - Tools for quality analysis and pre-processing of sanger and NGS data - De novo Assembler - Mapper of resequencing data - Tools for creating and managing multiple alignments including reconstruction of consensus sequences - Graphical tasks to plot restriction and ORF maps - Other tools for statistical and downstream analysis	- Manual Available - Commercial application - Runs on Windows, MaCOS and Linux. - Able to browse sequences of less than to 32 mega-bases. - Performs de novo assemblies - Processes resequencing data in RNA-Seq and Variant-Seq experiments - Protocols for calling and annotation of SNP/indel variants - Protocols for differential expression analysis - Protocols for de novo assembly and - Protocols for comparative analysis and annotation

With respect to graphical browsing capabilities, SeqEditor is significantly more efficient and faster in browsing tasks than all other tools tested. It is worth to clarify first that the capacity of graphic screens usually relies on the RAM power of the specific hardware.

For example, in a PC with 25 Gigabytes of RAM SeqEditor is able to manage sequences of up to 300 mega-bases without significant slowdown (remembering that the largest human chromosomes are around 250 mega-bases) while GeneRunner and Geneious only managed sequences up to 32 mega-bases, TIME up to 25 mega-bases and Sequencher up to 128 mega-bases but with delays in reading the sequence. The advantage of SeqEditor in graphical power respect to the other tools lies in the following aspects of the tools. First, the sequence browser of SeqEditor is optimized to manage very large sequences; second SeqEditor permits the user to edit the RAM assignation depending on the analysis and as convenience (details for configuring the RAM assigned to SeqEditor are provided in the SeqEditor' manual at https://gpro.biotechvana.com/tool/seqeditor/manual). In sharp contrast, GeneRunner, Geneious and Sequencher implement other graphical functions to read chromatograms of Sanger sequencing that permit the users to evaluate the sanger quality of the bases as well as tools for plotting restriction sites, motifs and ORFs. These graphic functions are not yet supported by SeqEditor but we are committed to implement them in future updates.

Regarding management and processing tasks of sequence files and databases, we have previously noted that the fasta task manager of SeqEditor permits the user to work simultaneously with multiple fasta files with a comparable efficiency to CLI tools. In contrast, TIME, GeneRunner, Sequencher and Geneious can only work with a single sequence at a time. In addition, the GTF/GFF viewer of SeqEditor lets the user mine contents from reference sequences or even manage the complexity of GTF or GFF files to edit (or to extract subsets) reference sequences and their associated GTF, GFF or files. The GTF/GFF viewer has been tested and performed efficiently with distinct GTF and GFF files including those provided by the UCSC

(Kent *et al.* 2002), NCBI (Sayers *et al.* 2020), Ensembl (Cunningham *et al.* 2019), or the Candida Genome Database (Skrzypek *et al.* 2017) or even those formats created by AUGUSTUS 3.3 (Stanke *et al.* 2008) like the GTF of to the recently published *Sparus aurata* genome (Pérez-Sánchez *et al.* 2019). To our knowledge, the GTF/GFF viewer has no equivalent in any of the four applications or any other currently available tool for sequence analysis. Users interested in editing or preparing personalized datasets from reference genomes with GTF or GFF files normally need to manage online repositories like the UCSC genome browser or Ensembl or alternatively, use the command line (in cases of users with expertise on the Linux syntax). Therefore, the fasta task manager and the GTF/GFF viewer together make SeqEditor a suitable option for users that need to edit and manage reference genomes or transcriptomes with or without GTF/GFF files.

The SinglePlexPCR, MultiPlexPCR, and PrimerPooler tools for primer design of SeqEditor are based on three distinct interface adaptations of two cutting edge CLI tools: Primer3 and PrimerPooler. Interestingly, the tools for primer design utilized by Geneious and Sequencher are also based on an interface adaptation of Primer3, which is one of the most efficient tools for primer design together with Primer-BLAST (Ye *et al.* 2012). Understandably, this indicates that adaptation friendly-to-use interfaces to manage Primer3 is not a novel. Indeed, Primer3 has already been adapted as an online public web server (Untergasser *et al.* 2007). Nevertheless, it is worth highlighting that the tools for primer design used by GeneRunner, Geneious, Sequencher, and even those of the web server of Primer3 are all oriented to singleplex experiments. This would mean that users interested in designing primers for multiplex experiments must do so manually or to use other specific tools like the PrimerSuite (Lu *et al.* 2017) or Oli2go (Hendling *et al.* 2018). The difference of SeqEditor with respect to the other available applications for sequence analysis is thus the implementation of MultiPlexPCR - a tool that remains unique to SeqEditor and is the is the only application to our knowledge that adapts Primer3 to design multiplex oligos. MultiPlexPCR based on two algorithms (referred above as Algorithm 3.1 and Algorithm 3.2). This search strategy is specifically optimized to find target- and species-specific primers and has been satisfactorily validated by performing a comprehensive primer design test using five human yeast pathogens for which fast and accurate diagnostics is necessary (Consortium Opathy and Gabaldón 2019). The results obtained from this validation are provided (and can be thus reproduced) in the tutorial we prepared with case study examples and step-by-step indications about how to use SinglePlexPCR, MultiPlexPCR and PrimerPooler. This tutorial is accessible online at https://gpro.biotechvana.com/software/tutorials/candida_target_specific_2020. Finally, it is worth stressing the importance of target-specific primers, which are central for species identification in many microbiological processes or for determining antimicrobial susceptibility and load infection. In fungal research for example, target-specific primers are frequently used for the identification of a yeast pathogen that is responsible for a given infection. This is clinically relevant because, despite all currently available antifungal drugs, invasive fungal infections have a mortality rate of ≥ 50% (Brown *et al.* 2012). Indeed, Candida-related bloodstream infections have a mortality rate between 30%-60% (Hirano *et al.* 2015). As longer hospital stays and the need for multiple analyses that accounted for over $7.2 billion USD in the USA in 2017 (Benedict *et al.* 2019), there is a clear benefit of using SeqEditor for accurate diagnoses.

3.9.7 Tutorial for primer design with SeqEditor

Downloading SeqEditor and getting familiar

- SeqEditor can be freely downloaded at https://gpro.biotechvana.com/download
- For general use indications, please refer to the SeqEditor manual available at https://gpro.biotechvana.com/tool/seqeditor/manual

3.9.7.1 Case study: Primer Search for Singleplex PCR using the Primer3 implementation of SeqEditor.

Data requirements:

To reproduce the examples, the following data are required:

- **target_sequences.fasta**: this file is available in the data folder of Supplementary file 2 and contains rDNA sequences for 5 *Candida* species as summarized in Table 3.2. These target sequences belong to ribosomal gene domains such as 28s and Internal Transcribed Sequences (ITS). The presence of sufficient polymorphism within these loci allow the design of species-specific primers. Being a multicopy gene, one might expect higher PCR amplification when compared to a single copy gene, usually protein-coding genes such as Actin, Beta-tubulin, Elongation factor, RPBII, etc.

- **non_targets.fasta**: this file is available in the data folder of Supplementary file 2 and contains sequences for other species which are closely related to the target species as summarized in Table 3.3. This was created by blasting the target sequences in the NCBI database and selecting those subjects with the highest similarity available.

Table 3.2 Target species and sequences.

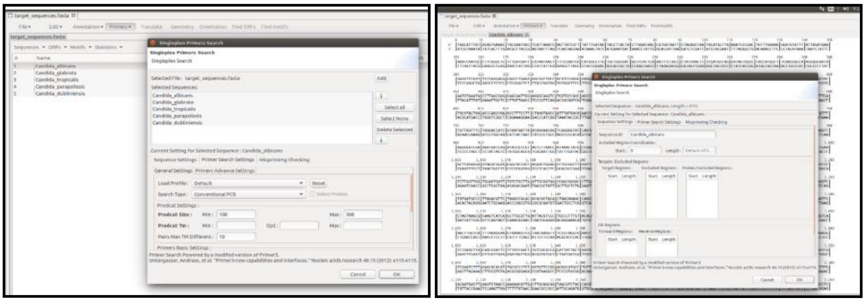
Species	Sequence Name	NCBI Accessions
Candida albicans	Candida_albicans_CBS_1949_NS1_LR5_53910_Lodderomyces_clade	CP025165.1 CP025157.1
Candida glabrata	Candida_glabrata_CBS_138_NS1_LR5_58190_Saccharomycetaceaea	CR380958.2 MK394140.1
Candida tropicalis	Candida_tropicalis_CBS_8072_NS1_LR5_53938_Lodderomyces_clade	MK394119.1
Candida parapsilosis	Candida_parapsilosis_CBS_604_NS1_LR5_81205_Lodderomyces_clade	HE605209.1
Candida dubliniensis	Candida_dubliniensis_CBS_7987_NS1_LR5_53918_Lodderomyces_clade	FM992695.1

Table 3.3 Non target species and sequences.

Species	NCBI Accessions
Homo sapiens	NG_054875.1; MF164260.1
Saccharomyces cerevisiae	KJ806314.1
Debaryomyces hansenii strain ATCC	GQ458041.1
Pichia norvegensis	NG_063278.1
Pichia norvegensis	AY497674.1
Pichia kudriavzevii	CP028774.1
Candida lusitaniae	M55526.1
Clavispora lusitaniae	JQ698900.1; KY106935.1; KY106931.1
Candida haemulonis	NG_063413.1; JN941107.1
Candida rugosa	KT336717.1; AB013502.1
Candida inconspicua	EF152417.1; KY106513.1
Diutina rugosa	KY563206.1
Aspergillus fumigatus	MF379664.1; KJ809565.1

Primer search for target region detection:

The objective of this example is to show how to use SeqEditor for the design of those singlePlex PCR primers, which will be able to detect the target DNA sequences. To do this, first open the target_sequences.fasta file with SeqEditor and then go to the Singleplex path via the Editor Top menu (Figure 3.6).



#	Name
1	Candida_albicans
2	Candida_glabrata
3	Candida_tropicalis
4	Candida_parapsilosis
5	Candida_dubliniensis

Singleplex Primers Search

Singleplex Search

Selected File : target_sequences.fasta

Selected Sequences:

Candida_albicans
Candida_glabrata
Candida_tropicalis
Candida_parapsilosis
Candida_dubliniensis

Current Setting for Selected Sequence : Candida_albicans

Sequence Settings | Primer Search Settings | Mispriming checking

General Settings | Primers Advance Settings

Load Profile : Default

Search Type : Conventional PCR

Prodcut Settings :

Prodcut Size : Min : 100 Max : 300

Prodcut Tm : Min : Opt : Max :

Pairs Max TM Different : 10

Primers Basic Settings

Primer Search Powered by a modified version of Primer3.
Untergasser, Andreas, et al. "Primer3-new capabilities and interfaces." Nucleic acids research 40.15 (2012): e115-e115.

Singleplex Primers Search

Singleplex Search

Selected Sequence : Candida_albicans, Length : 3115

Current Setting for Selected Sequence : Candida_albicans

Sequence Settings | Primer Search Settings | Mispriming Checking

Sequence ID : Candida_albicans

Included Region Coordinates :

Start : 0 Length : Default All S...

Targets Excluded Regions

Target Regions : Start Length

Excluded Regions : Start Length

Probes Excluded Regions : Start Length

Ok Regions

Forward Regions : Start Length

Reverse Regions : Start Length

Primer Search Powered by a modified version of Primer3.
Untergasser, Andreas, et al. "Primer3-new capabilities and interfaces." Nucleic acids research 40.15 (2012): e115-e115.

Figure 3.6 Singleplex Primer Search Dialog. Left batch mode. Right Single Mode.

[**Primers -> Singleplex Primers**].

There are two different ways to run the Singleplex Primers design tool (***Batch/Single run mode)*** **:**

- To run the search in batch mode (*i.e*. for all sequences in the file), select Singleplex Primers while the fasta sequences file summary view is active.
- If a single sequence fasta is opened in the Sequence Browser, you can still run Singleplex Primers. However, please keep in mind the browser will operate only on the sequence that has been opened.

Once you have selected the correspondent primer search mode from the menu, a search dialog will be opened to configure the search parameters. For each input target sequence, you can edit the search parameters. There are three main tabs for the editing of these search parameters:

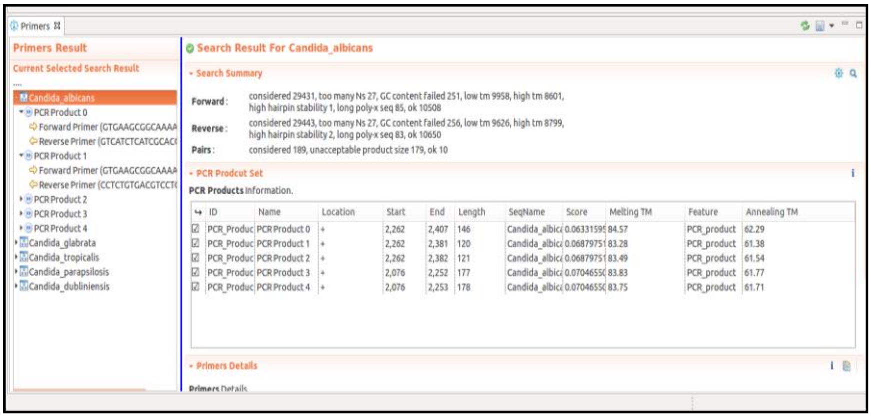
- Sequence Setting: target sequence related info including:
  - A custom ID for the search result.
  - **Included region**: A subregion in the given input sequence inside which the search for primers is performed.
    - All region inputs are defined by two parameters:

        - **Start**: index of the first base of the region in the target sequence.
        - **Length**: total length of the subregion starting at the **Start** parameter.
    - Targets and excluded regions:
        - **Target regions**: subregions of the input sequence within which a suitable primer pair must flank at least one of them.
        - **Excluded Regions**: subregions within the input sequence excluded from the primer search (*i.e*. rejecting all primers or probes that overlap any of those regions).
        - **Probes Excluded Regions**: same as described for the excluded regions but only applying for probes selection.
    - **OK Regions**: possible sub regions in the sequence the search is directed to when designing primers (*i.e*. constrained search regions). Each entry in this configuration includes two sub-regions, including one for the forward primers and one for the reverse primers. It is also possible to constrain a search region for one of the primers only (forward or reverse) and not the other.
- Primer Search Settings:
    - General Settings: This section includes parameters such as primer length, Tm and GC content as well as product size. PCR type and load can also be changed by selecting one of the predefined PCR configurations (Profiles).
    - Primers Advanced Settings: These include thermodynamic calculations parameters as well as other primers properties and score calculation settings.
    - Probes Advanced Settings: this section will be shown when probe selection is enabled. These are similar to the Primers Advanced Settings, however, to be applied for probes only.
- Mispriming Checking: provides the search with a library of sequences to mispriming.

Once all parameters are set, click on the OK button. The primer search will be launched in the background and you will be able to track it or cancel it from the progress view.

Primers Result view:-

The bottom left side panel of Figure 3.7 shows a tree-like view of the primer search result sorted by the custom ID/sequence name that is provided in the search dialog. Candidate PCR products and primer pairs are listed under each group. The bottom right side panel of the figure displays the result in different sections, including a search summary at the top, a PCR product section containing a table of all candidate PCR products and their properties, and a table of all primers/probes of all products and their properties at the bottom. This window also includes a “save the result to file” option where the user can export the results in either CSV, Fasta or GFF file format.



Forward : considered 29431, too many Ns 27, GC content failed 251, low tm 9958, high tm 8601, high hairpin stability 1, long poly-x seq 85, ok 10508

Reverse : considered 29443, too many Ns 27, GC content failed 256, low tm 9626, high tm 8799, high hairpin stability 2, long poly-x seq 83, ok 10650

Pairs : considered 189, unacceptable product size 179, ok 10

PCR Products Information.

↪	ID	Name	Location	Start	End	Length	SeqName	Score	Melting TM	Feature	Annealing TM
☑	PCR_Produc	PCR Product 0	+	2,262	2,407	146	Candida_albic	0.0633159	84.57	PCR_product	62.29
☑	PCR_Produc	PCR Product 1	+	2,262	2,381	120	Candida_albic	0.0687975	83.28	PCR_product	61.38
☑	PCR_Produc	PCR Product 2	+	2,262	2,382	121	Candida_albic	0.0687975	83.49	PCR_product	61.54
☑	PCR_Produc	PCR Product 3	+	2,076	2,252	177	Candida_albic	0.0704655	83.83	PCR_product	61.77
☑	PCR_Produc	PCR Product 4	+	2,076	2,253	178	Candida_albic	0.0704655	83.75	PCR_product	61.71

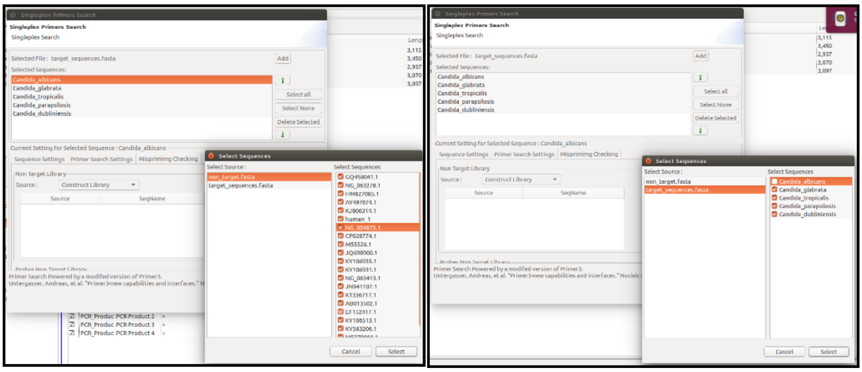
Primers Details

Figure 3.7 Primers Result View.

Species-specific primer search:-

The aim of this example is to describe the process that SeqEditor follows to search for species-specific PCR primers. That is, those primers that will allow the amplification of DNA fragments via PCR experiments in the target species but not in any others. To do so, SeqEditor allows the user to include a mispriming library containing sequences belonging to the non-target species in the search. This process is shown in Figure 3.8 as follows:

Open both “***non_targets.fasta***” and “***target_sequences.fasta***” fasta files with SeqEditor. From the ***target_sequences.fasta*** file, select [**Primers -> Singleplex Primers**] in the menu toolbar. Next, adjust the input parameters in the search dialog as described in previous example steps. From there, select any of the input sequences to select their parameters (*e.g*. select *C. albicans*) In the “Mispriming checking” tab -> non target library section, choose “construct library” in the source field. From the table context menu select “add new”. In the selection dialog, select “all sequences” in the non-target fasta files. To avoid including any of the other target species, select all sequences in the same file except the input target sequence (*e.g. C. albicans* in this case). Repeat the same steps again for each target species.



Selected File: target_sequences.fasta

Selected Sequences:

Candida_albicans
Candida_glabrata
Candida_tropicalis
Candida_parapsilosis
Candida_dubliniensis

Current Setting for Selected Sequence : Candida_albicans

Non Target Library

Source: Construct Library

Source	SeqName

Primer Search Powered by a modifed version of Primer3.
Untergasser, Andreas, et al. "Primer3-new capabilities and interfaces." N

Select Source: non_target.fasta

Select Sequences:

☑ GQ458041.1
☑ NG_063278.1
☑ HM627085.1
☑ AY497674.1
☑ KJ806314.1
☑ human_1
☑ NG_054875.1
☑ CP028774.1
☑ M55526.1
☑ JQ698900.1
☑ KY106935.1
☑ KY106931.1
☑ NG_063413.1
☑ JN941107.1
☑ KT336717.1
☑ AB013502.1
☑ EF152417.1
☑ KY106513.1
☑ KY563206.1

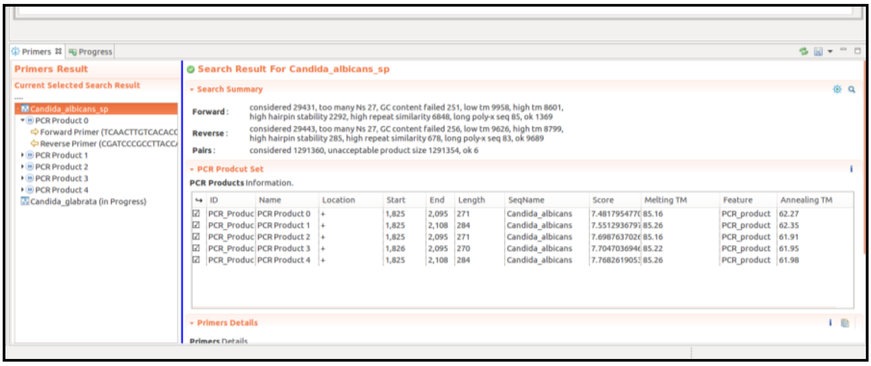
Select Source: target_sequences.fasta

Select Sequences:

□ Candida_albicans
☑ Candida_glabrata
☑ Candida_tropicalis
☑ Candida_parapsilosis
☑ Candida_dubliniensis

Figure 3.8 Adding a mispriming library to the the primer search parameters. (Left) Adding all sequences in the none_targets.fasta file. (Right) Adding all target sequences except Candida albicans as it is the currently selected sequence.

After configuring all search parameters, click on the “OK” button. The Primers result view as depicted in Figure 3.9 will be shown upon completion. Note that the results obtained in this case will be different from those in the previous example since the search summary for *C. albicans* that the tool considered in this search was **1291360** different primer pair combinations *versus* only **189** combinations obtained in the previous example shown in Figure 3.7. This is because most of the forward and reverse primers have a high similarity to one or more of the non-targets' sequences.


Primers
Progress
Primers Result
Current Selected Search Result
Candida_albicans_sp
PCR Product 0
Forward Primer (TCAACTTGTCACACC
Reverse Primer (CGATCCCGCCTTACC
PCR Product 1
PCR Product 2
PCR Product 3
PCR Product 4
Candida_glabrata (in Progress)
Search Result For Candida_albicans_sp
Search Summary
PCR Prodcut Set
Primers Details

Forward : considered 29431, too many Ns 27, GC content failed 251, low tm 9958, high tm 8601, high hairpin stability 2292, high repeat similarity 6848, long poly-x seq 85, ok 1369

Reverse : considered 29443, too many Ns 27, GC content failed 256, low tm 9626, high tm 8799, high hairpin stability 285, high repeat similarity 678, long poly-x seq 83, ok 9689

Pairs : considered 1291360, unacceptable product size 1291354, ok 6

PCR Products Information.

↪	ID	Name	Location	Start	End	Length	SeqName	Score	Melting TM	Feature	Annealing TM
☑	PCR_Produc	PCR Product 0	+	1,825	2,095	271	Candida_albicans	7.4817954770	85.16	PCR_product	62.27
☑	PCR_Produc	PCR Product 1	+	1,825	2,108	284	Candida_albicans	7.5512936797	85.26	PCR_product	62.35
☑	PCR_Produc	PCR Product 2	+	1,825	2,095	271	Candida_albicans	7.6987637026	85.16	PCR_product	61.91
☑	PCR_Produc	PCR Product 3	+	1,826	2,095	270	Candida_albicans	7.7047036946	85.22	PCR_product	61.95
☑	PCR_Produc	PCR Product 4	+	1,825	2,108	284	Candida_albicans	7.7682619053	85.26	PCR_product	61.98

Primers Details

Figure 3.9 Primers Result View of the second example with mispriming checking.

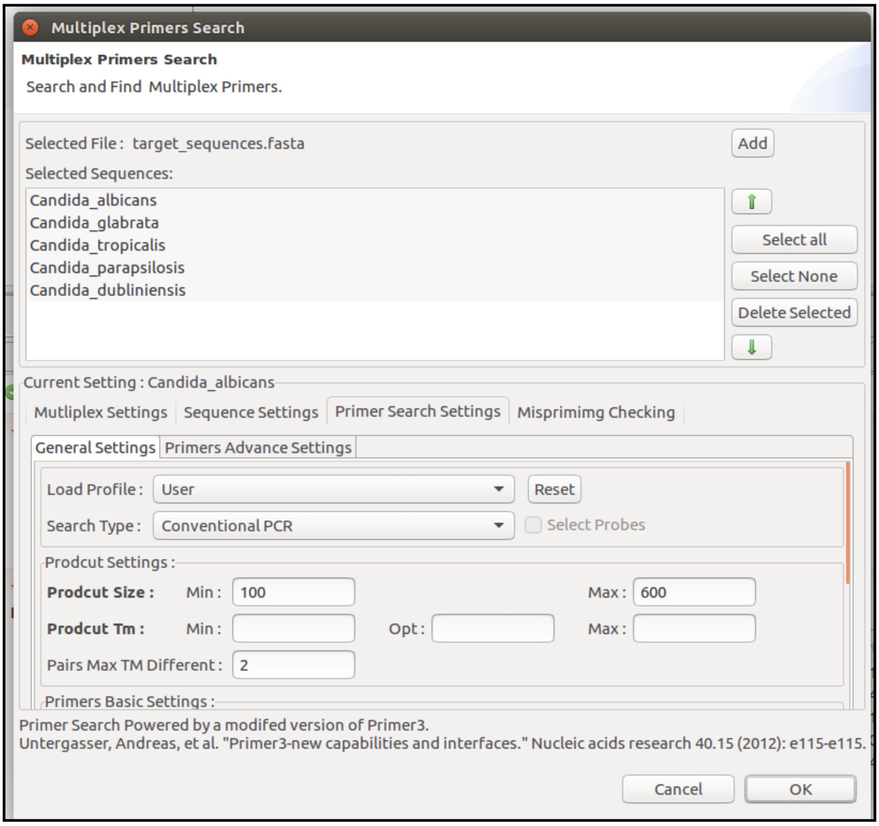
3.9.7.2 Case study: Primer Search for Multiplex PCR.

The example shown above described the design of species-specific primers. These primers, however, could not be used in a multiplex setting as some of the PCR products share the same length, making them indistinguishable from each other. One way to adopt this singleplex search to multiplex PCR is to change the desired product size range for each species in the search to have both distinguishable product sizes in the PCR for each targeted to predict any dimer formation between primers. This might require running the search multiple times until a suitable set of primers is found. Taking this into consideration, SeqEditor provides a Multiplex search option that runs this process automatically yielding a valid primer set for either conventional PCR or qPCR.

To run this test, you will need the two files used in the previous example.

- In the ***target_sequences.fasta*** file tab, select [**Primers -> Multiplex Primers**] from the top toolbar menu. Adjust the search parameters in the search dialog shown in Figure 3.10. In this case, all input parameters in both “Primer Search Settings” and “Mispriming Checking” will remain the same for all input sequences, however the “Sequence Settings” parameters can be different for each target as those parameters will control the search regions in the provided sequence.
- If Conventional PCR is selected, adjust the product size so that the final PCR product can be distinguishable from the others. In this example, a range of 100-600bps would be suitable for 5 targets. However, broadening the size range will allow the tool to find a higher number of suitable solutions.
- If intercalating dye-based qPCR is selected (*e.g*. SYBR Green), the preferred PCR product length should range from70-150bps up to a maximum of 200bps. The most important criterium here will be the product Tm, for which the tool will try to return PCR products with Tm differences of at least 1ºC.

For Mispriming checking -> non-target library section, choose “construct library” in the source field, then select “add new” from the table context menu. In the selection dialog, select all sequences included in the non-target fasta files. Do not add any sequence from our targets as mispriming check against the target sequences will be performed in the multiplex search by default.



Multiplex Primers Search

Search and Find Multiplex Primers.

Selected File : target_sequences.fasta

Selected Sequences:

Candida_albicans
Candida_glabrata
Candida_tropicalis
Candida_parapsilosis
Candida_dubliniensis

Current Setting : Candida_albicans

Load Profile : User

Search Type : Conventional PCR

☐ Select Probes

Prodcut Settings :

Prodcut Size : Min : 100 Max : 600

Prodcut Tm : Min : Opt : Max :

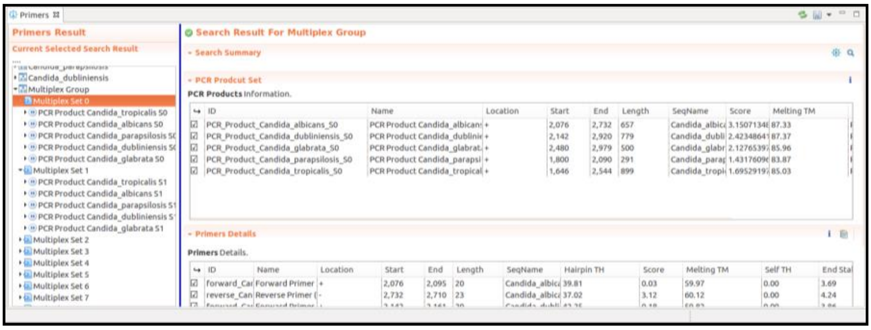
Pairs Max TM Different : 2

Primers Basic Settings :

Primer Search Powered by a modifed version of Primer3.
Untergasser, Andreas, et al. "Primer3-new capabilities and interfaces." Nucleic acids research 40.15 (2012): e115-e115.

Figure 3.10 Multiplex Primer search dialog.

After configuring all search parameters, click on the “OK” button and wait for the search to complete. Once the search is finished, the Primer result view (Figure 3.11) will be shown. The result view in this case will be different from the Singleplex result. The left side panel shows the result primers grouped by candidate multiplex sets (valid Multiplex PCR primers). The right-hand side panel of the results view will display the same information shown in the singleplex case. From this point, the user can navigate the search results by selecting any of the sets in the tree view to show only those primers included in the selected set.



Primers

Primers Result

Current Selected Search Result

- Candida_dubliniensis
- Multiplex Group
  - Multiplex Set 0
    - PCR Product Candida_tropicalis S0
    - PCR Product Candida_albicans S0
    - PCR Product Candida_parapsilosis S0
    - PCR Product Candida_dubliniensis S0
    - PCR Product Candida_glabrata S0
  - Multiplex Set 1
    - PCR Product Candida_tropicalis S1
    - PCR Product Candida_albicans S1
    - PCR Product Candida_parapsilosis S1
    - PCR Product Candida_dubliniensis S1
    - PCR Product Candida_glabrata S1
  - Multiplex Set 2
  - Multiplex Set 3
  - Multiplex Set 4
  - Multiplex Set 5
  - Multiplex Set 6
  - Multiplex Set 7

Search Result For Multiplex Group

- Search Summary

- PCR Prodcut Set

PCR Products Information.

↪	ID	Name	Location	Start	End	Length	SeqName	Score	Melting TM
☑	PCR_Product_Candida_albicans_S0	PCR Product Candida_albican	+	2,076	2,732	657	Candida_albic	3.1507134	87.33
☑	PCR_Product_Candida_dubliniensis_S0	PCR Product Candida_dublini	+	2,142	2,920	779	Candida_dubli	2.4234864	87.37
☑	PCR_Product_Candida_glabrata_S0	PCR Product Candida_glabrat	+	2,480	2,979	500	Candida_glabr	2.1276539	85.96
☑	PCR_Product_Candida_parapsilosis_S0	PCR Product Candida_parapsi	+	1,800	2,090	291	Candida_para	1.4317609	83.87
☑	PCR_Product_Candida_tropicalis_S0	PCR Product Candida_tropical	+	1,646	2,544	899	Candida_tropi	1.6952919	85.03

- Primers Details

Primers Details.

↪	ID	Name	Location	Start	End	Length	SeqName	Hairpin TH	Score	Melting TM	Self TH	End Stal
☑	forward_Car	Forward Primer	+	2,076	2,095	20	Candida_albic	39.81	0.03	59.97	0.00	3.69
☑	reverse_Can	Reverse Primer (	-	2,732	2,710	23	Candida_albic	37.02	3.12	60.12	0.00	4.24

Figure 3.11 Primer Result View for a Multiplex PCR example.

Validating Multiplex primers sets:-

For the experimental validation of the primer set results obtained in the test described above, a conventional PCR experiment was performed using the multiplex primers set shown in Table 3.4.

Table 3.4 Multiplex primer set used in the validation PCR experiment.

Species name	Primer sequence	PCR product length (bps)
C. albicans	forward: CCAAAAACATTGCTTGCGGC	613
	reverse: CAGAGGCTATAACACACAGCAG	
C. glabrata	forward: CGACTCCACTTCAGAGCGG	290
	reverse: ACACTCCCAGGTCTTTGTCG	
C. tropicalis	forward : GAGCAATCCTACCGCCAGAG	363
	reverse: TGGTGGCCACTAGCAAAATAAG	
C. parapsilosis	forward: GGTAGGCCTTCTATATGGGGC	977
	reverse: GCCAACATCCTAGGCCGAA	
C. dubliniensis	forward: CACCACATGTGTTTTGTTCTGG	412
	reverse : CCAGAGACCGCCTTAGCAAT	

The correspondent gel electrophoresis is shown in Figure 3.12. The correspondent PCR experimental conditions were as follows:

- Gel electrophoresis conditions: 6µl of each PCR product were run in a 2% agarose gel in TAE buffer (40 mM Tris-Acetate, 1mM EDTA) for 50 min at 130V.
- PCR mix (per sample): 2.5U of Biotaq DNA polymerase (Bioline), 10X Polymerase reaction buffer, 1.5mM MgCl2, 0.2mM of dNTP mix, 5 pmol of each primer, 1 ng of genomic DNA, miliQ-H_2O up to 50µl of final reaction volume.
- PCR program:
    - 95 C for 5min
    - 95ºC for 30 sec (x35)
    - 60ºC for 30 sec
    - 72ºC for 1min
    - 72ºC for 8 min.

3.9.7.3 Case study: Splitting Sequencing primers to pools using the PrimerPooler implementation of SeqEditor.

The objective of this example is to show how to use SeqEditor to either split or distribute amplicons between multiple pools in order to reduce dimer formation.

Prerequisites to run the following example are as follows:

- Download the pooler_example.fasta file (available in the data folder included in this Supplementary File 1). This file contains a list of sequencing primers for the human genome (taken from (Brown *et al*. 2017).
- Download the human genome hg38.2bit available at http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.2bit . You may also download other genomes in 2bit as provided in the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html).

- Once you have the two files, open pooler_example.fasta in SeqEditor.

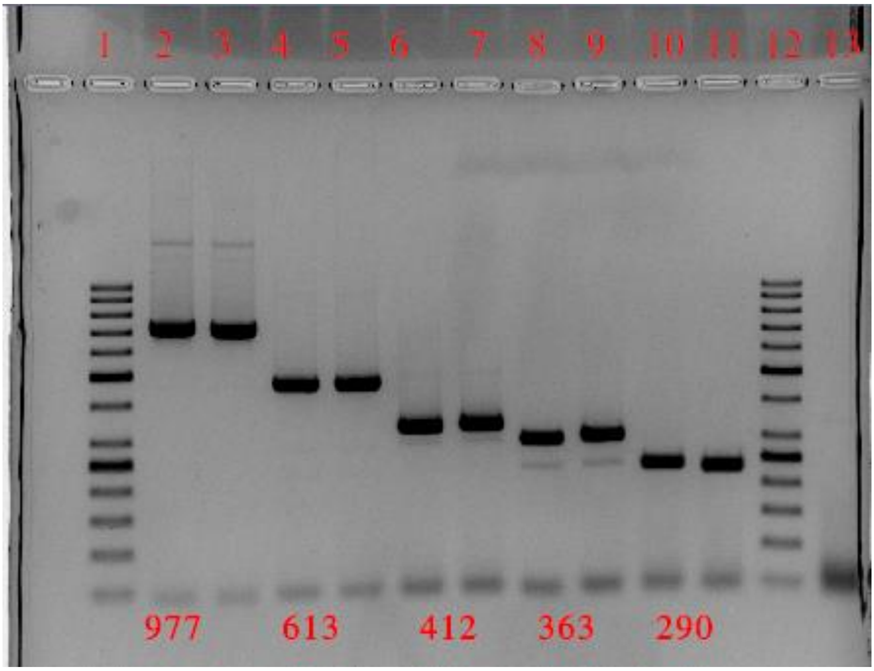


Figure 3.12 Gel electrophoresis of the PCR products amplified through the use of primer pairs designed with “Multiplex primer search” in SeqEditor. Lane 1. 50 bp DNA ladder (ThermoFisher). Lanes 2-3; *C. parapsilosis*, Lanes 4-5: *C. albicans*, Lanes 6-7: *C. dubliniensis*, Lanes 8-9: *C. tropicalis*, 10-11; *C. glabrata*, Lane 12 (Ladder) 13: Negative control (-genomic DNA).

Data input format:-

Please note that input files should be a fasta file containing primer sequences, and that each sequence name should end up in either F or R (standing for forward and reverse respectively).Primer pairs should all have the same name prefix (*e.g.* “primer”). Degenerate bases are allowed when using the usual DNA code letters.

Example of valid input data:

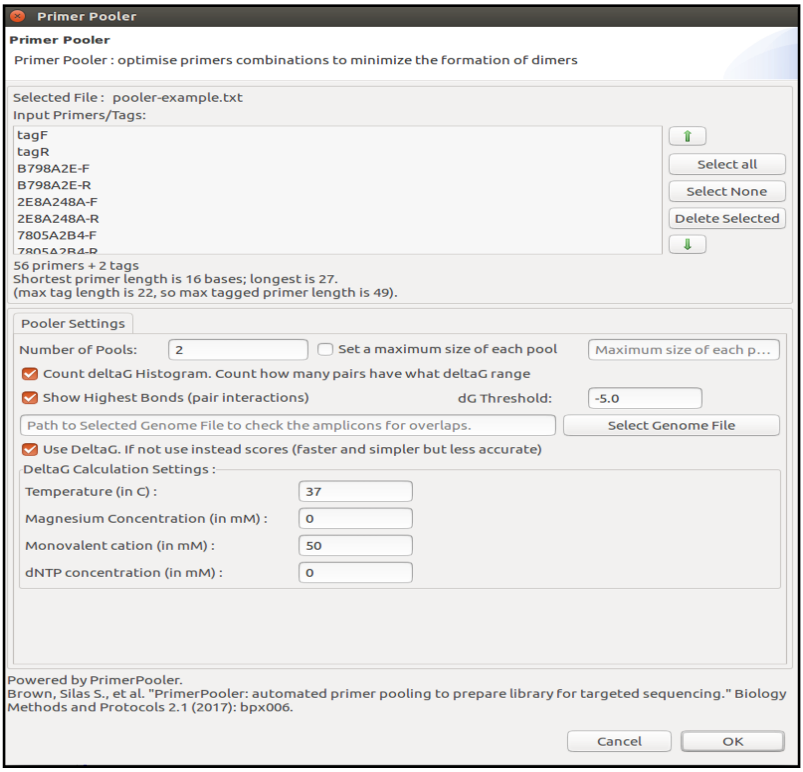
```
>primer1-F
CGCCGTCTTCCACCAACCA
>primer1-R
GGTAGGCGCTGCGGTT
>primer2-F
TCACAAAACACTTCATCTTTACTCAT
>primer2-R
CTCCAGTCCTCTCAGCCT
```

The “examples” folder contains a valid data input example denoted as “pooler.example.fasta” which contains primer sequences for the human genome. The user may also add tags to the primers using >tagF and >tagR. These primers will then be referred to as “tailed primers” and this process is known as barcoding). The primer tags can be changed part-way through the file). Should the user have either any Taq probes or other primers that do not yield amplicons, these can end with other letters (*e.g.* >probe1-P). Any set of names differing in only the last letter will be kept in the same pool. From the directory browser or File Menu open the example file “pooler.example.fasta”. Once open, select Primers -> Primer Pooler from the Editor Top menu. A Primer Pooler dialog will be displayed as shown in Figure 3.13. In the Input Primers/Tags List, show all input sequences and check that the input contains the required information.

From the PrimerPooler dialog you can the configure the following tool parameters:

- Number of pools: number of pools to divide the input into.
- Set Max size of pools: setting a maximum size of each pool can make the pools more even.
- Count deltaG/Score: Create histograms from the deltaG/Score of all pairwise interactions of all primers.

- Show Highest Interaction: shows a summary and the dimer structure formed from the pairwise interaction with the highest interaction.
    - Threshold: sets either a maximum or a minimum dG score for the display of the pairwise interaction.



Primer Pooler

Primer Pooler

Primer Pooler : optimise primers combinations to minimize the formation of dimers

Selected File : pooler-example.txt

Input Primers/Tags:

tagF
tagR
B798A2E-F
B798A2E-R
2E8A248A-F
2E8A248A-R
7805A2B4-F
7805A2B4-R

56 primers + 2 tags
Shortest primer length is 16 bases; longest is 27.
(max tag length is 22, so max tagged primer length is 49).

Pooler Settings

Number of Pools: 2 ☐ Set a maximum size of each pool Maximum size of each p...

☑ Count deltaG Histogram. Count how many pairs have what deltaG range

☑ Show Highest Bonds (pair interactions) dG Threshold: -5.0

Path to Selected Genome File to check the amplicons for overlaps.

☑ Use DeltaG. If not use instead scores (faster and simpler but less accurate)

DeltaG Calculation Settings :

Temperature (in C) : 37

Magnesium Concentration (in mM) : 0

Monovalent cation (in mM) : 50

dNTP concentration (in mM) : 0

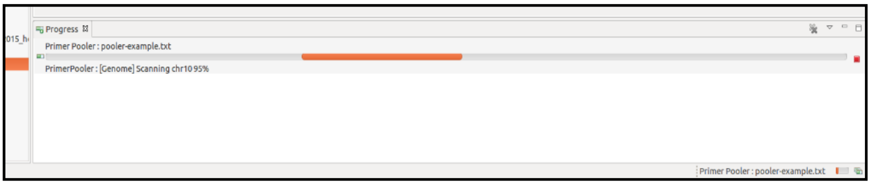
Powered by PrimerPooler.
Brown, Silas S., et al. "PrimerPooler: automated primer pooling to prepare library for targeted sequencing." Biology Methods and Protocols 2.1 (2017): bpx006.

Figure 3.13 PrimerPooler dialog

- Select Genome: select a genome file in 2bit format to check the amplicons for overlaps. Primer pairs that amplify overlapping regions of the genomes can produce an unwanted shorter amplicon if used in the same pool. (Fasta file use support will be added in future releases of the tool).
    - For this example: select hg38.2bit from the examples folder.

- Use DeltaG: use thermodynamic principles to calculate the correspondent ΔG for the pairwise interaction. If not selected, a score will be automatically calculated based on alignment. However please note that such a score will be calculated in a faster but less accurate way. To use deltaG you will need to input Temperature, Concentration of magnesium, Concentration of monovalent cation, and Concentration of deoxynucleotide (dNTP) parameters.

Upon parameter selection, click “OK” to launch the task. You can track its progress or cancel it from the progress view. To open the progress double click on the progress bar at the bottom of the screen (Figure 3.14). Alternatively, open the progress view from the view’s menu.



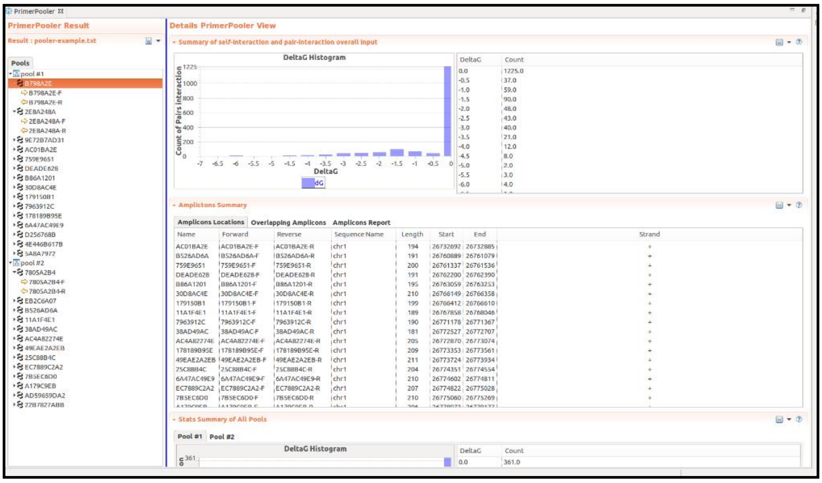
Primer Pooler : pooler-example.txt

PrimerPooler : [Genome] Scanning chr10 95%

Primer Pooler : pooler-example.txt

Figure 3.14 Progress View showing the progress of a task launched with the PrimerPooler tool

Upon completion, a new PrimerPooler view will appear containing a full report of the results as shown in Figure 3.15. The left side of the view will display a tree view of the pools and primers assigned to each pool. The right-hand side panel will contain different sections of the result depending on the options selected by the user in the run dialog.



Result : pooler-example.txt

Pools
- pool #1
  - B798A2E
    - B798A2E-F
    - B798A2E-R
  - 2E8A248A
    - 2E8A248A-F
    - 2E8A248A-R
  - 9E72B7AD31
  - AC01BA2E
  - 759E9651
  - DEADE628
  - B86A1201
  - 30D8AC4E
  - 179150B1
  - 7963912C
  - 178189B95E
  - 6A47AC49E9
  - D256768B
  - 4E446B617B
  - 5A8A7972
- pool #2
  - 7805A2B4
    - 7805A2B4-F
    - 7805A2B4-R
  - EB2C6A07
  - B526AD6A
  - 11A1F4E1
  - 38AD49AC
  - AC4A82274E
  - 49EAE2A2EB
  - 25C88B4C
  - EC7889C2A2
  - 7B5EC6D0
  - A179C9EB
  - AD59659DA2
  - 22B7827ABB

DeltaG	Count
0.0	1225.0
-0.5	37.0
-1.0	59.0
-1.5	90.0
-2.0	48.0
-2.5	43.0
-3.0	40.0
-3.5	21.0
-4.0	12.0
-4.5	8.0
-5.0	2.0
-5.5	3.0
-6.0	4.0

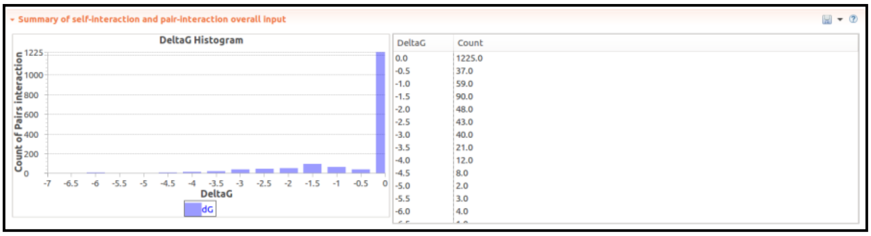
Amplicons Locations | Overlapping Amplicons | Amplicons Report

Name	Forward	Reverse	Sequence Name	Length	Start	End	Strand
AC01BA2E	AC01BA2E-F	AC01BA2E-R	chr1	194	26732692	26732885	+
B526AD6A	B526AD6A-F	B526AD6A-R	chr1	191	26760889	26761079	+
759E9651	759E9651-F	759E9651-R	chr1	200	26761337	26761536	+
DEADE628	DEADE628-F	DEADE628-R	chr1	191	26762200	26762390	+
B86A1201	B86A1201-F	B86A1201-R	chr1	195	26763059	26763253	+
30D8AC4E	30D8AC4E-F	30D8AC4E-R	chr1	210	26766149	26766358	+
179150B1	179150B1-F	179150B1-R	chr1	199	26766412	26766610	+
11A1F4E1	11A1F4E1-F	11A1F4E1-R	chr1	189	26767858	26768046	+
7963912C	7963912C-F	7963912C-R	chr1	190	26771178	26771367	+
38AD49AC	38AD49AC-F	38AD49AC-R	chr1	181	26772527	26772707	+
AC4A82274E	AC4A82274E-F	AC4A82274E-R	chr1	205	26772870	26773074	+
178189B95E	178189B95E-F	178189B95E-R	chr1	209	26773353	26773561	+
49EAE2A2EB	49EAE2A2EB-F	49EAE2A2EB-R	chr1	211	26773724	26773934	+
25C88B4C	25C88B4C-F	25C88B4C-R	chr1	204	26774351	26774554	+
6A47AC49E9	6A47AC49E9-F	6A47AC49E9-R	chr1	210	26774602	26774811	+
EC7889C2A2	EC7889C2A2-F	EC7889C2A2-R	chr1	207	26774822	26775028	+
7B5EC6D0	7B5EC6D0-F	7B5EC6D0-R	chr1	210	26775060	26775269	+

DeltaG	Count
0.0	361.0

Figure 3.15 PrimerPooler result View.

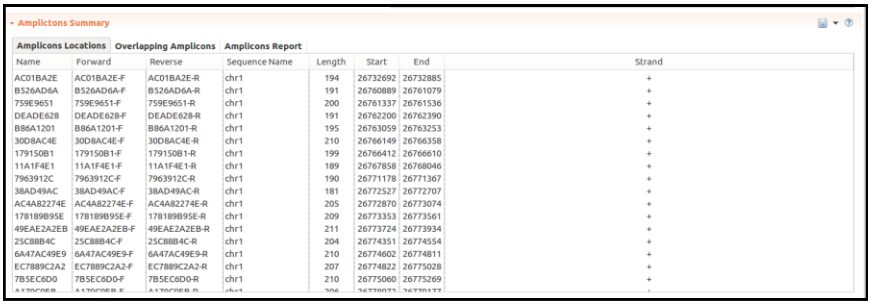
If "Count deltaG/Score" is selected, a count summary of such interactions will be displayed in both a graph and a table as shown in Figure 3.16. The user may export these results from the correspondent "save" menu.


Summary of self-interaction and pair-interaction overall input
DeltaG Histogram
Count of Pairs interaction
1225
1000
800
600
400
200
0
-7
-6.5
-6
-5.5
-5
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
DeltaG
dG

DeltaG	Count
0.0	1225.0
-0.5	37.0
-1.0	59.0
-1.5	90.0
-2.0	48.0
-2.5	43.0
-3.0	40.0
-3.5	21.0
-4.0	12.0
-4.5	8.0
-5.0	2.0
-5.5	3.0
-6.0	4.0

Figure 3.16 Summary of the "Count of pair interactions" overall input.

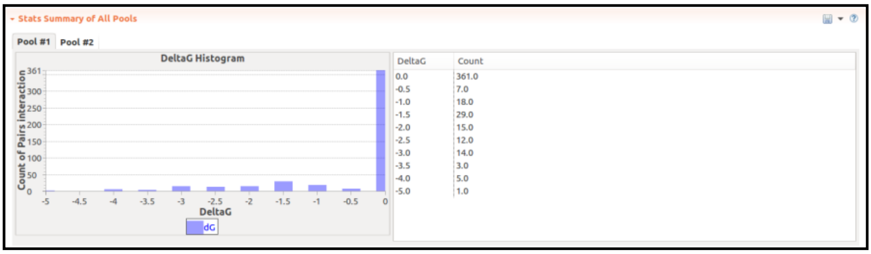
If a genome file is provided, an amplicons summary section will be displayed as shown in Figure 3.17, containing three tabs for amplicon locations in the genome, overlapping amplicons details and an amplicon summary report in text format. These results may also be exported via the "save" menu. In the Amplicon locations table, those amplicons that cannot be located in the genome will show no corresponding location.



Name	Forward	Reverse	Sequence Name	Length	Start	End	Strand
AC01BA2E	AC01BA2E-F	AC01BA2E-R	chr1	194	26732692	26732885	+
B526AD6A	B526AD6A-F	B526AD6A-R	chr1	191	26760889	26761079	+
759E9651	759E9651-F	759E9651-R	chr1	200	26761337	26761536	+
DEADE628	DEADE628-F	DEADE628-R	chr1	191	26762200	26762390	+
B86A1201	B86A1201-F	B86A1201-R	chr1	195	26763059	26763253	+
30D8AC4E	30D8AC4E-F	30D8AC4E-R	chr1	210	26766149	26766358	+
179150B1	179150B1-F	179150B1-R	chr1	199	26766412	26766610	+
11A1F4E1	11A1F4E1-F	11A1F4E1-R	chr1	189	26767858	26768046	+
7963912C	7963912C-F	7963912C-R	chr1	190	26771178	26771367	+
38AD49AC	38AD49AC-F	38AD49AC-R	chr1	181	26772527	26772707	+
AC4A82274E	AC4A82274E-F	AC4A82274E-R	chr1	205	26772870	26773074	+
178189B95E	178189B95E-F	178189B95E-R	chr1	209	26773353	26773561	+
49EAE2A2EB	49EAE2A2EB-F	49EAE2A2EB-R	chr1	211	26773724	26773934	+
25C88B4C	25C88B4C-F	25C88B4C-R	chr1	204	26774351	26774554	+
6A47AC49E9	6A47AC49E9-F	6A47AC49E9-R	chr1	210	26774602	26774811	+
EC7889C2A2	EC7889C2A2-F	EC7889C2A2-R	chr1	207	26774822	26775028	+
7B5EC6D0	7B5EC6D0-F	7B5EC6D0-R	chr1	210	26775060	26775269	+

Figure 3.17 Amplicon location and overlap summary.

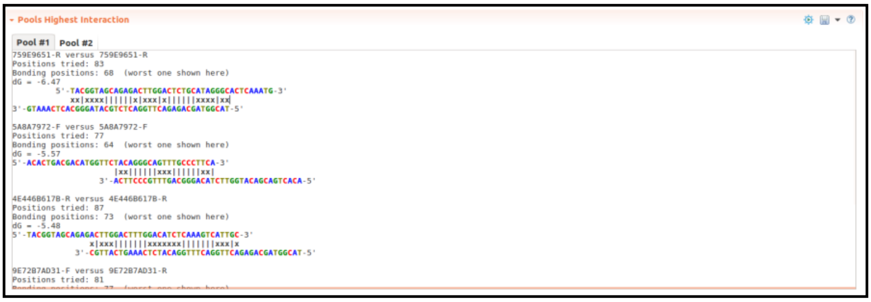
Figure 3.18 shows the “Stats Summary of All Pools”, showing the same information as that included in “count of pair interaction” but this time at the level of each pool (*e.g*. only consider interaction between primers within the same pool). In this section a tab for each pool will be added, so the user can navigate the results by selecting the pool in the corresponding tab.



DeltaG	Count
0.0	361.0
-0.5	7.0
-1.0	18.0
-1.5	29.0
-2.0	15.0
-2.5	12.0
-3.0	14.0
-3.5	3.0
-4.0	5.0
-5.0	1.0

Figure 3.18 . Pools Stats Summary.

Lastly, if “Show Highest Interaction” is selected, a section for the interaction will be displayed showing the highest interaction and highlighting the dimer structure formed between the two primers (Figure 3.19). From the section menu the user may then re-run this task selecting a different score/dG threshold value, then export the result to text files if desired.



```
759E9651-R versus 759E9651-R
Positions tried: 83
Bonding positions: 68  (worst one shown here)
dG = -6.47
         5'-TACGGTAGCAGAGACTTGGACTCTGCATAGGGCACTCAAATG-3'
            xx|xxxx||||||x|xxx|x||||||xxxx|xx
3'-GTAAACTCACGGGATACGTCTCAGGTTCAGAGACGATGGCAT-5'

5A8A7972-F versus 5A8A7972-F
Positions tried: 77
Bonding positions: 64  (worst one shown here)
dG = -5.57
5'-ACACTGACGACATGGTTCTACAGGGCAGTTTGCCCTTCA-3'
                     |xx||||||xxx||||||xx|
                  3'-ACTTCCCGTTTGACGGGACATCTTGGTACAGCAGTCACA-5'

4E446B617B-R versus 4E446B617B-R
Positions tried: 87
Bonding positions: 73  (worst one shown here)
dG = -5.48
5'-TACGGTAGCAGAGACTTGGACTTTGGACATCTCAAAGTCATTGC-3'
                x|xxx|||||||xxxxxxx|||||||xxx|x
             3'-CGTTACTGAAACTCTACAGGTTTCAGGTTCAGAGACGATGGCAT-5'

9E72B7AD31-F versus 9E72B7AD31-R
Positions tried: 81
Bonding positions: 77  (worst one shown here)
```

Figure 3.19 Pools Interactions Summary.

4 CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies

Hovhannisyan, Hrant*, **Ahmed Hafez***, Carlos Llorens, and Toni Gabaldón. 2020. "CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies." *Bioinformatics* 36 (3):925-927. doi: 10.1093/bioinformatics/btz626.

* - equal contribution
H. Hovhannisyan has designed the project, written the initial data analysis pipeline, documented, tested, and maintains the software. A. Hafez has implemented the final pipeline and its automation, tested, and maintains the software package.

4.1 Abstract

4.1.1 Motivation

Numerous sequencing studies, including transcriptomics of host-pathogen systems, sequencing of hybrid genomes, xenografts, mixed species systems, metagenomics, and meta-transcriptomics, involve samples containing genetic material from divergent organisms. A crucial step in these studies is identifying from which organism each sequencing read originated, and the experimental design should be directed to minimize biases caused by cross-mapping of reads to incorrect source genomes. Additionally, pooling of sufficiently different genetic material into a single sequencing library could significantly reduce experimental costs but requires careful planning and assessment of the impact of cross-mapping. Having these applications in mind we designed CrossMapper, the first to our knowledge tool able to assess cross-mapping prior to sequencing, therefore allowing optimization of experimental design.

4.1.2 Results

Using any combination of reference genomes, CrossMapper performs read simulation and back-mapping of those reads to the pool of references, quantifies, and reports the cross-mapping rates for each organism. CrossMapper performs these analyses with numerous user-specified parameters, including, among others, read length, read layout, coverage, mapping parameters, genomic or transcriptomic data. Additionally, it outputs the results in highly interactive and publication-ready reports. This allows the user to perform multiple comparisons at once and choose the experimental setup minimizing cross-mapping rates. Moreover, CrossMapper can be used for resource optimization in sequencing facilities by pooling different samples into one sequencing library.

4.1.3 Availability and implementation

CrossMapper is a command line tool implemented in Python 3.6 and available as a conda package, allowing effortless installation. The source code, detailed information and a step-by-step tutorial is available at our GitHub page https://github.com/Gabaldonlab/crossmapper.

4.2 Introduction

There are various biological problems addressed by next-generation sequencing (NGS) in which the samples contain genetic material from multiple species. These include, but are not limited to studies involving host-pathogen interaction (Westermann *et al.* 2017), symbiont-host or microbial interaction (Burns *et al.* 2017, González-Torres *et al.* 2015), metagenomics (Quince *et al.* 2017), or hybrid organisms (Metzger *et al.* 2017). A challenging step in these experimental setups is to assign each sequencing read to the corresponding source organism, which is usually done by mapping the reads to the set of reference genomes (Wolf *et al.* 2018). A similar strategy is applied in allele-specific expression (ASE) studies in the case of phased reference genomes (Yuan and Qin 2012). Successful read separation depends on numerous factors, including mainly read length, read layout, similarity of sequenced genomes and different mapping parameters. Thus, if these parameters are not carefully planned, downstream analyses can be biased by cross-mapping of reads to non-corresponding references. For example, in a human-*Salmonella* interaction study it was observed that ~1.44% of total reads map equally well (multi-mapped) to both reference genomes (Westermann and Vogel 2018). While the amount of erroneously mapped reads can be low for highly divergent species, in metagenomics (Petersen *et al.* 2017) and ASE studies, erroneously mapped and multi-mapped reads constitute the majority of the data (Yuan and Qin 2012). Despite the importance of sequencing design in aforementioned studies, today there are no computational tools to assist in their planning so that optimal results are obtained.

To overcome this, we developed CrossMapper – a pipeline assessing, prior to sequencing, the potential rates of multi-mapping and erroneous mapping for various combinations of sequencing parameters and any number of reference sequences.

4.3 Workflow and implementation

CrossMapper proceeds as follows (Figure 4.1A). It first takes as input any number of reference genomes and allows to simulate DNA and RNA reads in a wide range of experimental setups. This step is performed by wgsim (Li *et al.* 2009) with the possibility to define different parameters such as read length, error rates, outer distance, among others. CrossMapper allows to simulate many different sequencing configurations at once. The user can specify genome annotations to limit read simulations from specific parts of the genomic regions (*i.e.* for transcriptomic or exome sequencing studies).

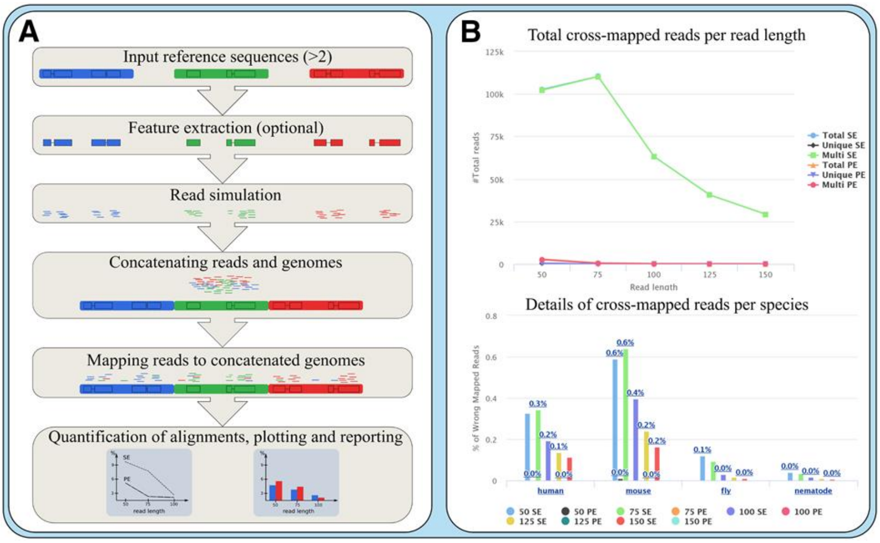


Figure 4.1 A. The general workflow of CrossMapper (see main text for details). B. An example of CrossMapper output.

After read simulation, CrossMapper concatenates fastq files from different organisms and maps the reads back to a concatenated set of reference genomes. By default CrossMapper uses BWA-MEM (Li and Durbin 2009), and STAR (Dobin *et al.* 2013) for mapping DNA and RNA data, respectively. However, we also implemented the --*mapper-template* option allowing to use any desired mapping software with custom parameters by supplying the configuration file to the CrossMapper (a documentation for creating a configuration file is given in the GitHub page). The final bam file for each read length and layout contains alignments of all simulated reads collectively mapped to all source reference genomes. Since simulated data preserve information regarding the source genome and exact location, CrossMapper can calculate the rate of multi-mapped and erroneously mapped reads for all source genomes. After the quantification step CrossMapper produces an extensive html report, which includes several interactive, publication-ready plots summarizing mapping rates, as well as tables with detailed mapping statistics for each experimental configuration. Based on this report users can decide the optimal experimental and mapping parameters prior to the actual sequencing. In addition, coordinates of cross-mapped reads are reported so these regions can be filtered, if necessary, in downstream analyses.

4.4 Usage case

Several examples of CrossMapper usage are available in the GitHub site of this tool. Here we explain how to use CrossMapper to optimize resources by pooling genetic material of different organisms into a single sequencing library. Indeed, the cost of sequencing has dropped dramatically in the past decade (Goodwin *et al.* 2016), largely due to throughput increase. However, the costs for library preparation do not follow the same trend and often constitute a financial bottleneck. A simple pooling of genetic materials of different species into one library could save a substantial amount of resources, provided reads from different sources could effectively be separated computationally. This has to be carefully planned to avoid aforementioned biases in downstream analyses. CrossMapper can achieve this task in a single run. Below is an example of sequencing design optimization for pooling genetic material of widely analysed organisms – human, mouse, fly and nematode – in a single library.

Command syntax

```
# Command to run
> crossmapper DNA -t 8 -gb -rlay both -g homo.fasta
mus.fasta dros.fasta caeno.fasta -gn human mouse fly
nematode -N 2500000 2500000 2500000 2500000 -rlen
50,75,100,125,150 -r 0.01
```

This command lets CrossMapper to simulate 2.5 million DNA reads per organism at 50, 75, 100, 125 and 150 read lengths at both single- and paired-end layouts, map the data to the pool of reference genomes (obtained from Ensembl (Zerbino *et al.* 2018)) and report mapping rates for all sequencing configurations (Figure 4.1B). Using Intel Xeon 3.5GHz, 64GB of RAM and 8 cores the analysis takes ~11 hours. In this case of a very large reference genome (ca 6.5 GB) and 10 mapping jobs, the main bottleneck for the speed of the analysis is genome indexing and read mapping, which collectively takes ~8 hours.

The results of this analysis (Suppl. File S1) demonstrate that by pooling the DNA of the 4 species reads can be effectively separated by mapping. However, single-end sequencing produces relatively high rates of multi-mapping (maximum 4.95% and 6.06% for 150 and 50 bp, respectively) and erroneous mapping (maximum 0.16% and 0.52%, for 150 and 50 bp, respectively) which potentially can bias differential expression or variant calling analysis. On the other hand, paired-end sequencing with 75 bp reads significantly reduces multi- and erroneous mapping (0.01% and 0%, respectively) rates. Thus, the pooling strategy with 2x75 bp reads can be the most efficient balance between accuracy and sequencing cost. Repeating this test with a higher number of reads (40, 30, 20 and 10 mln reads for human, mouse, fly and nematode, respectively), showed similar rates of cross-mapping (Suppl. File S2), which indicates that low-coverage simulations are sufficient to properly estimate cross-mapping rates.

Supplementary material 2 list more usage cases for CrossMapper.

4.5 Conclusion

CrossMapper allows to design numerous types of NGS experiments that share a common feature of sequencing several organisms as one sample. CrossMapper is easy to install and use. It is highly customizable and outputs the results in intuitive, interactive and publication-ready reports. We believe that CrossMapper will benefit both research and industrial communities by helping to optimize sequencing strategies and available resources.

5 Recurrent neural networks for classification and regression problems in DNA and RNA sequences with rnnXna

Hafez, Ahmed, Essam H. Houssein, Carlos Llorens, Toni Gabaldón. "Recurrent neural networks for classification and regression problems in DNA and RNA sequences with rnnXna." *(In preparation).*

5.1 Abstract

5.1.1 Summary

Recurrent neural network (RNN) models have shown promising results in machine translation, speech recognition and computational biology. Here, we propose the use of RNN models for a wide class of regression and classification problems involving DNA and RNA sequences. We tested this idea by applying RNNs to the complex problem of predicting the structural features of RNA molecules. RNAs play important roles in virtually every cellular process. Determining the secondary structures of RNAs is central to understanding their function and evolution, as their functions are mediated through the adoption of specific structures that enable RNAs to interact with other molecules. Our results showed a great performance of RNN classification to predict the preferred state of each residue in an RNA molecule either being single-stranded or double-stranded in the RNA secondary structure. We developed rnnXna, a tool with a simple command line interface that allows to train and use the RNN models introduced in this work in a variety of other applications.

5.1.2 Availability and implementation

The rnnXna tool is a command line tool implemented in Python 3 and available as a conda package, allowing effortless installation. The source code, training datasets, examples and detailed information and a step-by-step HowTo is available at our GitHub page https://github.com/Gabaldonlab/rnnXna.

5.2 Introduction

Neural networks are used in a wide range of machine learning and deep learning applications especially after the recent advances in computation hardware and techniques *e.g.* specialized AI accelerators such as Tensor Processing Units (TPU). Neural networks are a scalable and effective solution for many complex problems and are able to identify complex patterns from feature-rich datasets (LeCun *et al.* 2015). With recent developments in next-generation DNA sequencing technologies, more and more diverse data types are produced, thus introducing many computational challenges that could be tackled by deep learning applications (Koumakis 2020). For example, neural networks and deep learning have already been adapted for genomics problems such as inferring DNA sequences' function (Quang and Xie 2016) and motif discovery (Alipanahi *et al.* 2015).

Recurrent neural network (RNN) (Sherstinsky 2020) is a variant of neural network that captures the sequential information present in the input data -*e.g.* dependency between words in a text- while making predictions. RNN models are naturally suited to temporal or sequential data, and several variants have been developed for sequenced features such as long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) which have shown impressive performance in numerous sequence-based modelling and sequence to sequence mapping tasks such as natural language modelling (Cho *et al.* 2014), early detection of heart failure (Choi *et al.* 2017) or predicting transcription factor binding sites in DNA sequences (Shen *et al.* 2018).

RNA is a single-stranded nucleotide molecule which folds into complex secondary and tertiary structures (Mortimer *et al.* 2014). The structural conformation of an RNA molecule is key for its function and regulation (Cruz and Westhof 2009). Over recent years, several methods have been developed to probe RNA structures *in vivo* in a genome-wide scale (Leamy *et al.* 2016, Strobel *et al.* 2018). Some methods combine high-throughput sequencing technologies (NGS) with traditional enzymatic- or chemically-based assays to enable structural profiling of transcribed RNAs at genome-wide scales (Kertesz *et al.* 2010, Saus *et al.* 2018, Lu *et al.* 2018). Such methods can determine secondary structures at a single base resolution, indicating whether a nucleotide is likely to be single-stranded or double-stranded in the folded RNA structure. Computational prediction methods can help improving the accuracy of empirical methods or replace them as an alternative *in silico* approaches, if limitations to perform such experiments are present (Ouyang *et al.* 2013, Saus *et al.* 2018, Wu *et al.* 2015). For example, the nextPARS method (Saus *et al.* 2018) uses RNN models to improve the final prediction score.

In this work, we investigate and explore how RNN models can be applied to DNA or RNA sequence classification problems that depend on the structure information of a given sequence and the relationship between adjacent nucleotides such as RNA secondary structure, oligonucleotide melting temperature or other similar problems. To this end, we developed and performed a computation assessment of RNN classification models that can predict structural profiling of RNA molecules. Such models can be trained for other prediction problems given sufficient training data. In addition, we explored the applicability of RNN models to regression problems, for example to predict or calculate a physical/chemical property of a given oligonucleotide DNA sequence such as the melting temperature, which is of special interest for the screening and designing of primers for PCR reactions. Finally, we developed rnnXna, a general-purpose tool for training and validating RNN models for problems that share a similar structure to the ones introduced in this work. rnnXna tool has a simple command line interface allowing the non-experienced users to train and deploy recurrent neural network models in their research.

5.3 Results

5.3.1 Classification Model

We developed a RNN classification model that can predict the preferred structural state (single-stranded SS or double-stranded DS) of each base in an RNA molecule's secondary structure. The model takes a short *k*-mer RNA fragment and predicts if the base in the middle of the fragment is single-stranded or double-stranded in the secondary structure. To train this model RNA sequences and their secondary structure were collected for the RNA Strand database (Andronescu *et al.* 2008) to construct training datasets. *k*-mer datasets were constructed with different *k*-mer lengths (k=[11,13,15,17,19,21]) to build different RNN *k*-classifiers. 5-fold cross validation was performed to assess *k*-classifiers' performance. For the purpose of performance comparison an artificial neural network (ANN) was constructed and trained with the same dataset to build a corresponding ANN *k*-classifier (see Materials and Methods for more details). In addition, a selection of different classification methods widely used in machine learning were also used for the comparison: Nearest Neighbours (Sammut and Webb 2010b), Decision Tree (Fürnkranz 2010), Random Forest (Sammut and Webb 2010c), AdaBoost (Sammut and Webb 2010a), Naive Bayes (Webb 2010), Linear Support Vector Machine (SVM) (Zhang 2010, Cristianini and Shawe-Taylor 2000) and Quadratic Discriminant Analysis (QDA) (McLachlan 1992). Some methods are not feasible for large datasets such as SVM with non-linear kernels (Cristianini and Shawe-Taylor 2000). The classification methods implemented by Scikit-learn (Pedregosa *et al.* 2011) are used to build different *k*-classifiers using the same dataset used in training RNN k-classifiers.

Receiver operating characteristic curve (ROC) and Area under the ROC Curve (AUC) (Metz 1978) was measured to assess the performance of the classification models at all classification thresholds on the cross validation stage. Supplementary Figure 5.4 reports ROC and Area under the ROC Curve (AUC) on 5-fold cross validation with different *k*-mer length as an aggregate measure of performance across all possible classification thresholds for all classification methods included in the comparison. Figure 5.1 summarizes the mean AUC for each *k-classifier* on cross validation for all methods. In both RNN and ANN models we can observe that for smaller *k*-mer length models have a lower performance in classification power, and as the *k*-mer length increases so does the classification performance. This is due to the fact that as the *k*-mer length increases, the model gets more features and more discrimination power. RNN models have slightly better performance (on average a 5% increase in AUC) with all different *k*-mers over the corresponding ANN models with the same *k*-mer length. Other methods have significantly lower performance compared to neural networks in general, except for nearest neighbour classifiers which show similar performance to ANN classifiers. However, one major advantage of neural networks over the Nearest neighbour method and other methods in general is its scalability to large datasets and the ability to reduce training and prediction runtime using hardware accelerators such as GPU or TPU.

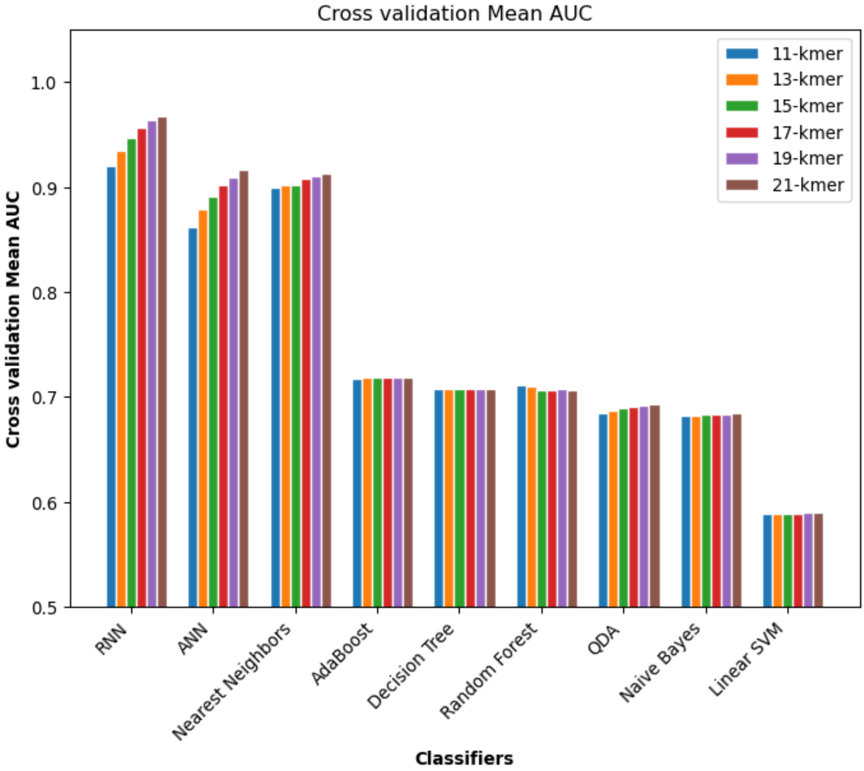


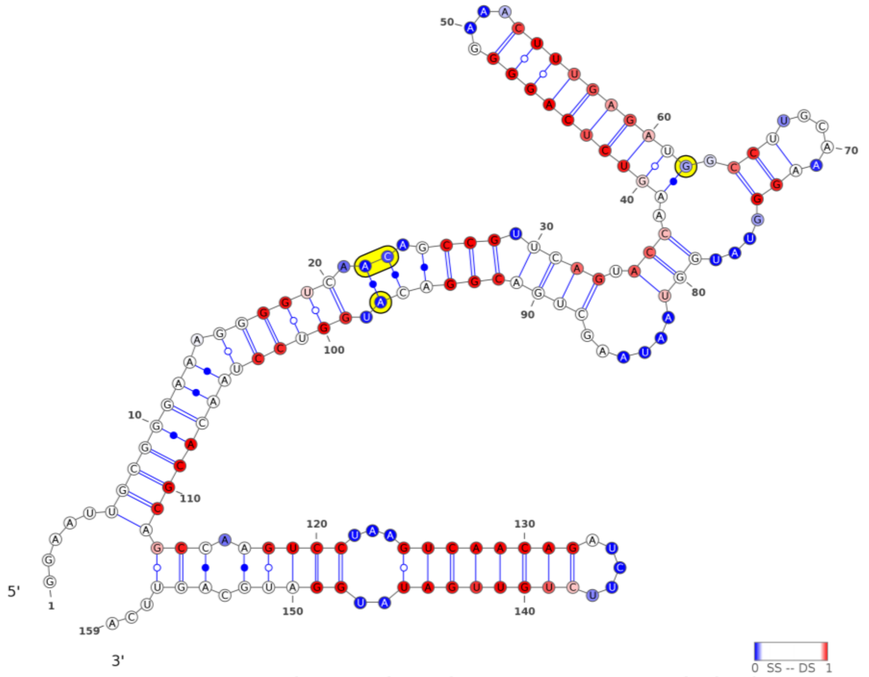
Figure 5.1 Mean AUC on Cross Validation for each k-classifier for RNN models compared to other classifiers. Individual ROC plot and mean AUC for each k-classifier for all methods are listed in the supplementary material.

The RNN model was validated by a test RNA molecule, which was excluded from the training dataset. Assigning DS/SS class to each base depends on the prediction probability and the threshold used for assigning the final call. Model prediction assigns each base a score for being DS/SS in the range of [0,1], where 0 means SS and 1 means DS. We assessed the model classification at three probability thresholds (50%, 80% and 90%) used as a confidence level of the final call of the classifier. As an example at 80% threshold classifier will assign SS to a nucleotide if the prediction score is lower than 0.2 and DS if the prediction score is greater than 0.8, otherwise it will be a undetermined call *i.e.* cannot assign a class. Table 5.1 summarizes the test performance for *k-classifier* [k=21] in terms of accuracy, precision, and recall. Accuracy is the number of times the classifier assigns the correct label DS/SS to each nucleotide, precision in this context presents the ability of the classifier to correctly identify DS sites, recall is the ability of the classifier to correctly identify SS sites.

At 50% threshold the final number calls that a classifier can determine is 139 bases, as the first 11 bases and the last 11 bases cannot be determined by a 21 *k*-mer classifier. The accuracy of the model is about 89%. The Precision is 97% and recall is 87%. At this level the classifier has a problem identifying SS sites and about 23% of SS sites are identified as DS instead of SS. To increase model confidence another threshold can be used to call a DS/SS site. 80% and 90% thresholds were used as threshold, and we can observe that by increasing the threshold we can obtain a better classification performance. However, the number of bases that the model can classify decreases. Figure 5.2 shows RNN *21-classifier DD/SS* calls at 90% threshold for the structure of TETp4p6 RNA molecule determined by high resolution X-ray and visualized by VARNA tool (Darty *et al.* 2009).

Table 5.1 RNN 21-classifier performance in terms of accuracy, precision and recall for the TETp4p6 RNA molecule.

Threshold	Accuracy	Precision	Recall	# Calls
50%	0.89	0.97	0.87	139
80%	0.94	0.98	0.92	118
90%	0.96	1.0	0.94	101



TETp4p6 - RCSB Protein Data Bank 2R8S (PDB_01255 - RNA SSTRAND database)

Figure 5.2 Visual representation by VARNA (Visualization Applet for RNA) of DS/SS sites assigned by RNN 21-classifier at 90% threshold to TETp4p6 RNA molecule and its secondary structure. DS sites assigned by the classifier are in red. SS sites assigned by the classifier are in blue. All sites that cannot be called by the classifier at this threshold are in white. SS sites that are misclassified as DS are highlighted in yellow colour.

To obtain a better performance and to increase classification confidence, Ensemble classification (Ren *et al.* 2016) can be used to calculate the final score/probability. In this context ensemble of classifiers can be built by aggregating scores from all RNN models with different *k*-mer length to get a consensus classification score. Ensemble classification was used in nextPARS (Saus *et al.* 2018) using a similar RNN model (one LSTM layer) and trained only will all RNA secondary structures that were validated by NMR or X-Ray techniques. The ensemble classifier used RNN *k*-classifiers of different *k*-mer length (*k* = [7,9,11,13,15]) on and the final score was used as prior in conjunction with a score obtained by experimental digestion profile using NGS technology *i.e.* raw number of enzymatic cuts per position. The nextPARS RNN classification score is calculated according to the following formula:

$$S_{RNN} = \sum_{k \epsilon k} w_k S_k \, , \%$$

where S_{RNN} is the RNN final classification score, S_{k} is the classification score from each *k*-classifier for *k*-mer sequences fragments, and w_{k} is a weight associated with each *k*-classifier, which can be used to assign more weight to *k*-classifier with larger *k* length and less weight to smaller *k*.

5.3.2 Regression Model

RNN models can also be used for regression problems, in which the model response is a scalar value. In the context of RNA and DNA sequences, this could be a physical or a chemical property of the sequence that requires an experiment to calculate. As a synthetic example we used here the melting temperature (Tm) of the given sequences as the response of the model. The Tm synthetic dataset (see Materials and Methods for more details) represents a good example of a wide range of problems that RNN models can be applied to, in which the final output depends on the sequence structure and local dependency between neighbour nucleotides. Therefore, RNN models can be trained to infer such functions.

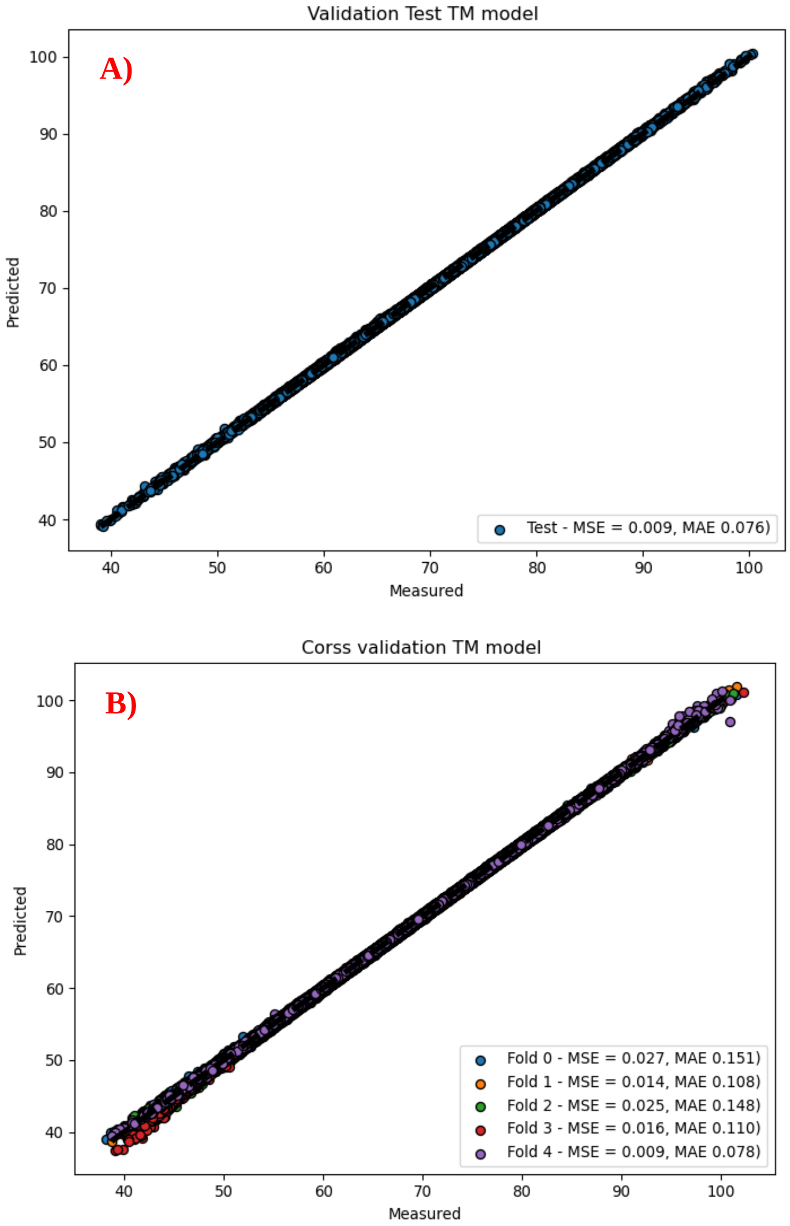


Figure 5.3 RNN regression model performance (Predicted Tm vs Measured Tm). A) During cross validation: the figure reports MSE and MAE on each fold. B) On a separated test dataset.

The RNN regression model was assessed by performing 5-fold cross validation on the training dataset. The average mean square error on cross validation was 0.019 with a standard deviation of 0.007 and mean absolute error of 0.119 with standard deviation of 0.027. Figure 5.3A highlights the predicted Tm by the model versus the actual Tm measured by primer3 tool (Untergasser *et al.* 2012) in each cross validation fold. It can be observed that most samples in the middle range of Tm, around 50º to 90º, have a low error rate, however sequences with very low and high melting temperature have a slightly higher error rate of prediction. The reason for this is that such sequences are less represented in the training dataset, for example about 15 sequences in the training dataset have Tm over 99º and have high percent of CG content. As a result, the model cannot learn accurately how to model such sequences in the training stage as they are distributed in different folds during cross validation. Figure 5.3B shows the model prediction on a separated test dataset, where it performs better around the low and high end of the Tm ranges, this is because the model was trained with the entire training dataset and was able to correctly model such sequences. The model performance in the test dataset has an MSE of 0.009 and MAE was 0.078. RNN was able to learn an accurate model for Tm calculation and was successful to simulate a similar behaviour to primer3 tool.

5.3.3 rnnXna tool

We developed rnnXna, which provides a simple command interface to train RNN models from input sequences. The tool has a training mode, and a prediction mode that uses previously trained models. To train a new model the tool accepts an input file in csv/tsv format where a short sequence is listed with the corresponding class (for classification models) or an associated score or property value *e.g.* melting temperatures (for regression models). The tool exposes some control parameters to configure the model such as number of layers, option to add drop out layers to minimize overfitting, number of iterations in the training, etc. For prediction, the tool accepts a simple text/csv file containing input sequences for prediction or a fasta file with a long input sequence, that will be divided into *k*-mers to perform the prediction step.

Command syntax

```
# Command 1 : training model
> rnnXna train -i traing_data.csv -o rnn_model

# Command 2 : perform prediction on new data
# new data can be csv file or fasta file
> rnnXna predict -model rnn_model -fasta
input_sequences.fasta

# Command 3 : perform cross- validation
> rnnXna train -i traing_data.csv --cv 5 --lstm-layers 2
```

rnnXna can also perform a cross validation on the training data or validation test on a separated test dataset under different parameter configuration in order to assess model performance before final training. The current implementation supports training and prediction on CPU, or a graphical processing unit (GPU) if available for a better performance. More usage examples are available on the online user guide in our GitHub page.

5.4 Discussion

Recurrent neural networks are an attractive solution to the structural profiling of RNA secondary structure. The proposed and developed RNN model in this work shows a good performance in predicting double and single stranded sites in the folded RNA structure. In addition, using RNN models for regression show a great performance in predicting Tm of short DNA sequences. The reported performance in terms of MSE and MAE of the regression model show that the model was able to learn the nearest neighbour thermodynamic model parameters (SantaLucia 1998) used in primer3 to calculate the melting temperature without any prior knowledge of the nature of the problem.

However, the current implementation of RNN model poses some limitations. One of them is that it only accepts as input the DNA or RNA sequences. In the case of the Tm dataset it would be desirable to incorporate other input parameters such as condition parameters (*e.g.* salt or dNTP concentrations in the PCR solution) to be able to predict the Tm under different conditions similar to primer3 tool. Another limitation is that it only takes a fixed length sequence, the model can be extended to accept a variable length sequence by implementing word embedding and padding solution similar to the model proposed in (Alipanahi *et al.* 2015) making the model more flexible in training and prediction. RNN models can be trained for sequence to sequence application (Sutskever *et al.* 2014). Therefore, the possibility to adopt the current RNN model to predict directly the presence of stable hairpin structure or dimers formation under given conditions *e.g.* for screening and selecting best oligonucleotide for PCR primers.

5.5 Materials and Methods

5.5.1 RNA secondary structure Dataset

RNA secondary structures were downloaded from RNA STRAND Database (Andronescu *et al.* 2008) for all RNA molecules larger than 30 nucleotides (nt) from all sources. In total 3118 RNA molecules were retrieved with a total of 2,323,909 nt. The average length of RNA sequences was 745 nt and ranged from 30 nt to 4380 nt. TETp4p6 molecule (2R8S entry in RCSB Protein Data Bank and PDB_01255/PDB_00082 entries in RNA STRAND database) was removed from the training data set to be used as a testing dataset of RNN classifiers.

To prepare the RNA sequences for training the classifiers. Each sequence is fragmented into *k*-mer sub-sequences and the corresponding class DS/SS from the molecule secondary structure is assigned for the nucleotide in the middle of *k*-mer sub-sequence. Multiple training dataset of different *k*-mer sizes (k = [11,13,15,17,19,21]) were constructed for each *k*-classifier. Table 5.2 lists the number of fragments in the constructed training set for different values of *k* after removing any fragment with ambiguous nucleotide characters.

Table 5.2 Total number of sequence fragments for each k-classifier training dataset.

k	Number of k-mer fragments
11	2241023
13	2231952
15	2223094
17	2214442
19	2206045
21	2197824

5.5.2 Melting Temperature (Tm) Dataset

A synthetic dataset was constructed by selecting short oligonucleotide DNA sequences of 32-nucleotides from the human genome. In total unrepeated 2.5 million fragments were randomly selected, then for each oligonucleotide the melting temperature we calculated using primer3 tool (Untergasser *et al.* 2012) with the default parameter which are 0mM divalent cations, 0mM dNTPs concentration, 50mM monovalent cations, 50nM concentration of DNA strands and thermodynamic parameters from (SantaLucia 1998). The dataset was divided into a training dataset used for training and cross validation of size 2 million and a testing dataset of 0.5 million that was never used in training nor cross validation.

5.5.3 RNN models

The recurrent neural network structure is constructed with 2 Long Short-Term Memory (**LSTM)** (Hochreiter and Schmidhuber 1997) layers and one fully connected neural network layer (Dense layer). LSTM is used as it overcomes the problem of long-term dependency in normal RNN and resolves vanishing/exploding gradients. Sequences fragments used for input in the network are represented by a binary vector in which each nucleotide character is represented by 4-binary one-hot encoding vector (Rodríguez *et al.* 2018), and final input vector size for a *k*-mer sequence fragment is *n* = 4**k*. The input vectors are fed to the first hidden LSTM layer which consists of 4**n* LTMS cells. The output of the first LSTM layer is then propagated to the second LSTM layer with half the size of the first one, then to a dense layer. For classification models a sigmoid activation function is used in the final dense layer. However, for regression models a linear activation function is used in the final dense layer. Training the RNN classifier is performed using Adam optimizer (Kingma and Ba 2015) with a binary cross entropy as the loss function. However, regression model training is performed using root mean square prop optimizer (RMSprop) (Tieleman and Hinton 2012) with mean squared error (MSE) as the loss function. For comparison performance an ANN was constructed for the classification model by replacing LSTM layers in RNN model with dense layers with 4**n* and 2**n* units in the first and second layers, respectively.

5.5.3.1 Cross validation and Testing

Training was performed using 25 epochs and batch size of 128. 5-fold cross validation was performed and for each fold performance metrics were measured and recorded. Stratified 5-folds (Forman and Scholz 2010) were used for RNA secondary structure dataset to preserve the percentage of samples for each class in each fold.

5.5.4 rnnXna tool implementation

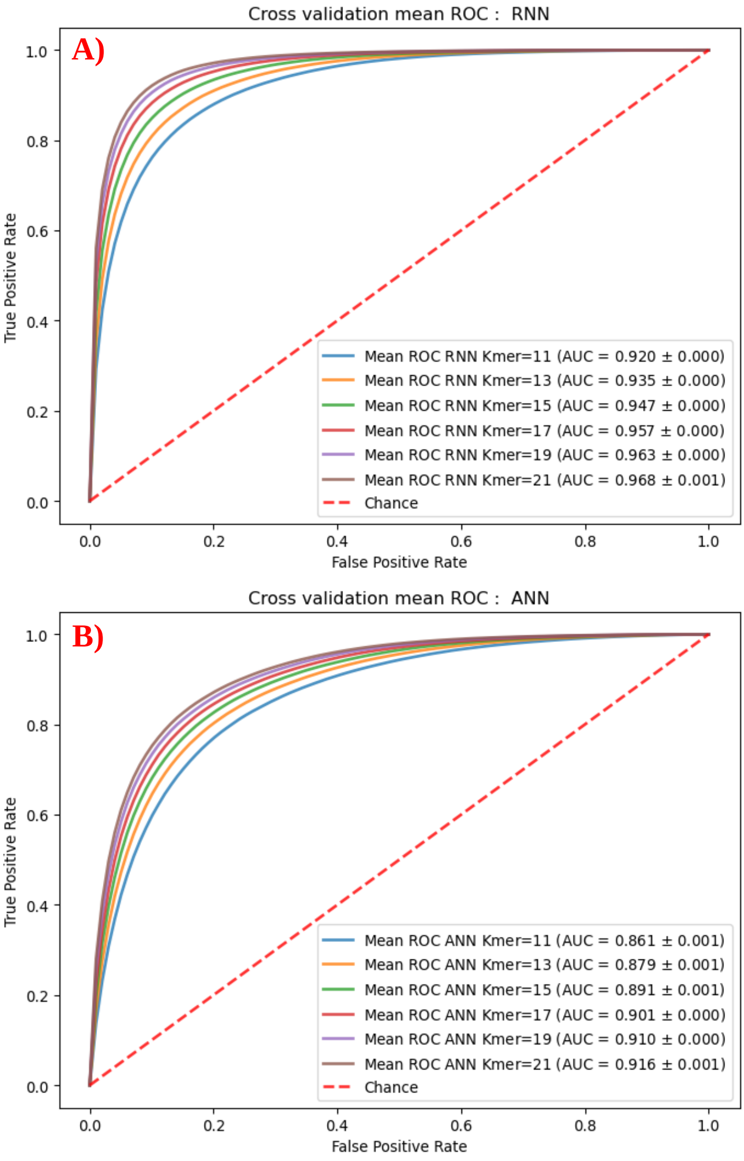
rnnXna has been implemented as a python module and is distributed with conda package manager. rnnXna is implemented using Keras (Chollet 2015) and TensorFlow (Martín *et al.* 2015) as the backend. rnnXna has a simple command line interface to train RNN models and use the trained model for further prediction. rnnXna current implementation includes RNN models introduced in this work. The tool accepts the csv/tsv format with *k*-mer fragment list and the corresponding class/score each fragment. The tool allows the user to configure a different number of parameters to train the network such as the number of layers, options to add dropout layers to overcome overfitting especially for smaller training dataset. The tool also allows the user to perform cross-validation on the training data to measure the performance before training and deploying the model in other applications. Training dataset and usage examples are available in https://github.com/Gabaldonlab/rnnXna.

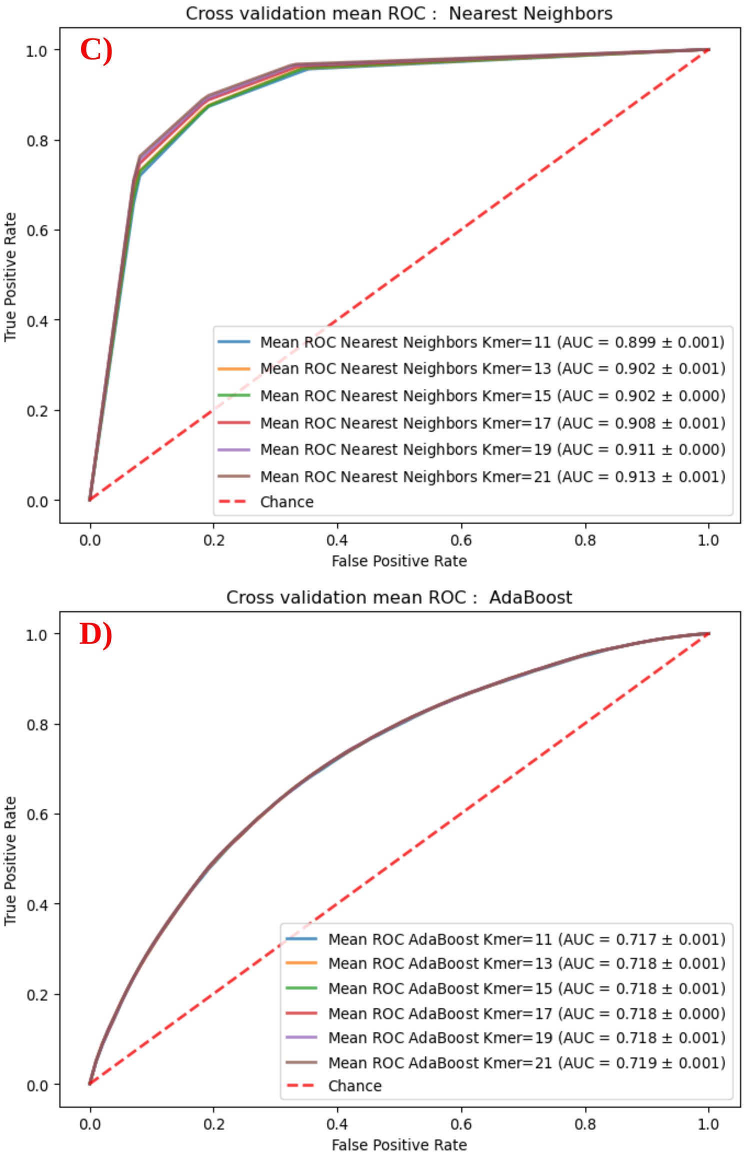
5.6 Acknowledgements

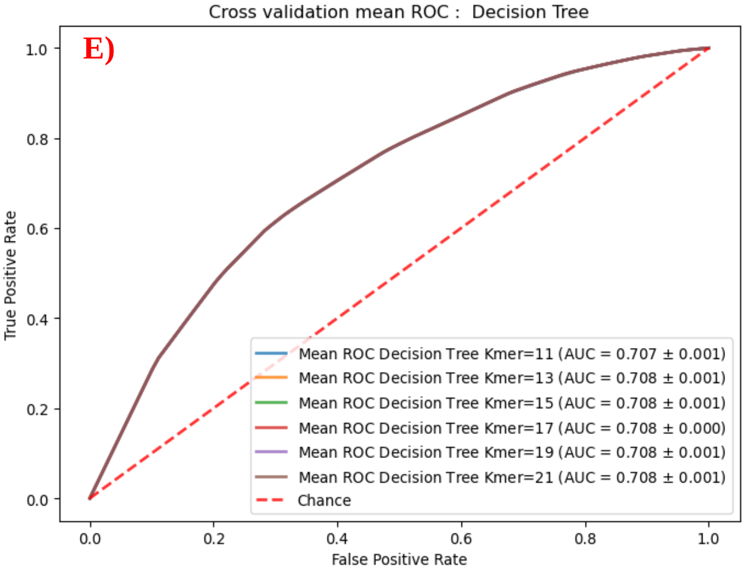
This work was supported by the European Union’s Horizon 2020 Research and Innovation programme under the Marie Sklodowska-Curie grant agreement (No. 642095) of the OPATHY consortium.

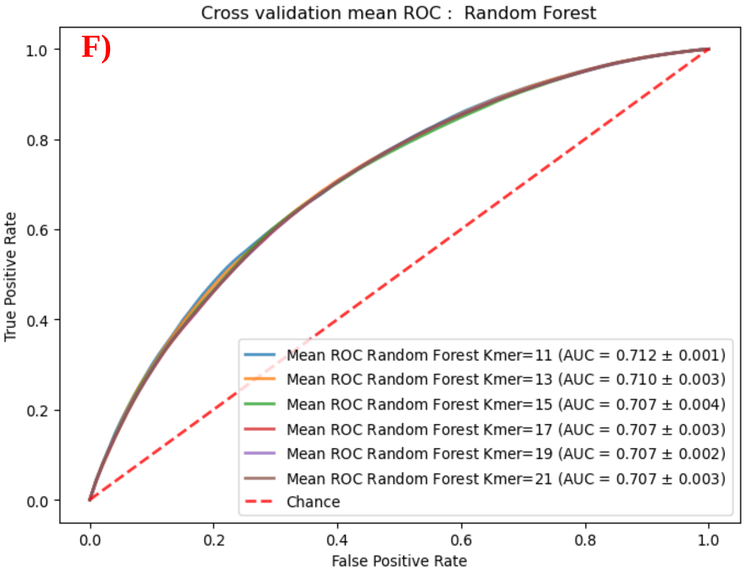
5.7 Supplementary Materials

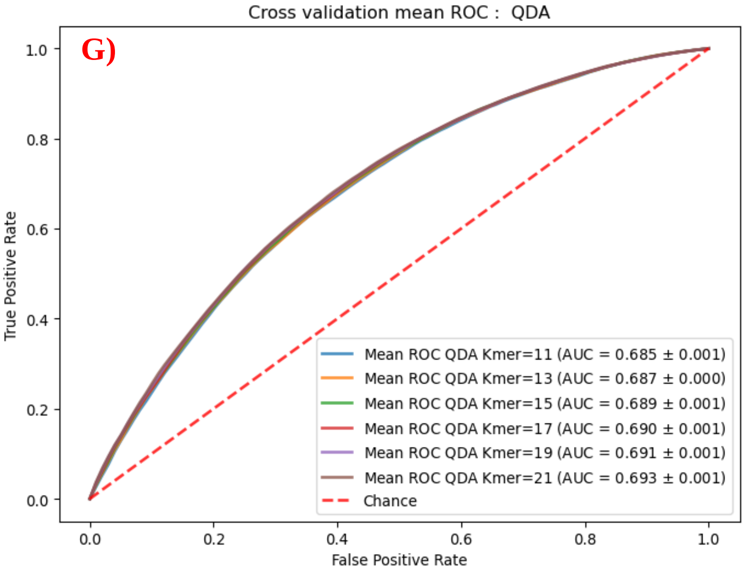
The following figure report Mean ROC and AUC for all k-classifiers for each method on cross validation.

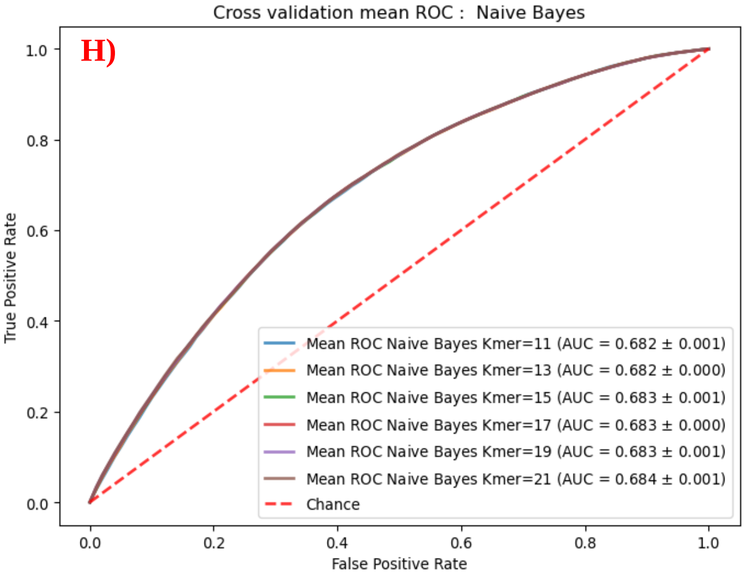



Cross validation mean ROC : Nearest Neighbors
C)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC Nearest Neighbors Kmer=11 (AUC = 0.899 ± 0.001)
Mean ROC Nearest Neighbors Kmer=13 (AUC = 0.902 ± 0.001)
Mean ROC Nearest Neighbors Kmer=15 (AUC = 0.902 ± 0.000)
Mean ROC Nearest Neighbors Kmer=17 (AUC = 0.908 ± 0.001)
Mean ROC Nearest Neighbors Kmer=19 (AUC = 0.911 ± 0.000)
Mean ROC Nearest Neighbors Kmer=21 (AUC = 0.913 ± 0.001)
Chance
Cross validation mean ROC : AdaBoost
D)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC AdaBoost Kmer=11 (AUC = 0.717 ± 0.001)
Mean ROC AdaBoost Kmer=13 (AUC = 0.718 ± 0.001)
Mean ROC AdaBoost Kmer=15 (AUC = 0.718 ± 0.001)
Mean ROC AdaBoost Kmer=17 (AUC = 0.718 ± 0.000)
Mean ROC AdaBoost Kmer=19 (AUC = 0.718 ± 0.001)
Mean ROC AdaBoost Kmer=21 (AUC = 0.719 ± 0.001)
Chance


Cross validation mean ROC : Decision Tree
E)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC Decision Tree Kmer=11 (AUC = 0.707 ± 0.001)
Mean ROC Decision Tree Kmer=13 (AUC = 0.708 ± 0.001)
Mean ROC Decision Tree Kmer=15 (AUC = 0.708 ± 0.001)
Mean ROC Decision Tree Kmer=17 (AUC = 0.708 ± 0.000)
Mean ROC Decision Tree Kmer=19 (AUC = 0.708 ± 0.001)
Mean ROC Decision Tree Kmer=21 (AUC = 0.708 ± 0.001)
Chance


Cross validation mean ROC : Random Forest
F)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC Random Forest Kmer=11 (AUC = 0.712 ± 0.001)
Mean ROC Random Forest Kmer=13 (AUC = 0.710 ± 0.003)
Mean ROC Random Forest Kmer=15 (AUC = 0.707 ± 0.004)
Mean ROC Random Forest Kmer=17 (AUC = 0.707 ± 0.003)
Mean ROC Random Forest Kmer=19 (AUC = 0.707 ± 0.002)
Mean ROC Random Forest Kmer=21 (AUC = 0.707 ± 0.003)
Chance


Cross validation mean ROC : QDA
G)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC QDA Kmer=11 (AUC = 0.685 ± 0.001)
Mean ROC QDA Kmer=13 (AUC = 0.687 ± 0.000)
Mean ROC QDA Kmer=15 (AUC = 0.689 ± 0.001)
Mean ROC QDA Kmer=17 (AUC = 0.690 ± 0.001)
Mean ROC QDA Kmer=19 (AUC = 0.691 ± 0.001)
Mean ROC QDA Kmer=21 (AUC = 0.693 ± 0.001)
Chance


Cross validation mean ROC : Naive Bayes
H)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC Naive Bayes Kmer=11 (AUC = 0.682 ± 0.001)
Mean ROC Naive Bayes Kmer=13 (AUC = 0.682 ± 0.000)
Mean ROC Naive Bayes Kmer=15 (AUC = 0.683 ± 0.001)
Mean ROC Naive Bayes Kmer=17 (AUC = 0.683 ± 0.000)
Mean ROC Naive Bayes Kmer=19 (AUC = 0.683 ± 0.001)
Mean ROC Naive Bayes Kmer=21 (AUC = 0.684 ± 0.001)
Chance

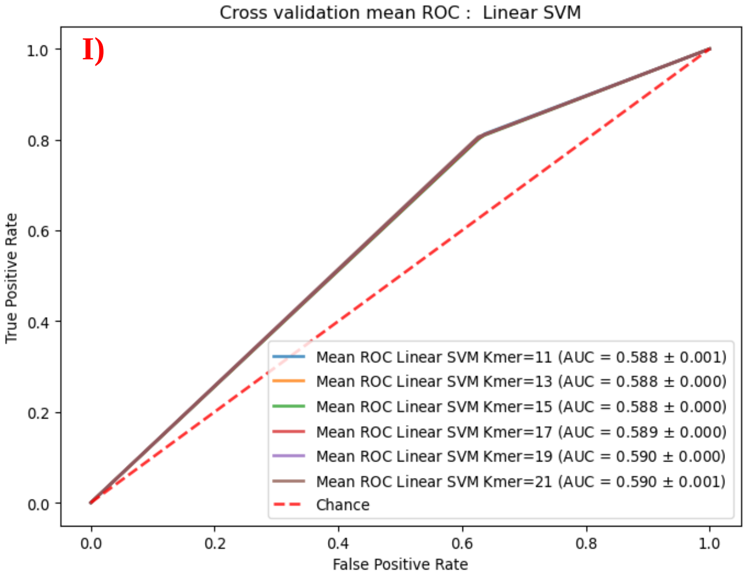


Figure 5.4 Mean ROC and AUC for all k-classifiers for each method on cross validation. True positives here are considered to be those sites that are double-stranded in the reference structure with a score greater than the given threshold, while true negatives are those sites that are single-stranded in the reference structure with a score lower than the given threshold.

6 Applications and pipeline infrastructure of the GPRO suite for RNASeq, DeNovoSeq and VariantSeq analysis

Hafez, Ahmed, Ricardo Futami, Beatriz Soriano, Francisco J. Roig, Ana Miguel, Aya A. Elsayed, Ricardo Ramos-Ruiz, Miguel A. Torres-Font, Fernando Naya-Català, Josep Calduch-Giner, Lucia Trilla-Fuertes, Angelo Gamez-Pozo, Vicente Arnau, Jaume Perez-Sánchez, Jose M. Sempere, Toni Gabaldón, Carlos Llorens. "*Applications and pipeline infrastructure of the GPRO suite for RNASeq, DeNovoSeq and VariantSeq analysis*." *(In preparation).*

6.1 Abstract

6.1.1 Summary

GPRO Suite ™ is a developing bioinformatic platform for omics data analysis that provides three client-desktop applications – *"RNASeq", "DeNovoSeq" and "VariantSeq"* - for analysis of Next Generation Sequencing data via cloud computing in remote servers. The three applications are coupled with a server-side infrastructure managing and running pipelines and tools for; i) de novo assembly and annotation; ii) differential expression and functional enrichment analysis, and iii) calling and annotation of Single Nucleotide Polymorphism (SNPs) and Insertion/Deletions (INDELs) variants. The server-side infrastructure can be installed natively on PC or a remote server or can be easily deployed using a Docker container image managed and maintained by our side. The three applications provide the users with two operational modes - *pipeline-like* or *step-by-step-like* - to manage the analyses; the first mode automatically executes all the steps considered by the selected pipeline one after another, while the other mode provides a workflow that permits the user to execute the different analytical steps analysis by analysis, or just run one or a subset of tasks of the workflow.

6.2 Availability and implementation:

"*RNASeq", "DeNovoSeq"* and *"VariantSeq"* have been developed in Java using Eclipse Rich Client Platform while the server-side infrastructure implementation involves PHP, python and bash scripting. Executable of *"RNAseq", "DeNovoSeq"* and *"VariantSeq"* are publicly available at the following URLs;

RNASeq: https://gpro.biotechvana.com/download/RNAseq

DeNovoSeq: https://gpro.biotechvana.com/download/DeNovoSeq

VariantSeq: https://gpro.biotechvana.com/download/VariantSeq

Detailed instructions to setup, configure and deploy the server-side infrastructure are available at https://gpro.biotechvana.com/tool/gpro-server/manual. A user manual is also available for each tool and for the container as well, at https://gpro.biotechvana.com.

6.3 Introduction

Advances in Next Generation Sequencing (NGS) have revolutionized the way and strategies with which researchers manage their bioinformatic tools to perform their molecular biology analyses and produce new knowledge. Currently, it is possible to reconstruct "de novo" the whole genome and the transcriptome of a given species using NGS and graph-based algorithms to assemble sequencing reads into contigs, scaffolds or chromosomes and subsequently predict and annotate genes (Miller *et al.* 2010). Once a genome or transcriptome has been annotated, it can be used as a reference sequence in comparative transcriptomics (via RNA-seq) or genotyping (mainly SNPs and INDELs) by mapping on the reference sequence the newly obtained sequencing reads from genomes or transcriptomes belonging to strains or variants from the same species (Li and Homer 2010).

Tools for NGS data analysis do not consist of a single software but rather of a complex workflow of computational steps referred to as a “pipeline” where distinct applications are sequentially or simultaneously executed. From a bioinformatics perspective, pipelines involve many different intermediate steps that interact with each other to go from raw sequencing reads to biological insights. The implementation of these protocols in routine application of NGS data analysis is still a challenge due to the technical complexity of the pipelines, whose management requires bioinformatic experts with advanced skills on the command line to manage and install third-party software, write pipeline scripts, and use a command-line interface (CLI). In an attempt to make this area of bioinformatics accessible for every researcher, distinct Graphical User Interface (GUI) solutions have arisen over the last years to provide users with friendly tools to manage pipelines for NGS data analysis. Most of these are commercial packages distributed under payment licenses and that perform multiple tasks (Smith 2015). However, these packages can be expensive and in some cases are overhyped as for their real potential. This is because while conventional GUI based solutions usually run in personal computers, NGS technologies normally require more powerful computational infrastructures (like computational servers running under Linux environments) to convincingly manage and store the huge amounts of data generated via NGS, especially if the NGS experiments are based on multiple samples with replicates. An alternative solution provided by the academy to manage server-side pipelines via GUIs, is the Galaxy Project (Afgan *et al.* 2018) which is a comprehensive platform based on a collection of web-based modules that can be combined to personalize the pipelines. Galaxy is a powerful system for managing

NGS analysis. However, it also needs advanced bioinformatic skills and some level of management and customization to reach specific analysis goals. With the aim to provide the users with a publicly available solution able to combine the generalized and versatility features of packages like Galaxy and the custom dedication usually provided by commercial GUI tools, here we present a bioinformatic solution provided by the GPRO suite (Futami *et al.* 2011) based on three client-desktop applications designated as *"DeNovoSeq", "RNAseq",* and *"VariantSeq"* which are linked to a server-side infrastructure of pipelines for processing and analysis of NGS data.

6.4 General overview and functions

"*DeNovoSeq", "RNASeq"* and *"VariantSeq"* are three client-desktop applications for execution and management of bioinformatic pipelines and workflows of NGS data analysis. The three applications are coupled with a server-side infrastructure of pipelines, tools and databases that allow users to run and manage the pipelines via user-friendly interface environments. The server-side infrastructure is distributed in a container that can be installed in the personal computers of the user (provided that the computer has enough RAM to support the required analysis and the input data) or as a cloud computing solution in remote servers (see the section below, Back-end implementation for more details). In particular, *"DeNovoSeq"* supports de novo sequences assembly and annotation approaches providing the user with GUI-access to pipelines and tools for assembly/scaffolding and annotation of new genomes and/or transcriptomes with no previous reference sequence and using both short and long NGS reads. As shown in **Figure 6.1**, *"DeNovoSeq"* allows different pipeline combinations based on a workflow of steps with three alternative paths; one oriented to predict (and then annotate) genes from contigs and scaffolds assembled from eukaryotic genomes, second to directly annotate the transcripts obtained from transcriptome assemblies and third to extract (and then annotate) open reading frames (ORFs) contig assembled from prokaryotic genomes, metagenomes and meta transcriptomes. *"RNASeq"* and *"VariantSeq"* support re-sequencing approaches in transcriptomics and genotyping; *"RNASeq"* provides GUI-access to pipelines and tools for differential expression and enrichment analyses while *"VariantSeq"* offers pipelines and tools for calling and annotation of SNPs and INDELs.

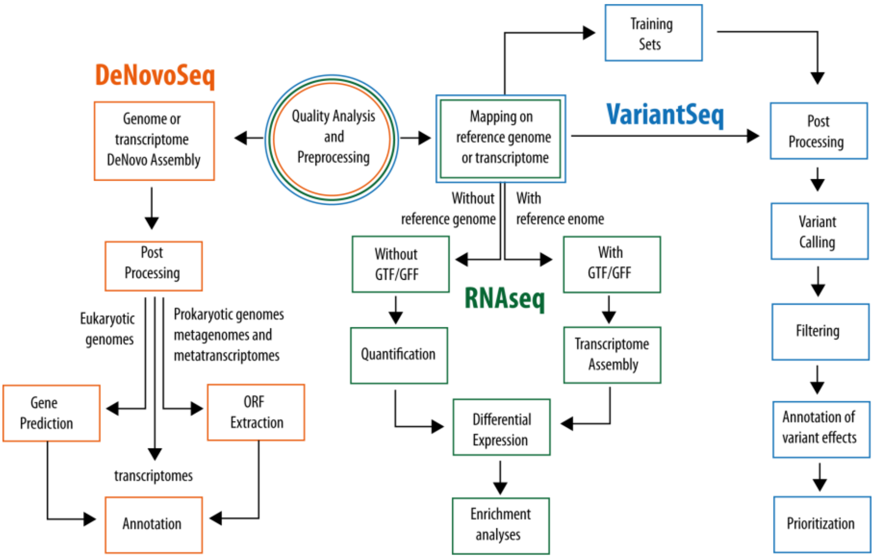


Figure 6.1 Workflow steps conducted by the server-side infrastructure of the GPRO suite for the "*RNASeq", "DeNovoSeq"* and *"VariantSeq"* applications. Steps highlighted in orange correspond to the workflow implemented by *"DeNovoSeq"*; in green the workflow of steps implemented by *"RNASeq"*; and in blue those implemented by *"VariantSeq"*. The step of quality analysis and preprocessing is highlighted with three colors because that step is implemented by the three applications. The step of "mapping on reference genome or transcriptome" is highlighted green and blue because is common to "*RNASeq"* and *"VariantSeq"*. A dynamic version of the scheme is also available online at https://gpro.biotechvana.com/genie; by clicking on any step on the dynamic diagram online the user can find information of the tools and options available for that particular step and application.

As also shown in Figure 6.1,*"RNASeq"* allows different pipeline combinations based on one workflow of steps with two alternative paths; one for RNAseq analyses using reference genome with gene annotations provided in GTF/GFF file and the other based on read quantification based on transcriptomes with no available reference genome. Finally, "*VariantSeq"* offers different pipeline combinations based on the typical workflow for genome or exome variant analysis.

6.5 Infrastructure design and implementation

The GPRO infrastructure consists of two components; client-side and server-side which are described in the following subsections.

6.5.1 Client-side

"*RNAseq", "VariantSeq"* and *"DeNovoSeq"* constitute the client-side component of the GPRO suite solution for NGS data analysis which are cross-platform desktop applications implemented using Eclipse Rich Client Platform (RCP) (Beaton and McAffer 2007). The client-side applications design and implementation incorporate a shared pipeline framework to encapsulate third party tools *i.e.* task wrapper, dynamically generate GUI views for each tool, handle and generate executable scripts and enable composable pipelines. Recent advances in bioinformatics have led to the development of numerous tools dedicated to specific analyses with possible alternative tools for one analysis, implementing direct GUI for each tool or task is time consuming and further maintenance is a cumbersome task. Thus, an intermediate framework was implemented to reduce the development time and the associated cost. The framework was developed to enable fast tasks prototyping following modular software architecture patterns. Figure 6.2A shows a basic schematic representation of the framework internal design and architecture. The pipeline framework enables application developers to write an abstract task representation of each third party tool and a corresponding task's handler with a CLI template responsible to execute the task, these templates are then translated to bash scripts to be executed on the server using plain engine components for representing bash commands. In addition, the pipeline framework enables complex workflows to be easily composed by adding single tasks together and configuring input and output data flow between different tasks, then GUIs and pipeline scripts for the workflow can be generated. The pipeline framework provides several advantages including extensibility, composable pipelines and modular representation. In addition, the framework provides extensive tracking and logging of running jobs to enable basic error reporting.

The pipeline framework is used to develop the client software applications RNASeq, VariantSeq and DeNovoSeq. The three applications present a common layout (shown Figure 6.2B) consisting of the following elements:

- Directory Browser to set any folder of the user' PC as a main directory to manage and store files in the application
- FTP Browser which is a File Transfer Protocol (FTP) letting the user to access the server-side account and to transfer files/folders from the directory browser to the server-side infrastructure or vice versa.
- Working space to call the distinct GUIs provided by each tool.
- Main Menu at the top of the layout with distinct tabs to select the protocol mode, to track the distinct pipeline jobs running or finished and/or for configuring access to the server
- Tasks Menu accessible only when selecting the *Step-by-Step-like* protocol mode, explained below.

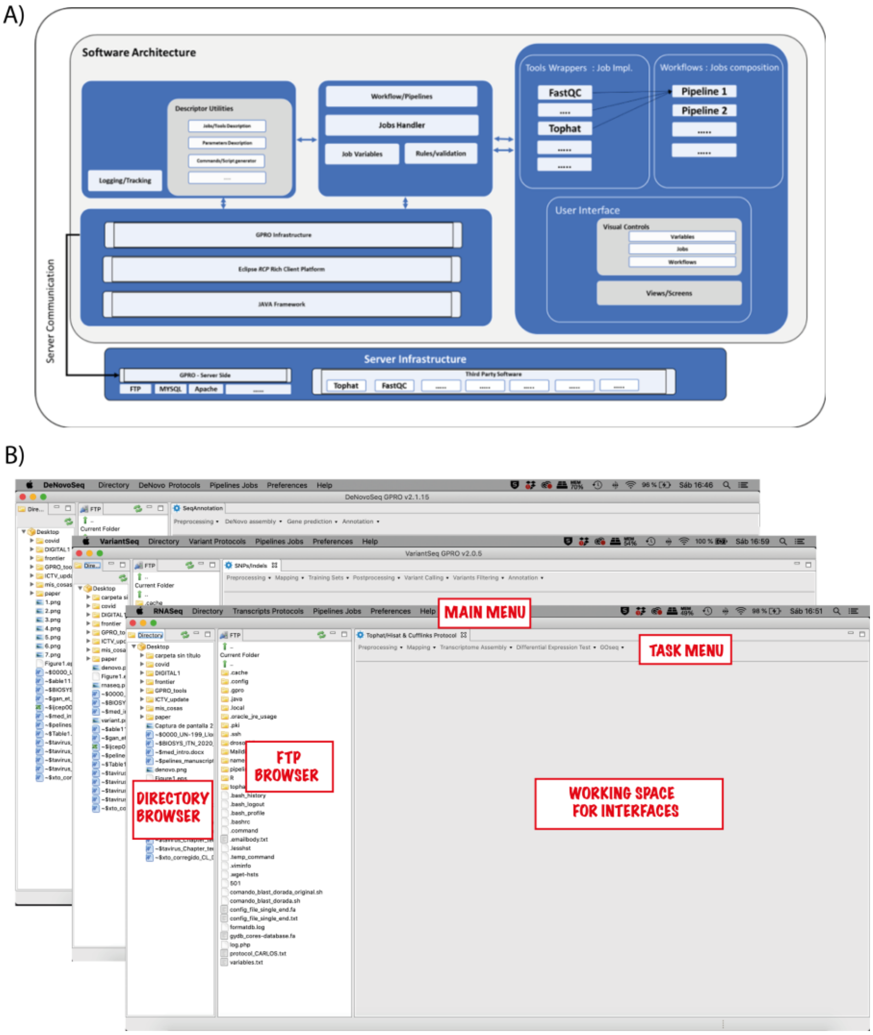

A)
Software Architecture
Descriptor Utilities
Jobs/Tools Description
Parameters Description
Commands/Script generator
Logging/Tracking
Workflow/Pipelines
Jobs Handler
Job Variables
Rules/validation
GPRO Infrastructure
Eclipse RCP Rich Client Platform
JAVA Framework
Tools Wrappers : Job Impl.
FastQC
Tophat
Workflows : Jobs composition
Pipeline 1
Pipeline 2
User Interface
Visual Controls
Variables
Jobs
Workflows
Views/Screens
Server Communication
Server Infrastructure
GPRO - Server Side
FTP
MYSQL
Apache
Third Party Software
Tophat
FastQC
B)
DeNovoSeq Directory DeNovo Protocols Pipelines Jobs Preferences Help
DeNovoSeq GPRO v2.1.15
SeqAnnotation
Preprocessing • DeNovo assembly • Gene prediction • Annotation •
VariantSeq Directory Variant Protocols Pipelines Jobs Preferences Help
VariantSeq GPRO v2.0.5
SNPs/Indels
Preprocessing • Mapping • Training Sets • Postprocessing • Variant Calling • Variants Filtering • Annotation •
RNASeq Directory Transcripts Protocols Pipelines Jobs Preferences Help
Tophat/Hisat & Cufflinks Protocol
Preprocessing • Mapping • Transcriptome Assembly • Differential Expression Test • GOseq •
MAIN MENU
TASK MENU
FTP BROWSER
DIRECTORY BROWSER
WORKING SPACE FOR INTERFACES

Figure 6.2 A) Internal framework design and architecture. B) Layout of "*DeNovoSeq*", "*VariantSeq*" and "*RNASeq*". The three applications present a common layout consisting of the following components: DIRECTORY BROWSER, FTP BROWSER, WORKING SPACE, MAIN MENU and TASK MENU. The interfaces communicate to the server dependencies via HTTP/SSH. The server-side handles and manage all dependencies (tools and databases) to execute pipelines for i) de novo assembly and annotation; ii) differential expression and enrichment, and iii) calling and annotation of SNPs and INDELs.

The default layout is adjustable, and users can arrange all components as they need. It is possible to drag and drop files between directory browser and FTP browser to start file transfer (upload/download) between local PC and the remote server, which is a conventional way of managing remote server files.

6.5.2 Server-side

The server-side contains all installed third-party software used in the client-side software and management system to handle users accounts, pipelines, databases, tracking and logging. As shown in Table 6.1, the server-side dependencies for implementation of pipelines are free or open-source software usually considered as the State-Of-The-Art (or the most popular) in the field. The server-side requires complex steps to setup Linux, Apache, MySQL, and PHP (LAMP stack), manual installation to third party software, and GPRO server-side application and scripts to handle incoming requests to run and manage third party tools and pipelines. For detailed instructions on how to get the server ready and running, one can refer to https://gpro.biotechvana.com/tool/gpro-server/manual. For faster deployment, we provide a docker image (Merkel 2014) ready to run with the GPRO server application. To run, first setup and configure the docker on a local PC or server. Then run following command

```
# One click Command to deploy and run GPRO Server
> docker run -d -p 80:80 -p 20-22:20-22 -p 65500-
65515:65500-65515 -v /path/to/local_home:/home/gpro_user
biotechvana/gpro
```

See the online manual for further configuration of the docker container https://gpro.biotechvana.com/tool/gpro-server/manual.

Table 6.1 Server dependencies for GPRO server-side infrastructure.

Pipeline Step	**Tools**	**RNAseq**	**DeNovoSeq**	**VariantSeq**
Quality analysis and Preprocessing	FastQC (Andrews and Others 2010)	✓	✓	✓
	FastqMidCleaner	✓	✓	✓
	Cutadapt (Martin 2011)	✓	✓	✓
	Prinseq (Schmieder and Edwards 2011)	✓	✓	✓
	Trimmomatic (Bolger *et al.* 2014)	✓	✓	✓
	FastxToolkit (Gordon and Hannon 2017)	✓	✓	✓
	CANU (Koren *et al.* 2017)		✓	
	FastqCollapser	✓	✓	✓
	FastqIntersect	✓	✓	✓
Mapping on reference genome or transcriptome	TopHat (Kim *et al.* 2013)	✓		✓
	Hisat2 (Kim *et al.* 2015)	✓		✓
	Bowtie2 (Langmead and Salzberg 2012)	✓		✓
	BWA (Li and Durbin 2009)(Li and Durbin, 2009)	✓		✓
	STAR (Dobin *et al.* 2013)			✓
Quantification	Corset (Davidson and Oshlack 2014)	✓		
	Htseq (Anders *et al.* 2015)	✓		
Post Processing	GATK (McKenna *et al.* 2010)			✓
	Picard (Wysoker *et al.* 2013)			✓
	SAMtools (Li *et al.* 2009)	✓		✓
Transcriptome assembly	Cufflinks (Trapnell *et al.* 2012)	✓		
	Oases (Schulz *et al.* 2012)		✓	
	SOAPdenovo-trans (Xie *et al.* 2014)		✓	
	Velvet (Zerbino and Birney 2008)		✓	

Pipeline Step	**Tools**	**RNAseq**	**DeNovoSeq**	**VariantSeq**
Genome assembly	SOAPdenovo2 (Luo *et al.* 2012)		✓	
	CANU (Koren *et al.* 2017)		✓	
	SPAdes (Bankevich *et al.* 2012)		✓	
Post Processing - Gap filling	Gap closer (Luo *et al.* 2012)		✓	
Post Processing - Scaffolding	BESST (Sahlin *et al.* 2014)		✓	
	OPERA (Gao *et al.* 2011)		✓	
Differential expression	DESeq (Love *et al.* 2014)	✓		
	EdgeR (Robinson *et al.* 2010)	✓		
	Cuffdiff (Trapnell *et al.* 2012)	✓		
Enrichment Analysis	GOseq (Young *et al.* 2010)	✓		
Gene Prediction	Augustus (Stanke *et al.* 2008)		✓	
Annotation	BLAST (Altschul *et al.* 1990)		✓	
	HMMER (Mistry *et al.* 2013)		✓	
	InterProScan (Jones *et al.* 2014)		✓	
Annotation of variant effects	Variant Effect Predictor (McLaren *et al.* 2016)			✓
Training Sets	GATK (McKenna *et al.* 2010)			✓
Variant Calling	GATK (McKenna *et al.* 2010)			✓
	VarScan2 (Koboldt *et al.* 2012)			✓
Variant Filtering	GATK (McKenna *et al.* 2010)			✓

✓ indicates if a tool is included in an application

6.6 Usage modes

The three applications provide the users with two operational modes -*pipeline-like* or *step-by-step-like* - to manage the analyses. The *pipeline* mode gives access to a GUI where the user can upload input data, select the appropriate pipeline from a list, prepare a configuration file with a series of parameters and instructions, declare the output, and then, run the pipeline. In doing so, the *pipeline* GUI will automatically execute sequentially all the steps considered by the selected pipeline using the data declared as input and the parameters and options declared in the configuration file. As an example, in Figure 6.3, we show the GUI of *"RNAseq"* for executing an analysis in pipeline mode. Please note that each application presents a specific GUI a particular configuration file for executing the pipeline mode. For specific details and tutorials about the pipeline mode execution, please refer to the respective manual of each application.

On the other hand, the *step-by-step* mode provides a working mode permitting users to execute the different steps of a workflow one after another (*i.e*. analysis by analysis) managed manually the distinct interfaces of each tool. Notice that by accessing the step-*by-step* mode, a task menu will appear organizing the workflow tools as tabs at the top of the working space in order that reflect the logical stages of the selected protocol. By clicking on each tab of the task menu a scroll down will appear permitting the user to choose from a list a particular tool to perform the desired task. Each tool provided by the *step-by-step* mode is presented to the user with a specific GUI presenting distinct fields to manage the input and configure output and optional parameters of the analysis. As an example, the GUI of *"RNAseq"* application for executing the step of transcriptome assembly with cufflinks is shown in Figure 6.4. For specific details and tutorials about each GUI per tool provided by the step-by-step mode in each application *("RNAseq", "DeNovoSeq"* and *"VariantSeq")* please refer the respective manual of each application.

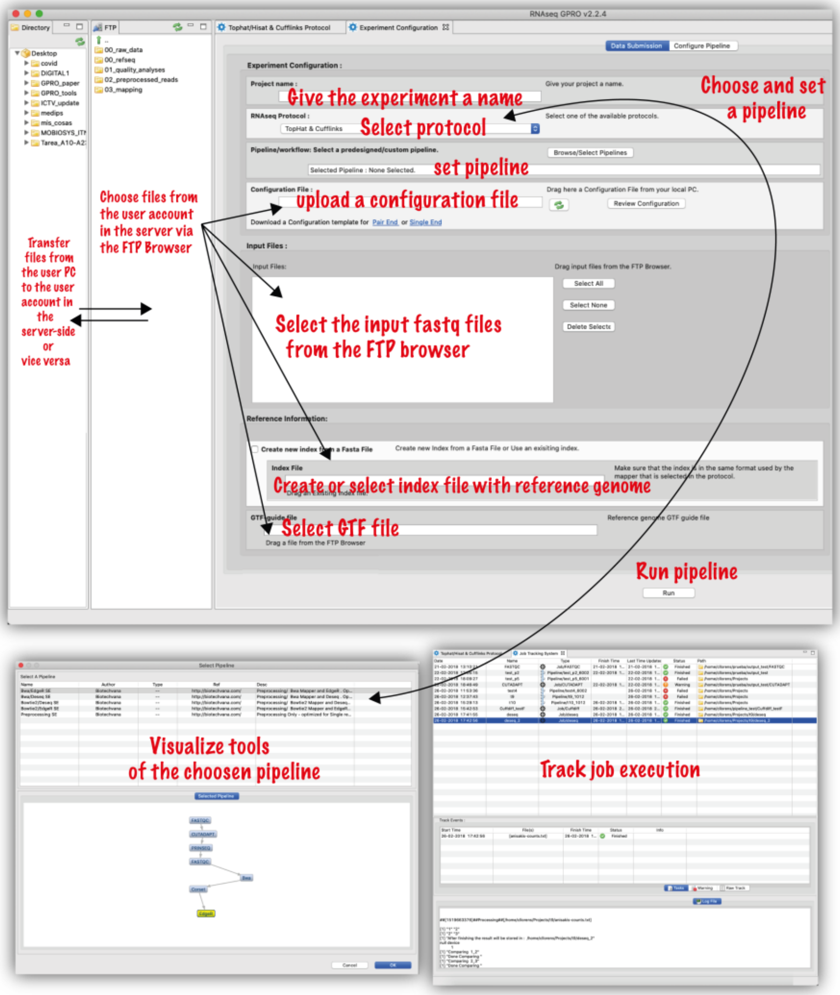
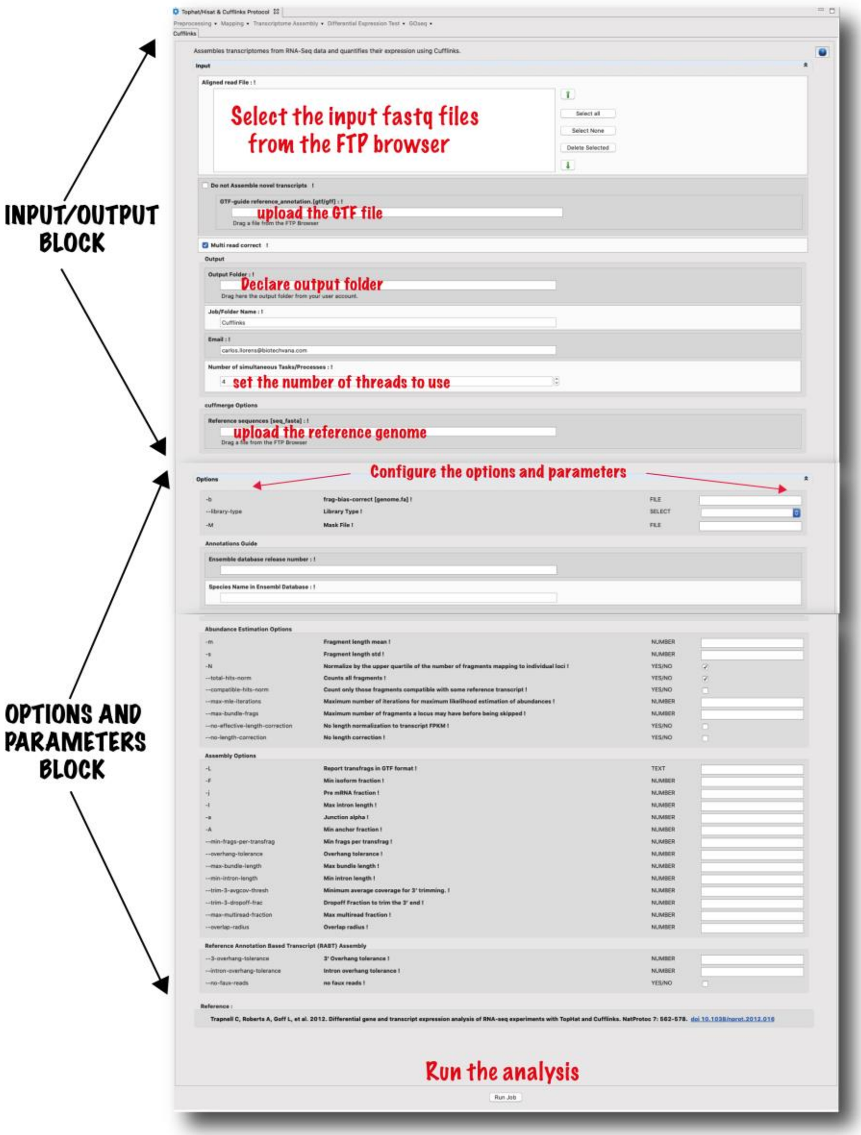

RNAseq GPRO v2.2.4
Directory
FTP
Tophat/Hisat & Cufflinks Protocol
Experiment Configuration
Data Submission
Configure Pipeline
Experiment Configuration :
Project name :
Give your project a name.
Give the experiment a name
Choose and set a pipeline
RNAseq Protocol :
TopHat & Cufflinks
Select protocol
Select one of the available protocols.
Pipeline/workflow: Select a predesigned/custom pipeline.
Browse/Select Pipelines
Selected Pipeline : None Selected.
set pipeline
Configuration File :
upload a configuration file
Drag here a Configuration File from your local PC.
Review Configuration
Download a Configuration template for Pair End or Single End
Choose files from the user account in the server via the FTP Browser
Transfer files from the user PC to the user account in the server-side or vice versa
Input Files :
Input Files:
Drag input files from the FTP Browser.
Select All
Select None
Delete Selects
Select the input fastq files from the FTP browser
Reference Information:
Create new index from a Fasta File
Create new Index from a Fasta File or Use an exisiting index.
Index File
Make sure that the index is in the same format used by the mapper that is selected in the protocol.
Create or select index file with reference genome
Drag an Existing Index file.
GTF guide file
Select GTF file
Reference genome GTF guide file
Drag a file from the FTP Browser
Run pipeline
Run
Desktop
covid
DIGITAL1
GPRO_paper
GPRO_tools
ICTV_update
medips
mis_cosas
MOBIOSYS_IT
Tarea_A10-A2
00_raw_data
00_refseq
01_quality_analyses
02_preprocessed_reads
03_mapping
Select Pipeline
Select A Pipeline
Visualize tools of the choosen pipeline
Selected Pipeline
FASTQC
CUTADAPT
PRINSEQ
FASTQC
Bwa
Convert
EdgeR
Cancel
OK
Job Tracking System
Track job execution
Track Events
Log File

Figure 6.3 Top, interface of "*RNASeq*" for pipeline mode execution which is based on a single form where the user configures the analysis by declaring the input and output as well as a configuration file with the parameter's configuration. The user can choose a pipeline from a summary (interface on the left below) and finally runs all the pipeline tasks automatically one after another. The user can also monitor the workflow and the results in a job tracking interface (interface on the right below) available via the main menu.



INPUT/OUTPUT BLOCK
OPTIONS AND PARAMETERS BLOCK
Select the input fastq files from the FTP browser
upload the GTF file
Declare output folder
set the number of threads to use
upload the reference genome
Configure the options and parameters
Run the analysis
Tophat/Hisat & Cufflinks Protocol
Preprocessing • Mapping • Transcriptome Assembly • Differential Expression Test • GOseq •
Cufflinks
Select all
Select None
Delete Selected
Run Job

Assembles transcriptomes from RNA-Seq data and quantifies their expression using Cufflinks.

Input

Aligned read File : !

□ Do not Assemble novel transcripts !

GTF-guide reference_annotation.[gtf/gff] : !

Drag a file from the FTP Browser

☑ Multi read correct !

Output

Output Folder : !

Drag here the output folder from your user account.

Job/Folder Name : ! Cufflinks

Email : ! carlos.llorens@biotechvana.com

Number of simultaneous Tasks/Processes : ! 4

cuffmerge Options

Reference sequences [seq_fasta] : !

Drag a file from the FTP Browser

Options

-b	frag-bias-correct [genome.fa] !	FILE	
--library-type	Library Type !	SELECT	
-M	Mask File !	FILE	

Annotations Guide

Ensemble database release number : !

Species Name in Ensembl Database : !

Abundance Estimation Options

-m	Fragment length mean !	NUMBER	
-s	Fragment length std !	NUMBER	
-N	Normalize by the upper quartile of the number of fragments mapping to individual loci !	YES/NO	☑
--total-hits-norm	Counts all fragments !	YES/NO	☑
--compatible-hits-norm	Count only those fragments compatible with some reference transcript !	YES/NO	□
--max-mle-iterations	Maximum number of iterations for maximum likelihood estimation of abundances !	NUMBER	
--max-bundle-frags	Maximum number of fragments a locus may have before being skipped !	NUMBER	
--no-effective-length-correction	No length normalization to transcript FPKM !	YES/NO	□
--no-length-correction	No length correction !	YES/NO	□

Assembly Options

-L	Report transfrags in GTF format !	TEXT	
-F	Min isoform fraction !	NUMBER	
-j	Pre mRNA fraction !	NUMBER	
-I	Max intron length !	NUMBER	
-a	Junction alpha !	NUMBER	
-A	Min anchor fraction !	NUMBER	
--min-frags-per-transfrag	Min frags per transfrag !	NUMBER	
--overhang-tolerance	Overhang tolerance !	NUMBER	
--max-bundle-length	Max bundle length !	NUMBER	
--min-intron-length	Min intron length !	NUMBER	
--trim-3-avgcov-thresh	Minimum average coverage for 3' trimming. !	NUMBER	
--trim-3-dropoff-frac	Dropoff Fraction to trim the 3' end !	NUMBER	
--max-multiread-fraction	Max multiread fraction !	NUMBER	
--overlap-radius	Overlap radius !	NUMBER	

Reference Annotation Based Transcript (RABT) Assembly

--3-overhang-tolerance	3' Overhang tolerance !	NUMBER	
--intron-overhang-tolerance	Intron overhang tolerance !	NUMBER	
--no-faux-reads	no faux reads !	YES/NO	□

Reference :

Trapnell C, Roberts A, Goff L, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. NatProtoc 7: 562-578. doi 10.1038/nprot.2012.016

Figure 6.4 Interface of "RNAseq" for executing the step of transcriptome assembly using Cufflinks. The interface is divided in two blocks. One block provides the user with a form for uploading the input files and declaring the output folder. The second block permits the user to configure the analysis using all the parameters provided by Cufflinks for the step of transcriptome assembly.

6.7 Acknowledgements

This work was supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642095 for the OPATHY consortium, by the pre-doctoral research fellowship from Industrial Doctorates of MINECO (Grant 659 DI-17-09134); by the State Plan for Scientific and Technical Research and Innovation 2017-2020 under the Grant TSI-100903-2019-11 from the Secretary of State for Digital Advancement of Ministry of Economic Affairs and Digital Transformation, Spain; and by Spanish MICIUN Project Bream-AquaINTECH: From Nutrition and Genetics to Sea Bream Aquaculture Intensification and Technological Innovation (RTI2018-094128-B-I00).

7 Biomarkers of caspofungin resistance in *C. albicans* isolates: a proteomic approach

Buda De Cesare, Giuseppe, **Ahmed Hafez**, David Stead , Carlos Llorens and Carol A. Munro "*Biomarkers of caspofungin resistance in C. albicans isolates: a proteomic approach*." *(In preparation)*.

7.1 Abstract

7.1.1 Summary

Candida albicans is a clinically important polymorphic fungal pathogen that causes life-threatening invasive infections in immunocompromised patients. Antifungal therapy failure is a substantial clinical problem, due to the emergence of an increasing number of drug resistant isolates. Caspofungin is a common antifungal drug, often used as first line therapy, that inhibits cell wall β-(1,3)-glucan synthesis. In this work, the cell surface of different echinocandin-resistant *C. albicans* clinical isolates was compared with sensitive isolates and their responses to echinocandin treatment analysed. Protein analysis by LC-MS/MS detected changes in the repertoire of proteins involved in cell wall organization and maintenance in drug-resistant strains in comparison to susceptible isolates, after incubation with caspofungin. Moreover, with this data set, a differential expression analysis was performed, and an interaction network was created in order to identify useful biomarkers of echinocandin resistance. These results suggest drug resistance may involve not only a different cell wall architecture, but also a different response to drugs. Therefore, the identified protein subsets can be potentially used for a rapid diagnosis of drug resistance in clinical settings.

7.1.2 Importance

Candida albicans is the most common fungal species in clinical settings. In healthy subjects, it is usually harmless in perfect balance with the host microbiota, but under particular conditions that impair the host's immune system, such as stress, infections or immunosuppressant therapies, the fungus can escape from its normal commensal niches and invade the body causing disease. The alteration of the host environment (such as pH, immune response, and nutrients) induces a change in morphology from the oval yeast-form to filamentous hyphal cells, which facilitates invasion of host cells. The entire scenario is complicated by the inclination of some strains to form drug-resistant biofilms, which often correlates with higher mortality rates in bloodstream infections. Echinocandins, and in particular caspofungin, are the most used antifungal drugs for treatment of infections caused by *Candida* species in clinical settings. An early diagnosis of drug-resistant infections is important for an effective antifungal treatment.

7.2 Introduction

Candida albicans is a human opportunistic pathogen, able to switch from commensalism to pathogenicity in response to different cues from the host niches it colonizes (Dutton *et al.* 2016, Kumamoto 2011, Achkar and Fries 2010). A primary determinant of such a switch is the cell wall, which plays key roles in pathogenicity and interactions with host defences (Gow and Hube 2012, Hernández-Chávez *et al.* 2017). The dynamic structure of the cell wall mainly consists of three polysaccharides: chitin, glucan and mannan. Fundamentally, fungal cell walls are layered structures, with an inner conserved core of chitin and glucan and an outer layer of polysaccharides, variable according to the fungus. In particular, the wall of *C. albicans* consists of three different polysaccharides, chitin, β-(1,3)- and β-(1,6)-glucan. Cell wall proteins, often highly mannosylated, form the outermost layer (Plaine *et al.* 2008, Nather and Munro 2008). The major class of cell wall proteins are covalently attached to β-(1,6)-glucan by modified Glycosyl Phosphatidyl Inositol (GPI) anchors.

Echinocandins are a class of antifungal agents, available for more than a decade, which are recommended as first line treatment for many types of *Candida* infections (Cornely *et al.* 2012, Hope *et al.* 2012). These drugs affect the behaviour of the cell wall, and in particular the exposure of β-(1,3)-glucan on the surface and the phagocytosis from macrophages (Walker and Munro 2020). The fungicidal activity of this class of drugs has been shown *in vitro* to inhibit the activity of β-(1,3)-glucan synthase, the enzyme required for the synthesis of the glucan layer found in most medically important fungi (Douglas 2001). The reduced susceptibility to echinocandins has been primarily attributed to point mutations in *GSC1/FKS1* gene that encodes the catalytic subunit of the glucan synthase complex, which can decrease sensitivity to the drug by several log orders (Garcia-Effron, Park, *et al.* 2009, Garcia-Effron, Lee, *et al.* 2009). The compromised integrity of the wall has an effect also on the properties of the plasma membrane, as shown by (Kelly and Kavanagh 2010). They showed caspofungin treatment increased the permeability of the wall, with increased amino acid leakage from the treated cells compared to DMSO-treated cells, which had previously been reported to alter membrane permeability (Yu and Quinn 1998). Protein release was also increased in caspofungin-treated cells, and in particular release of cytoplasmic proteins involved in metabolic pathways (Pgk1, Gpm1, Fba1 and Eno1). The released proteins are highly immunogenic (Ahmed *et al.* 2015, Mundodi *et al.* 2008), with a possible role in immune response and inflammation *in vivo*.

The diagnostic “gold standard” for detecting *Candida* infections is blood culture, with other additional tests based on the detection of circulating polysaccharides from the fungal cell wall or antigens in blood samples (Chumpitazi *et al.* 2014, Prella *et al.* 2005). However, the diagnosis of antifungal resistance is an issue in the clinical settings, as it is not routinely performed and requires 48-72 hours, which is often too late to influence the treatment of the patient (Zhao and Perlin 2014).

Interaction networks have been created using transcriptomic data and applied to study host-pathogen interactions (Amorim-Vaz *et al.* 2015, Remmele *et al.* 2015) and to identify genes involved in filamentation response (Martin *et al.* 2013).

In this study, we aimed to characterize the cell wall of caspofungin-resistant isolates and to compare them with sensitive isolates, in order to find differences, between the groups, which may be helpful in diagnosing drug resistance in *C. albicans* isolates. We evaluated protein expression levels by mass-spectrometry, specifically for enzymes involved in cell wall synthesis and maintenance during drug treatment. We performed a differential expression (DE) analysis with this proteomics data set, and an interaction network was created to look at proteins associated with the cell wall that were differentially expressed in the strains analysed. The strains were also evaluated for the capacity to form biofilms, an important virulence attribute that contributes to drug-recalcitrant infections. The set of proteins identified by the DE analysis can help to elucidate the responses of the cell wall to echinocandin drugs. This information can be utilised to develop a rapid diagnostic assay for clinical use.

7.3 Results

7.3.1 Strain characterisation

The susceptibility of nine isolates of *C. albicans* to caspofungin (CAS) was evaluated by the broth microdilution method. The minimum inhibitory concentrations of caspofungin are shown in Table 7.1, six isolates were detected as susceptible (Ca1, Ca2, Ca3, Ca4, Ca5 and Ca6) and three as resistant (Car1, Car2 and Car3) according to CLSI breakpoints (Alexander 2017). The Ca2, Ca3, Ca4 and Ca6 isolates had the same range of IC_{50} despite coming from different sources.

Table 7.1 Caspofungin MIC against *C. albicans* isolates.

Strain ID	Name	IC_{50} CAS (µg/ml)	S/I/R [a]	Genotype	Reference
SC5314	Ca1	0.03-0.06	S	*Wild type*	(Fonzi and Irwin 1993)
CBS8758	Ca2	0.06-0.125	S	*Wild type*	(Gillum *et al.* 1984)
ATCC2091	Ca3	0.06-0.125	S	*Unknown*	(Morrison *et al.* 1993)
ATCC76615	Ca4	0.06-0.125	S	*Unknown*	(Thanos *et al.* 1996)
B17_009053	Ca5	0.125-0.25	S	*Unknown*	Munich, unpublished
B17_008835	Ca6	0.06-0.125	S	*Unknown*	Munich, unpublished
K063-3	Car1	2	R	*GSC1 (S645Y)/ GSC1 (S645Y)*	(Lee *et al.* 2012)
B15_004476	Car2	2-4	R	*Unknown*	(Munich, unpublished)
B12_007355_1	Car3	1-2	R	*GSC1 (R1361G)/GSC 1 (R1361G)*	(Munich, unpublished)

The caspofungin IC50 was calculated after 24h incubation in RPMI-1640 medium in a broth microdilution method according to the CLSI breakpoints.

[a] Interpretive category according to the breakpoints: S=susceptible, I=intermediate, R=resistant

Candida spp. infections are often associated with biofilm and biomaterial-related infections (Crump and Collignon 2000). One of the features of biofilms is their reduced antimicrobial susceptibility compared to planktonic cells (Crump and Collignon 2000, Sardi *et al.* 2013). In order to assess possible implications of drug resistance in *in vivo* infections, the biofilm formation capacity of the isolates was evaluated. Cells were grown for 6, 24, 48 and 72 hours in the absence of drug, or with 2 or 4 µg/ml caspofungin, at 37 ºC in RPMI-1640 medium, with 20 % FCS added, which has been shown to increase biofilm formation (Frade and Arthington-Skaggs 2011). Biomass was then measured through crystal violet absorbance at 570 nm (Figure 7.1, Figure **7.2**). The reference strain of *C. albicans* Ca1 steadily increased biofilm formation throughout the 72 h in absence of drug, whereas the addition of 2 or 4 µg/ml caspofungin totally abolished biofilm development (Figure 7.1). In general, the biofilm biomass of the caspofungin-susceptible isolates was comparable to Ca1 strain (Figure 7.1), with no substantial biomass increase at 72 h (Figure 7.1D). The addition of either 2 or 4 µg/ml caspofungin, instead, totally abolished the formation of biofilms for all the caspofungin-susceptible isolates (Figure 7.1A-C, yellow and cyan columns), except for some biomass increase observed at 72 h time point (Figure 7.1D). The three resistant isolates, except for the 6 h time point (Figure 7.2A), showed different trends in presence or in absence of the antifungal (Figure 7.2B-D). Without caspofungin (Figure 7.2, magenta columns), Car1 isolate produced biofilm similar to its isogenic parental isolate Ca1, whereas the other two resistant isolates did not significantly increase their biomass after the first 6 h time point (Figure 7.2B-D). The same situation was observed in presence of drug (Figure 7.2, yellow and cyan columns), with Car1 able to reach the same levels of biofilm biomass observed in the absence of drug (Figure 7.2, magenta columns). Car2 and Car3 also had similar biofilm growth under the no drug and drug conditions.

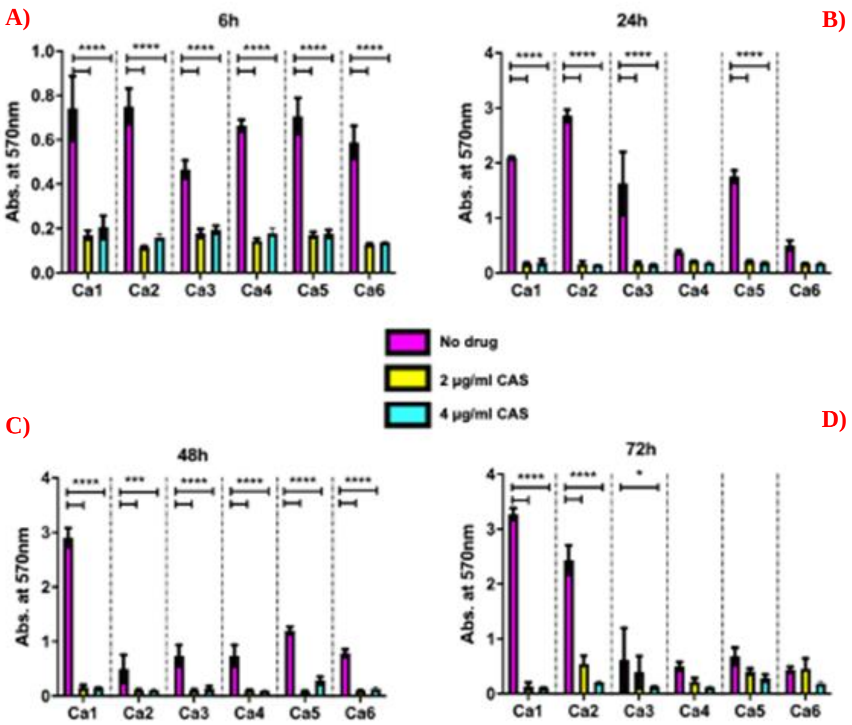


Figure 7.1 Biofilm formation by caspofungin-susceptible isolates of *C. albicans*. Absorbance from crystal violet staining of *C. albicans* Ca1, Ca2, Ca3, Ca4, Ca5, Ca6 isolates grown at 37ºC in RPMI-1640 + 20% FCS grown either without drug (magenta) or the addition of 2 µg/ml CAS (yellow) or 4 µg/ml CAS (cyan) and measured at: (A) 6h; (B) 24h; (C) 48h; (D) 72h. The values are expressed as absorbance at 570nm. The statistical analysis performed was one-way ANOVA (n = 1 (3 replicates), *P<0.05, **P<0.005, ***P<0.0005, ****P<0.00005)

7.3.2 Proteomic response to caspofungin

To highlight the differences in the cell wall proteome of *C. albicans* isolates, the cell walls of the isogenic strains Ca1 (caspofungin-susceptible) and Car1 (caspofungin-resistant) were extracted according to a modified protocol by (Kapteyn *et al.* 2000). A trypsin digestion of the extracted walls was carried out and the peptides released analysed by LC-MS/MS and sequences identified according to the CGD database. Cells were grown in RPMI-1640 medium at 37 ºC and incubated for 90 min with different caspofungin concentrations based on ICs (shown in Table 7.1), in order to achieve the same percentage of growth (~80 %). The cells were incubated with the drug during the exponential phase of growth and negligible differences in growth between the strains were observed ensuring the proteome was not altered by factors other than drug effects. The analysis was carried out with Proteome Discoverer v2.2 software, using a cut-off of at least 2 peptides detected per protein, with abundances determined using Area Under the Curve measurements.

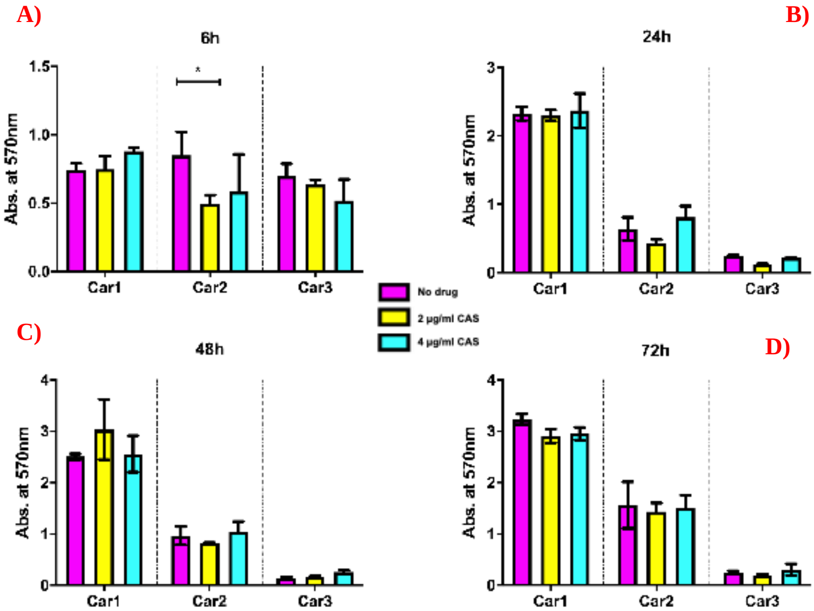


Figure 7.2 Biofilm formation assay for caspofungin-resistant isolates of *C. albicans*. Absorbance from crystal violet staining of *C. albicans* Car1, Car2 and Car3 isolates grown at 37ºC in RPMI-1640 + 20% FCS grown either without drug (magenta) or the addition of 2 µg/ml CAS (yellow) or 4 µg/ml CAS (cyan) and measured at: (A) 6h; (B) 24h; (C) 48h; (D) 72h. The values are expressed as absorbance at 570nm. The statistical analysis performed was one-way ANOVA (n = 1 (3 replicates), *P<0.05, **P<0.005, ***P<0.0005, ****P<0.00005).

Given the same genetic background, a limited set of proteins was found differentially expressed between the two isolates (Figure 7.3). Pga52 and Pga31, two GPI-anchored proteins of unknown function, were detected in increased abundances in Car1 compared to Ca1 isolate, in the presence (4.9- and 3.6-fold difference respectively) and absence of caspofungin (13.7- and 5.9-fold difference respectively). Another GPI-modified cell wall protein, Rbt5, involved in biofilm formation and iron homeostasis, was slightly more abundant in the echinocandin-resistant isolate Car1 without drug (1.7-fold difference) and with drug (5.9-fold difference). The peptides from two proteins of unknown function, C2_04780W_A and C3_07470W_A, were also found in higher amount in the drug-resistant isolate Car1 in the presence (7.2- and 4.4-fold difference respectively) and absence of echinocandin (3- and 1.4-fold difference respectively).

Other proteins were differentially expressed in the two strains based on the conditions: Slr1, involved in hyphal growth, and Msb2, cell wall damage sensor involved in activation of Cek1 phosphorylation pathway, were more abundant in Car1 upon drug incubation (2- and 2.4-fold difference respectively) and slightly less abundant without caspofungin (0.7- and 0.8-fold difference respectively). On the contrary, the cell surface 1,3-beta-glucanosyltransferase Phr2, involved in cell wall remodelling, was detected in higher amounts in Car1 in the absence of antifungal (3.4-fold difference) than with the drug (0.9-fold difference) the latter due to upregulation of this protein in Ca1 in response to drug treatment.

Four proteins were detected in lower amounts for both conditions (presence and absence of drug) in the caspofungin-resistant isolate: the phospholipase Plb5 (0.2- and 0.03-fold difference respectively); the glucose transporters Hgt7 (0.4- and 0.4-fold difference respectively) and Hgt8 (0.3- and 0.4-fold difference respectively); the glucan synthase Gsc1 (0.2- and 0.2-fold difference respectively).

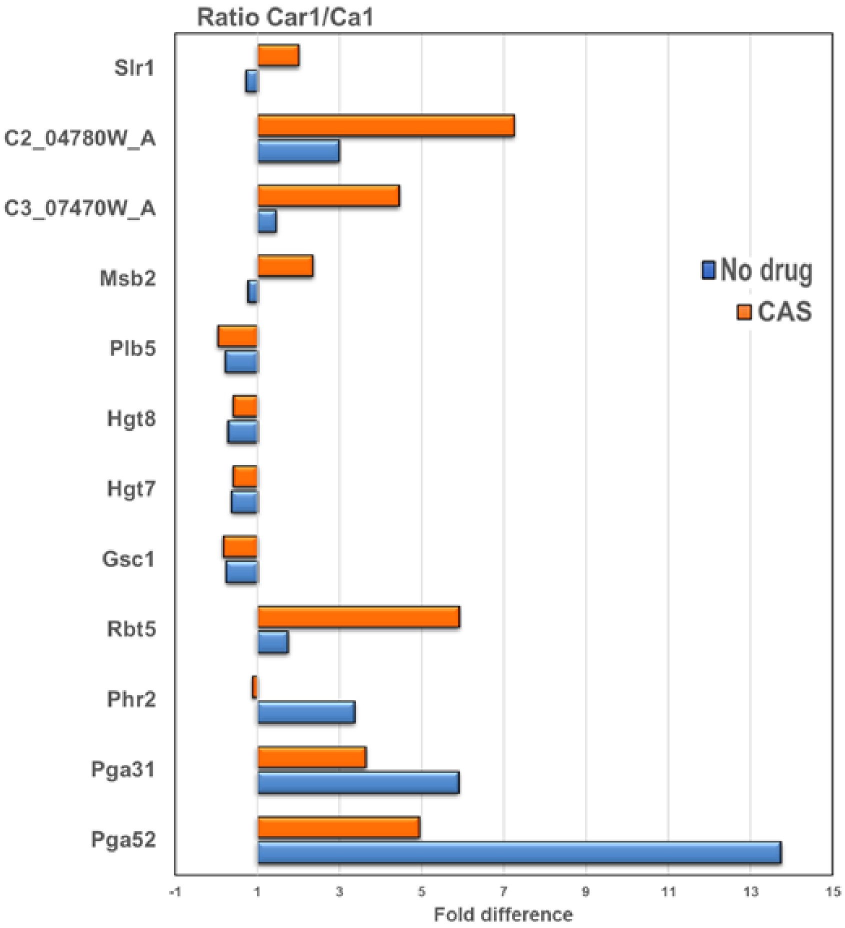


Figure 7.3 Proteomic analysis of cell wall fractions from *C. albicans* resistant and susceptible isolates exposed to caspofungin. Differential expression of relevant proteins identified by LC MS/MS in susceptible Ca1 and resistant Car1 isolates in absence (blue) and presence (orange) of caspofungin in RPMI 1640 medium. The values displayed are the ratios of the averages (n=3) of the peak areas from the LC MS/MS analysis of the two isolates (Car1 and Ca1). The full list is available in supplementary material.

7.3.3 Common response to caspofungin

Next, to ascertain whether the differences in the expression of cell surface proteins was strain specific or common among *C. albicans* isolates, the cell walls from seven additional isolates (listed in Table 7.1) were extracted and analysed by LC-MS/MS with the same strategy described in the previous section. A total of 842 proteins were detected and 566 of those were represented by at least 2 peptides. Thirty proteins were found exclusively in the resistant isolates in the absence of drug (Figure 7.4A), including cell wall proteins (Iff8, Eng1, Exg2, Pra1, Sep7, Plb4.5, Ihd1) but also cytoplasmic and plasma membrane proteins (Erg1, Hgt7, Hgt8, Mid1, Nce102), which are likely to be contaminants of the wall fraction. The 183 proteins detected only in the susceptible strains were mainly cytoplasmic contaminants. Amongst the 235 proteins in common, there were proteins involved in cell wall architecture, in particular with glucanosyl-transferase activity, such as Bgl2, Phr1 and Phr2, which are responsible for modifications of the glucan chain. The presence of caspofungin decreased, in both groups, the level of proteins involved in the host defence response and pathogenesis (*e.g.* Als1, Mp65, Als3) (Figure 7.4B). Moreover, differences in the abundances of several proteins responsible for cell wall organization and maintenance were measured in the two different conditions by performing a regression analysis and averaging the expression values between the resistant and the susceptible isolates (Figure 7.5C). In particular, important differences were observed for Sun41 (4.5-fold-difference), Mnt1 (0.27), Hyr1 (2.2), Als4 (0.34) in the group of resistant isolates in comparison with the susceptible isolates.

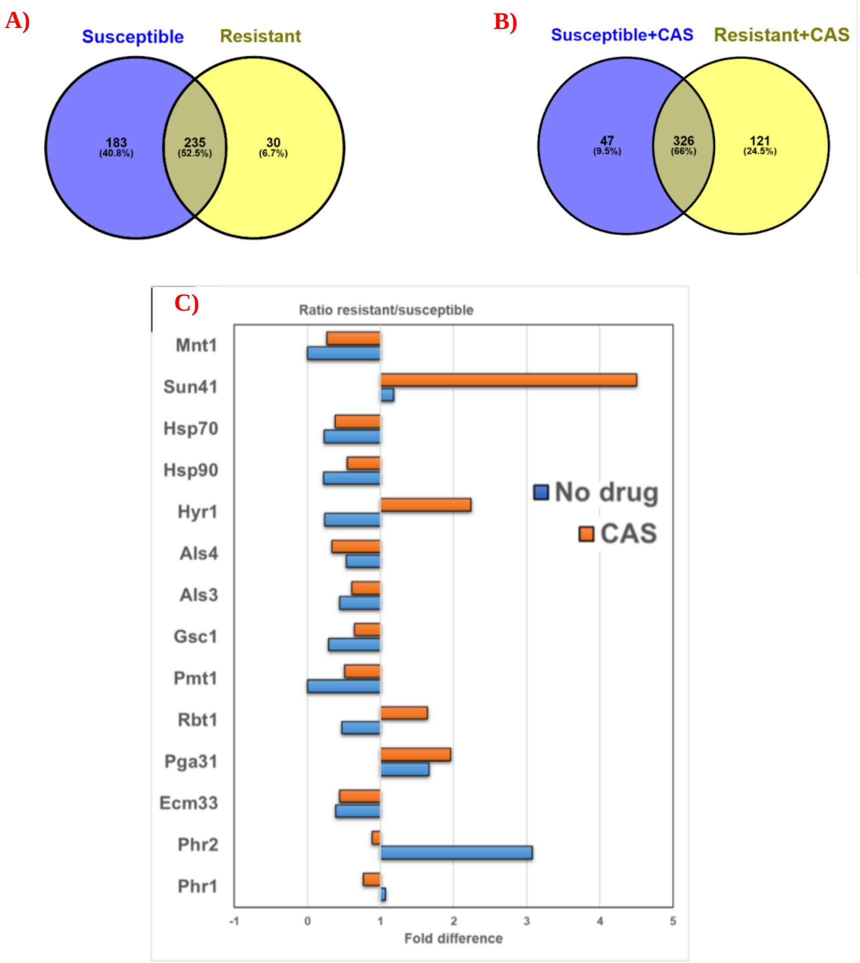


Figure 7.4 Proteomic analysis of cell wall fractions from *C. albicans* resistant and susceptible isolates exposed to caspofungin. (a, b) Total number of proteins identified by LC MS/MS in susceptible and resistant isolates in absence (a) and presence (b) of caspofungin in RPMI 1640 medium. (c) Differential protein expression for the resistant compared to the susceptible isolates with (orange) and without (blue) caspofungin. The values displayed are the ratios of the log 10 of the averages of the peak areas from the LC MS/MS analysis of the two groups (resistant and susceptible). Venn diagrams were created using Venny software. (n=1).

7.3.4 Interaction network of the cell wall proteome

In order to build an interaction network for the cell wall proteome of *C. albicans*, the data from LC-MS/MS were analysed with two different strategies: a statistical analysis and a network inferential analysis. Data were filtered, excluding from the analysis the proteins with less than two peptides and the missing values were randomly imputed using the *k*-NN algorithm. A differential analysis was performed in order to identify proteins differentially expressed between the two groups of isolates (caspofungin-resistant and caspofungin-susceptible). The results are shown in Figure 7.5. In absence of drug, the volcano plot shows a group of proteins differentially expressed (adjusted p-value<0.1; >1log$_2$ fold change difference in the ratios) with higher expression in the caspofungin-resistant compared to the caspofungin-susceptible isolates (Figure 7.5A). This group included: the cell wall 1,3-beta-glucanosyltransferase Phr2 (1.73 log$_2$ ratio) and glycosidase Sun41 (1.48 log$_2$ ratio); the GPI-anchored proteins Pga52 (1.64 log$_2$ ratio) and Rhd3/Pga29 (1.09 log$_2$ ratio); the proteins related to iron assimilation Rbt5 (1.54log$_2$ ratio) and Pga10 (1.49 log$_2$ ratio).

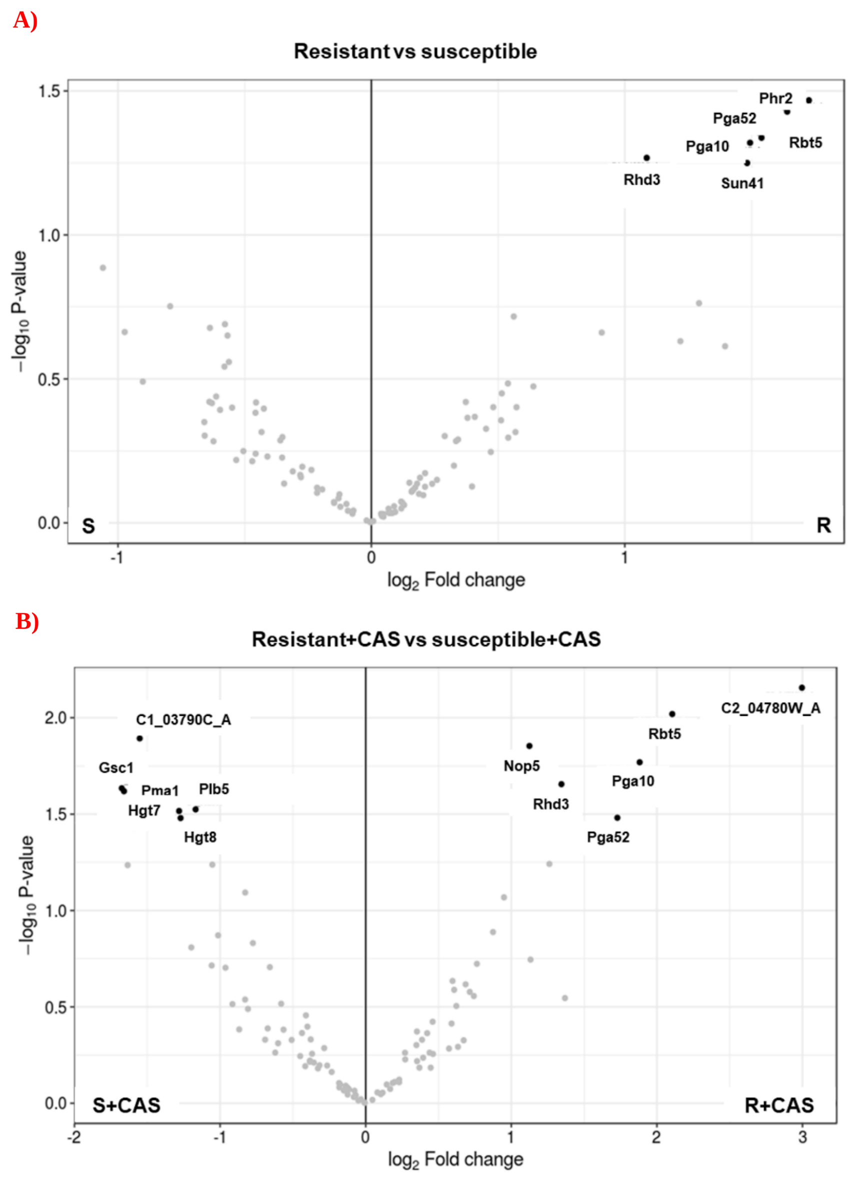


Figure 7.5 Volcano plots of cell wall proteome comparison of caspofungin-resistant and -susceptible isolate of *C. albicans* performed by differential expression (DE) analysis. The plots compare fold change and statistical significance of DE for caspofungin-resistant and -sensitive isolates of *C. albicans* (a) without drug and (b) with drug. The DE analysis was carried out using the limma package, part of Bioconductor software.

Incubation with caspofungin increased the number of proteins significantly and differentially expressed between drug-resistant and -susceptible isolates (Figure 7.5B). The group of proteins detected in reduced amounts in the drug-resistant isolates included: the cytoplasmic component C1_03790C_A (-1.55 $\log_2$ ratio); the plasma membrane proteins Pma1 (-1.66 $\log_2$ ratio), Hgt7 (- 1.28 $\log_2$ ratio) and Hgt8 (-1.27 $\log_2$ ratio); Gsc1 (-1.68 $\log_2$ ratio) and the wall enzyme Plb5 (-1.17 $\log_2$ ratio). The group of proteins detected in higher amount in the drug-resistant isolates included: the cytoplasmic component Nop5 (1.12 $\log_2$ ratio); the protein of unknown function C2_04780W_A, (3 $\log_2$ ratio); and the proteins already detected in no-drug condition (but with increased differences between the group of isolates) Rbt5 (2.11 $\log_2$ ratio), Rhd3 (1.34 $\log_2$ ratio), Pga10 (1.88 $\log_2$ ratio) and Pga52 (1.73 $\log_2$ ratio).

The network (Figure 7.6) was then created considering the interactions of the proteins as the level of the expression changes in two conditions (+/– caspofungin) between the two groups of isolates (drug-susceptible and drug-resistant). The green edges of the circles indicate the significant proteins for the DE analysis in all the conditions (adjusted p-value ≤ 0.1). The lines connecting the circles indicate the positive interaction (red, both increase or both decrease +/- caspofungin) or negative interaction (blue, one decreases and the other increases or vice versa). The thickness of the line indicates the strength of the interaction. The histograms inside the circles indicate the expression of each protein related to the normalised average between the two groups of isolates in the two conditions: from left to right, the columns represent susceptible, susceptible + caspofungin, resistant, resistant + caspofungin.

The proteins were clustered in seven groups and GO analysis performed. Each group was given a name based on the most abundant

GO terms. A general view of the network is presented in Figure 7.6a. A dense net of positive interactions was detected between the ribosomal proteins and the proteins involved in the modulation of the host response. In this group (Figure 7.6D) there were highly immunogenic enzymes, such as Pgk1 and Eno1, as well as cell wall-related proteins. Cell wall proteins Rbt5 and Pga10 (which also share a strong positive interaction) were found significantly overexpressed in the resistant isolates in the DE analysis, as well as Rhd3/Pga29, grouped into the ribosomal cluster (Figure 7.6B). The cell wall organisation cluster (Figure 7.6C) included proteins involved in modulation of cell wall (1,3)-glucan (such as Bgl2, Phr2 and Sun41) and Pga52, detected by the DE analysis to be highly expressed in caspofungin-resistant isolates in absence and presence of drug. A strong positive interaction between Pga52 and Phr2 was noticed. Another positive interaction between the cell wall damage sensor Msb2 and Rbt1, part of the group of proteins involved in adhesion to the host, was detected. Hyphal-associated proteins and adhesins, such as Als1, Hyr1 and Als3 (which shared a strong positive interaction), also belong to this group (Figure 7.6F) as well as the predicted GPI-anchored proteins Sap10 (which has a negative interaction with Hyr1) and Sod4. One of the most heterogeneous groups was the germ-tube formation cluster (Figure 7.6G), which included proteins with different functions, ranging from cell wall biosynthetic processes (*e.g.* Gsc1) to transmembrane transport (*e.g.* Pma1 and Cdr1), from cell wall adhesins (Als2 and Als4) to cytoplasmic proteins (*e.g.* Hmo1 and C1_03790C_A). A network of positive interactions of Cdr1 with other membrane proteins, such as Cdr2, Pma1, Ena21 and Sur7, as well as with Gsc1, was noticed. Gsc1 and Cdr1 were also involved in a negative interaction with Als4. Another negative interaction was detected between the adhesin Als2 and the unknown protein C2_04780W_A. Another group (Figure 7.6H) is made exclusively of plasma membrane glucose transporters (Hgt6, 7, 8), which did not have particular interactions with members of the other clusters. The last group, G0 (Figure 7.6e), consisted of proteins that were not found to have any interaction with any other group nor any protein within this group. Cell wall enzymes belonged to this group (*e.g*. Cht2, Crh11, Plb5 and Phr1), but also cytoplasmic proteins (*e.g*. Rpa34 and Rpl10).

In order to identify potential diagnostic markers for echinocandin resistance, the proteins from each group of the interactome, with a stable difference in the expression between drug-resistant and -susceptible isolates of *C. albicans,* are listed in Table 7.2. Moreover, to better understand the common cell wall response between susceptible and resistant isolates to caspofungin, the proteins differentially expressed between the two conditions are also listed in the table.

7.4 Discussion

The treatment options for invasive fungal infections are limited as there are relatively few classes of antifungal drugs available on the market. The main target for current antifungals is plasma membrane ergosterol and the effect of these drugs can be fungicidal but toxic to the host (polyenes) or fungistatic (azoles), hence the fungus is more prone to develop resistance (Berkow and Lockhart 2017). Echinocandins belong to a fungicidal class that targets the biosynthesis of cell wall β-(1,3)-glucan. Caspofungin was one of the first echinocandins discovered and patented (Rybowicz and Gurk-Turner 2002) and is recommended as first line treatment for *Candida* infections (Rybowicz and Gurk-Turner 2002, Hope *et al.* 2012, Cornely *et al.* 2012). Despite its fungicidal effect, the occurrence of drug resistance is a problem in the clinic due to acquisition of

GSC1/FKS1 mutations, which decrease the affinity of the drug towards the enzyme (Perlin 2007).

Table 7.2 Summary of the groups identified by the integration of a DE analysis and an interaction network built with the proteomics data.

GO Group	Marker	CAS changes
G1 (CW organisation)	Pga52	Msb2, C3_07470W_A
G2 (Germ-tube formation)	Gsc1, C2_04780W_A	Slr1
G3 (Modulation of host response)	Pga10, Rbt5	Pga31, Skn1
G4 (Adhesion to the host)	Sap10*, Als1*	Als3*
G5 (Ribosomal proteins)	Rhd3	Sim1
G6 (Glucose transporters)	Hgt7,8	
G0 (Ungrouped proteins)	Plb5	Cht2, Sik1, Phr1, Rpa34, Tos1

The marker column indicates the proteins for each group that were significant for the DE analysis and the CAS changes column the proteins changing in presence of caspofungin for both drug-resistant and –susceptible isolates. *not significant.

In this work, three caspofungin-resistant isolates of *C. albicans* were characterised and compared to six caspofungin-susceptible isolates, in order to better understand if the resistant isolates had modified call walls and if the 2 sets of isolates responded differently to the drug. The overarching aim was to identify possible diagnostic markers for drug resistance.

First, the caspofungin MIC of the isolates was assessed (Table 7.1). Car1 is a caspofungin-resistant strain recovered from the kidneys of mice treated with caspofungin, which were infected with the Ca1 strain (Perlin 2007, Lee *et al.* 2012). Car1 carries the most common host spot one (HS1) mutation in *FKS1* at codon *645* in which serine is replaced by tyrosine. This mutation has been shown to be responsible for a significant increase in the MIC, from 8- to more than 100-fold compared to the wild type allele, whereas other *FKS1* mutations accounted for smaller increases (4- to 30-fold) (Garcia-Effron, Park, *et al.* 2009). The Car3 isolate had a relatively high caspofungin MIC and carried a mutation in HS2 of *FKS1* (*R1361G*), with a non-polar residue (glycine) substituted for a positively-charged one (arginine); this is a mutation more common in *C. krusei* isolates (Park *et al.* 2005, Desnos-Ollivier *et al.* 2008).

An important feature of *Candida* infections *in vivo* is biofilm formation, which is associated with higher resistance to antifungal drugs, in some cases biofilms were 1000-fold more resistant to antifungal treatments than planktonic cells (Ramage *et al.* 2001, Rajendran *et al.* 2016). The capacity of the 9 isolates to form biofilms was assessed in the absence or presence of caspofungin (2 and 4 µg/ml) for 6, 24, 48 and 72 hours (Figure 7.1, Figure 7.2). In general, all the isolates seemed to reach the maximum level of biofilm biomass at the 24h time point (Figure 7.1B, Figure 7.2B). In the absence of drug, the drug-resistant isolates Car1, Car2 had biomass levels comparable to the susceptible isolates after the first time point, with no evident defects in the capacity to form biofilms. The one exception was the Car3 isolate, which did not increase biomass after 6h (Figure 7.2B-D). The presence of caspofungin completely inhibited the formation of biofilms in all the drug-susceptible isolates, probably due to the supra-MICs used, causing a decrease in cell viability. In the same conditions, the resistant isolates did not seem to be affected by the presence of the echinocandin. In this case the amount of drug used was above the IC_{50} of two out of three isolates, but their biomass levels were comparable to the no-drug condition. This study showed no effect of caspofungin on the capacity to form biofilms by caspofungin-resistant isolates in contrast biofilm formation of caspofungin -susceptible isolates was inhibited.

From the proteomic point of view, the changes in the cell wall following caspofungin treatment consisted of an increase in the expression of cell wall remodelling enzymes (such as the glucanosyl-transferases Phr1, Phr2) and the Crh family of chitin-glucanosyl-transferases (Fonzi 1999, Pardini *et al.* 2006). In this work, LC-MS/MS analysis detected altered amounts of several cell wall synthesis and remodelling enzymes in caspofungin-resistant isolates, such as Sun41, Phr1, Phr2, Gsc1, Pmt1 and Mnt1 (Figure 7.4C). Other proteins were also found differentially expressed in the walls of resistant isolates, such as Als3, Als4, Ecm33 and Pga31. Cytoplasmic proteins were detected in increased amounts in the susceptible isolates compared to resistant isolates in the absence of drugs (Figure 7.5A), which could indicate different permeability of the wall. Differences in the levels of Gsc1 protein, the echinocandin target, were also detected, with Gsc1 found at higher levels in drug-susceptible compared to drug-resistant isolates in the presence and absence of caspofungin (Figure 7.4C and Figure 7.4B). Previous work showed *FKS1/GSC1* mutations influenced the activity of the enzyme and the echinocandin IC_{50} for the mutated protein (Garcia-Effron, Park, *et al.* 2009), but there was no evidence of different expression patterns. The different Gsc1 levels detected by proteomics could be due to the different genetic backgrounds of the isolates or due to the point mutations affecting protein stability, this has still to be investigated.

These proteomic profiles suggest that the cell walls of resistant isolates are likely to be diverse, with altered remodelling and maintenance mechanisms not only upon incubation with drugs, but also in normal culture conditions.

The diagnostic “gold standard” for detecting *Candida* infections is still blood culture, with other additional tests based on the detection of circulating polysaccharides from the fungal cell wall or antigens in blood samples (Consortium Opathy and Gabaldón 2019). However, the diagnosis of antifungal resistance is an issue in the clinical settings, as it is not routinely performed and requires 48-72 hours, which is often too late to influence the treatment of the patient (Zhao and Perlin 2014). In this work, an interaction network was created from the mass spectrometry data. From this analysis several proteins were found differentially expressed between resistant and susceptible isolates, hence there is the possibility that they can be exploited as markers for antifungal resistance (Table 7.2).

The results presented here should be validated on a broader panel of clinical isolates to ensure the identified changes in the cell wall glycoproteome are consistent and common traits in drug resistant isolates. Moreover, a better characterisation of the fungal cell wall is required for therapeutic purposes, in order to investigate new targets for the future development of antifungal drugs.

7.5 Materials and methods

7.5.1 Strains and growth conditions

The isolates of *C. albicans* used in this study are listed in Table.1. Fungal cells were stored in 25% glycerol at -70ºC and re-cultured in YPD agar (1% [w/v] yeast extract [Oxoid], 2% [w/v] mycological peptone [Oxoid], 2% [w/v] glucose [Fisher Scientific], 2% [w/v] agar

[Oxoid]). For overnight cultures, unless indicated otherwise, a single colony of each strain was inoculated into YPD broth (1% [w/v] yeast extract, 2% [w/v] mycological peptone, 2% [w/v] glucose) and incubated overnight at 30 ºC with shaking at 200 rpm. For hyphal induction, cells were grown in RPMI-1640 modified medium (50% [w/v] RPMI-1640 [Sigma-Aldrich Co.], pH 0.8-1.5; 1.65M MOPS buffer [Melford], pH 7.2; 3.6% [w/v] glucose [Fisher Scientific]; 4.2mM L-glutamine [Sigma-Aldrich Co.]) with added 20% Fetal Calf Serum (FCS, Gibco) at 37ºC for 6h with 100 rpm shaking.

7.5.2 Antifungal susceptibility testing

The Minimum Inhibitory Concentration (MIC) of caspofungin (CAS) against *C. albicans* isolates was determined according to CLSI guidelines (Alexander 2017) and following the protocol described previously by (Walker *et al.* 2008). Cells were pre-grown overnight in YPD medium at 30 ºC and then diluted to $2*10^6$ cells/ml in 2X RPMI-1640 broth (Sigma Aldrich Co.) supplemented with 4.2 mM L-glutamine and grown at 37 ºC. Cells were incubated for 24h in flat bottomed 96 well plates (Nunc) containing serial dilutions of the drug in sterile water. Caspofungin (Cancidas, Merck and Co. Inc., USA) concentration ranged from 0.016 µg/ml to 16 µg/ml. After incubation optical densities were read in a VersaMax microplate reader (Molecular Devices, USA) at 405 nm.

7.5.3 Biofilm formation assay

The capacity of the strains to form biofilms was evaluated by a modified method from (Ramage *et al.* 2001). The cells were grown overnight in YPD medium at 30 ºC and transferred to a microtiter plate (Nunc) with RPMI-1640 plus 20 % Fetal Calf Serum (FCS, Gibco), incubated at 37 ºC for 6, 24, 48 h with 0, 2 or 4 µg/ml caspofungin. Planktonic cells were washed away with PBS

(Dulbecco’s Phosphate Buffered Saline [Sigma-Aldrich Co.]), and the remaining cells adhering to plastic surface (biofilm) were quantified by incubation with 0.05 % crystal violet for 20 min. They were then destained with 100% ethanol and the absorbance of the destain was measured, after transfer to a fresh microtiter plate, at 570nm in a VersaMax microplate reader (Molecular Devices, USA).

7.5.4 Cell wall proteomic analysis

The cell walls were isolated following a published protocol (Kapteyn *et al.* 2000) with some modifications. The cells were grown overnight in YPD medium at 30 ºC and transferred to RPMI-1640 at 37 ºC supplemented with 4.2 mM L-glutamine, until exponential phase was reached (OD_{600} = 0.4 ~ 0.6). Caspofungin was added to some cultures for 90 min. The caspofungin concentration used was based on the MIC tests, in order to achieve the same percentage of growth (~ 80 %) between the different isolates. Cells were then harvested by centrifugation at 3000 x*g* for 5 min and washed once in 10 mM Tris-HCl (pH 7.5). The mechanical breakage of the cells was accomplished using zirconia/silica 0.5 mm beads (Thistle Scientific) in a FastPrep machine (MP Biomedicals). The cell debris containing cell walls was washed 5 times in 1 M NaCl to remove cytoplasmic contamination, resuspended in buffer (500 mM Tris-HCl buffer [pH 7.5], 2 % [w/v] SDS, 0.3 M β-mercaptoethanol, and 1 mM EDTA), boiled 3 times at 100 ºC for 10 min and freeze-dried. The pellets were digested with trypsin according to the PRIME-XS protocol [43]. Mass spectrometry analysis was performed using a Q-Exactive Plus (Thermo Fisher Scientific) and tryptic peptides were identified using the MASCOT searching engine (Matrix Science). The analysis was carried out with Proteome Discoverer 2.2 software (Thermo Fisher Scientific), with the proteins matched from Candida Genome Database (Skrzypek *et al.* 2017) and a cut-off of at least 2 peptides detected per protein, with the area under the curve (AUC) used as semi-quantitative measure of abundance.

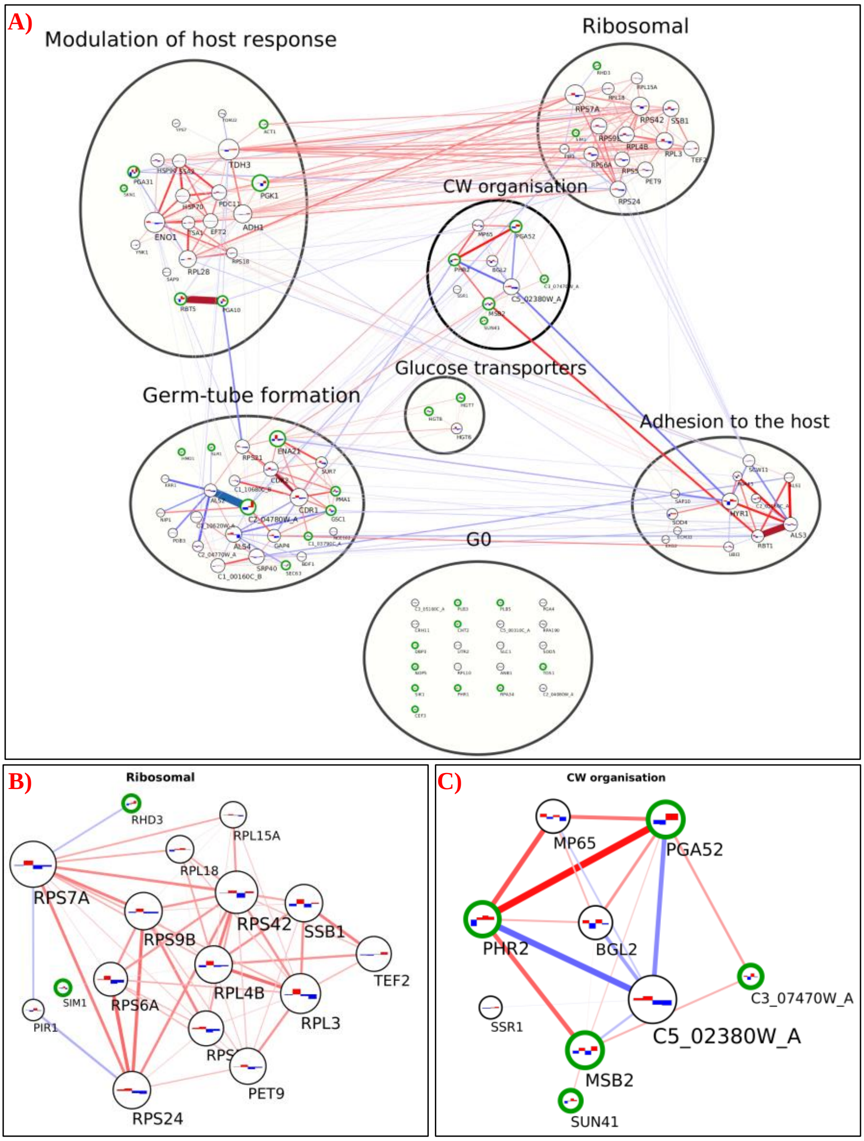
7.5.5 Statistical analysis

GraphPad Prism version 5.01 for Windows (GraphPad Software, USA) was used for all the statistical analyses in this work, unless specified. Differences in protein expression between different isolates were calculated using IBM SPSS Statistics (IBM, USA). A general linear model analysis was performed on the ranked values in order to avoid the missing hits in the data. Bonferroni was applied as a Post-Hoc test at the end of the analysis.

7.5.6 Cell wall network analysis

To build an interaction network for the cell wall proteome of *C. albicans*, the data from LC-MS/MS were analysed with two different strategies: a statistical analysis and a network inferential analysis. In order to overcome the problem of the missing values, the data were filtered and randomly imputed using the *k*-nearest neighbours algorithm. For statistics, a differential expression (DE) analysis was carried out using *limma* package, part of Bioconductor, an R-based software (Ritchie *et al.* 2015). For network inference, a multivariate Poisson log-normal (PLN) model, R-based package was used (Chiquet *et al.* 2019). Once the interaction network was built, cluster analysis was performed using fast greedy modularity optimization algorithm(Clauset *et al.* 2004) to identify major communities/groups in the cell wall proteom. Finally, Gene Ontology enrichment using CandidaMine enrichment tool (Chapter 8) was performed in order to assign a functional name to different groups, based on the most abundant terms, and the statistical analysis applied to the created network.

Data availability. All proteomic data is available at the PRIDE repository accession number: (Accession number will be available upon submission).



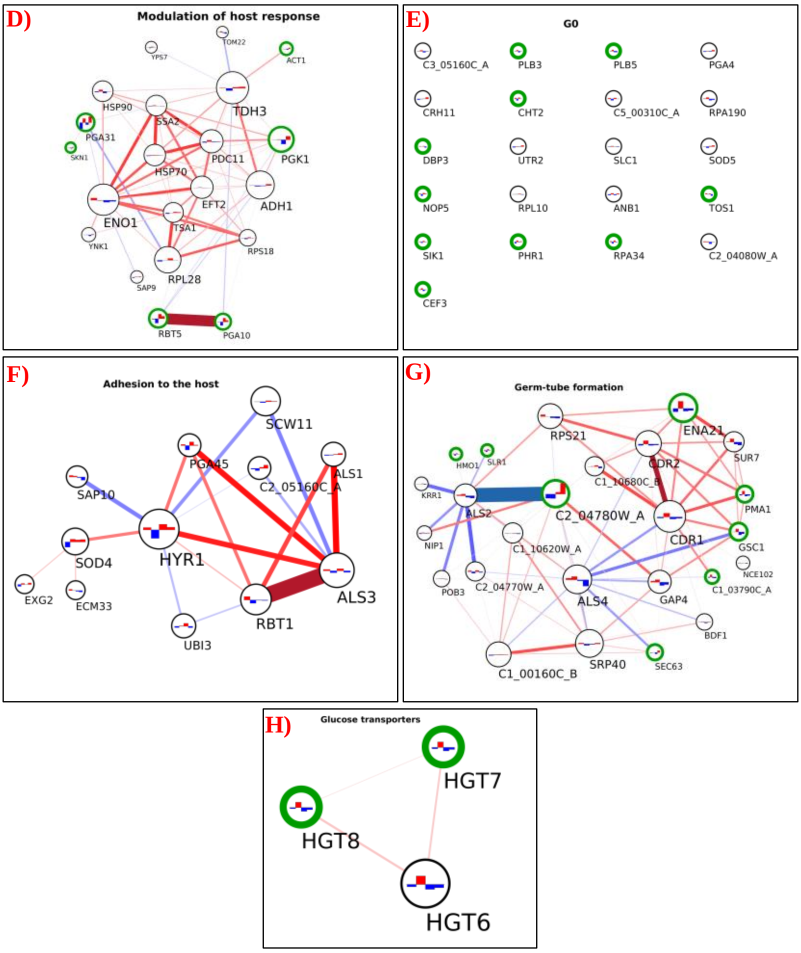


Figure 7.6 Interaction network of cell wall proteome from *C. albicans* isolates. The network was created using LC-MS/MS data from cell wall fractions of *C. albicans* isolates grown in absence and in presence of caspofungin. The green edges of the circles indicate the significant proteins for the DE analysis (adjusted p-value ≤ 0.1). The lines connecting the circles indicate the positive (red, both increase or decrease +/- caspofungin) or negative interaction (blue, one decreases and the other increases or vice versa). The thickness of the line indicates the strength of the interaction. The histograms inside the circles indicate the expression of the proteins related to the normalised average between the two groups of isolates in the two conditions: from left to right, the columns represent susceptible, susceptible+caspofungin, resistant, resistant+caspofungin. Proteins were clustered in seven groups and the GO analysis performed: (A) general view of the network; (B) ribosomal proteins; (C) cell wall organisation; (D) modulation of the host response; (E) G0; (F) adhesion to the host; (G) germ-tube formation; (H) glucose transporters.

8 CandidaMine, an integrative omics database for *Candida* yeast pathogens

Hafez, Ahmed, Hrant Hovhannisyan, Miquel Àngel Schikora-Tamarit, Manuel Molina, Carlos Llorens and Toni Gabaldón. "*CandidaMine, an integrative omics database for Candida yeast pathogens*." *(In preparation).*

8.1 Abstract

8.1.1 Background

High-throughput technologies such as next generation sequencing have become a widespread tool to study the physiology and evolution of organisms of interests, including human pathogens. The *Candida* clade comprises an evolutionary diverse group of pathogenic yeasts that cause a growing number of infections from superficial mucosal infections to life-threatening bloodstream infections in immunocompromised patients. High-throughput genome and transcriptome sequencing is increasingly being used to study the pathogenesis, epidemiology, and physiology of these yeasts, resulting in hundreds of diverse omics datasets. There is a need for centralized repositories that allow researchers to mine this wealth of information. CandidaMine is an integrative data warehouse for genomes, transcriptomes, and other data from yeast *Candida* pathogens. CandidaMine is powered by Intermine to enable advanced mining features for the development of the integrative analysis tools. InterMine is an open source data warehouse built specifically for the integration and analysis of complex biological data from a wide range of heterogeneous data sources.

8.1.2 Results

The main different data source for the current version is Candida Genome Database (CGD), Candida Gene Order Browser CGOB, Uniport, KEGG, Sequence Read Archive (SRA) and other published dataset studies related to *Candida* species. To build CandidaMine, *Candida* genomes and raw sequencing data are collected from CGD and SRA, Gene Homology information is collected from CGD and CGOB, proteins dataset extracted from Uniport and Interpro databases, Functional, Phenotypes and Pathways Annotations are collected from CGD and KEGG databases. Variant and gene expression dataset are computed from raw sequencing data available at SRA using common pipelines. Currently CandidaMine contains data for *Candida albicans*, *Candida glabrata*, *Candida parapsilosi*s, *Candida tropicalis* and *Candida auris*.

8.1.3 Conclusions

CandidaMine, comprehensively integrates information from various studies and resources to facilitate exploration, analysis, and interpretation of *Candida* related studies. It will be regularly updated and extended with new *Candida* species and datasets. It is aimed to serve as a dedicated resource for the broad research and clinical community. CandidaMine is freely available at http://candidamine.org

8.2 Introduction

High-throughput technologies such as next generation sequencing have become a widespread tool to study the physiology and evolution of organisms of interests. As mentioned in Chapter 0; the increasingly frequent recourse to omics and bioinformatics to generate valuable knowledge from high-throughput data has resulted in a wide repository of resources that have improved and updated previously existent reference databases. This includes a wide range of general databases such as UniProt Knowledgebase (Schneider *et al.* 2009, UniProt Consortium 2018), Ensembl (Kersey *et al.* 2016, Zerbino *et al.* 2018) Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa 2000) and others. In addition to the associated annotation in standard ontology vocabulary such as Sequence Ontology (SO), Gene Ontology (GO) (The Gene Ontology Consortium 2017) or the system of Enzyme Commission (EC) numbers (Bairoch 2000) (http://enzyme.expasy.org). Also raw sequencing reads are increasingly submitted to Sequence Read Archive (SRA) (Leinonen *et al.* 2011) by researchers worldwide. In total, the SRA database had 188545 submissions with 1871406 runs during 2019, which increased the number of bases stored in the SRA database by approximately 4566 tera-bases. In addition, the progression of fungal omics and bioinformatics has also promoted the emergence of a wide variety of databases and repositories specifically dedicated to fungi, including yeast pathogens as summarized in Chapter 0, these include the Candida Genome Database (CGD) (Skrzypek *et al.* 2017), the Candida Gene Order Browser (CGOB) (Maguire *et al.* 2013, Fitzpatrick *et al.* 2010) and the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe 2005).

The decentralized nature of the available database resources makes it a challenging task to mine and extract information especially if a complex analysis is involved. Each database has its own standard format to store data, different database schemes, or even some databases use unstructured, and use a text-based format. It is evident that centralizing such resources has many challenges; first, biological entities such as genes or proteins in different databases will have different key identifiers (Id), thus linking and cross-referencing such entities must be done carefully to avoid any linkage mistakes between different databases or even missing such cross-references. In addition, such centralized resources must provide an efficient and optimized storage mechanism, an optimized query engine and an automated updates mechanism to keep it updated with new releases of such different databases.

To provide researchers with centralized repositories for different *Candida* species and to enable them to mine this wealth of information, we developed CandidaMine which is an integrative data warehouse for *Candida* species genomes and transcriptomes. CandidaMine is powered by Intermine (Smith *et al.* 2012, Kalderimis, Lyne, *et al.* 2014) to host different *Candida* related omics' dataset to enable advanced mining features for the development of the integrative analysis tools. InterMine is an open source data warehouse built specifically for the integration and analysis of complex biological data. Using InterMine, large biological databases can be created from a wide range of heterogeneous data sources, and the extensible data model allows for easy integration of new data types. As part of the InterMOD (Sullivan *et al.* 2013) project, a number of InterMine data-warehouses have been developed and released to the public containing high-quality integrated data curated by the major model organism database (MOD) organisations (Kalderimis, Stepan, *et al.* 2014). In addition, the InterMine platform is widely used by other projects, such as the modENCODE (Contrino *et al.* 2012) project and metabolicMine (Lyne *et al.* 2013). Thus, providing reliable integrated data sets for researchers working in a wide range of fields in the life-sciences, which can be accessed by a common interface.

8.3 Materials and Methods

CandidaMine contains data from the four most common *Candida* species causing candidiasis as shown in Table 8.1 which account for more than 90% of invasive candidiasis (Fuchs *et al.* 2019), in addition to the multi-drug resistant *Candida auris* (Aznar-Marin *et al*. 2016, Rhodes and Fisher 2019). The Main different data sources for the current version is CGD, CGOB, Uniport, KEGG, and SRA as listed in Table 8.2.

Table 8.1 Candida Species stored in CandidaMine.

Name	Taxon Id
Candida albicans SC5314	237561
Candida glabrata CBS 138	284593
Candida parapsilosis CDC317	578454
Candida tropicalis MYA-3404	294747
Candida auris	498019

8.3.1 Building CandidaMine

To build CandidaMine, *Candida* reference genome assemblies and annotations were collected from CGD, Gene Homology information is collected for CGD and CGOB, proteins dataset extracted from Uniport and Interpro databases. Functional, Phenotypes and Pathways annotations are collected from CGD and KEGG databases. Gene Expression dataset is measured on various samples obtained from SRA under different experiment conditions.

In addition, CandidaMine loads Sequence Ontology (SO) (Eilbeck *et al.* 2005), Gene Ontology (GO) (Ashburner *et al.* 2000, The Gene Ontology Consortium 2017) and Ascomycete Phenotype

Ontology (APO) (http://www.yeastgenome.org/) as supporting ontology annotations for other datasets. This includes ontology vocabulary terms, corresponding descriptions, and relationships between ontology terms. Current version of CandidaMine loads Table 8.3 summarizes the list of dataset types and source of those datasets.

Table 8.2 Main database sources used to build CandidaMine.

Candida Genome Database **(CGD)** (Skrzypek *et al.* 2017)	http://www.candidagenome.org/
Candida Gene Order Browser **(CGOB)** (Maguire *et al.* 2013)	http://cgob3.ucd.ie/
Uniprot (UniProt Consortium 2018)	http://www.uniprot.org/
Interpro (Finn *et al.* 2017)	https://www.ebi.ac.uk/interpro/
KEGG (Kanehisa 2000)	http://www.genome.jp/
STRING (Szklarczyk *et al.* 2019)	https://string-db.org/
Sequence Read Archive **(SRA)** (Leinonen *et al.* 2011)	https://www.ncbi.nlm.nih.gov/sra

InterMine comes with ready to use data source loaders, such as genomic sequences loader from fasta file, genomic features from Generic Feature Format (GFF) files. The framework has a core genomic model where each entity is represented as an object (that will be translated into a table in the database) each object has defined attributes (translate to columns in the database). Attributes could be as simple as text or numeric values or complex entities such as other objects in the model (translate to relationship between tables in the database model). Custom data loaders can be implemented to extend the core model. CandidaMine has several custom implementations of data loaders to load expressions profiles, variants data, SRA metadata, STRING interactions dataset, phenotype and pathways annotations that are customized to the compiled data from different database sources. All custom implementations and configurations for CandidaMine are available at the public GitHub repository (https://github.com/Gabaldonlab/candidamine).

To resolve cross references and different ids in multiple data sources, CGD main Ids are used as the primary identifiers in CandidMine, Ids and gene names synonyms are collected across all used data sources and Ids map were compiled to resolve Ids cross-referenced.

Table 8.3 Dataset type and corresponding description stored in CandidaMine.

Data Set		**Source**	**Species**
Genomics	Fasta Sequences, Sequence Feature (GTF)	CGD	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis* *C. auris*
Homology	Gene Orthologues	CGD and CGOB	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis* *[C. dubliniensis,* *S. cerevisiae, A. nidulans]*
Proteins	Proteins	Uniprot	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis*
	Protein domains	Interpro	
	Uniprot Proteins to Proteins Domains	Interpro	
	IprScan *(Jones et al. 2014)*	CGD	

Data Set		Source	Species
Annotation	*Go Annotations*	CGD	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis*
	Phenotype annotation	CGD	*C. albicans,* *C. glabrata,* *C. parapsilosis*
	Pathways	KEGG	*C. albicans,* *C. glabrata,* *C. tropicalis*
Variants	*SNPs and INDELs (VCF)*	SRA*	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis* *C. auris*
Expression Profiles		SRA*	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis*
Interactions	*Protein to protein interactions*	String	*C. albicans,* *C. glabrata,* *C. parapsilosis,* *C. tropicalis*

*sequencing reads retrieved from SRA then processed as described in the following section.

8.3.2 Expression Profiles

The gene expression levels of the four main *Candida* species were estimated using all publicly available RNA-Seq datasets accessible in the SRA database as of July 2019. First, we obtained the RNA-Seq data from SRA using sratoolkit v. 2.9.6-1 with prefetch and fastq-dump functions. For *C. albicans*, we first have discarded the samples (n=64) with read length less than 49bp. Then we used FastQC v0.11.6 (Andrews and Others 2010) and Multiqc v. 1.0 (Ewels *et al.* 2016) software to perform quality control of the remaining data. Subsequently, we used trimmomatic v. 0.36 (Bolger *et al.* 2014) to process all the samples in a uniform manner with the following parameters: <ADAPTERS.fa>:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:49. Finally, we have calculated the gene expression levels, *i.e.* raw read counts and Transcripts Per Million (TPM) values, using salmon v. 0.12 (Patro *et al.* 2017) with --*gcBias* and -*g* parameters. Salmon was run using the corresponding transcriptomes of each species generated by RSEM prepare-reference v1.3 (Li and Dewey 2011). Reference genomes and annotations for transcriptome generation of *C. albicans* SC5314 (assembly 22, haplotype A), *C. glabrata* CBS138 and *C. parapsilosis* CDC317 and *C. tropicalis* MYA-3404 were obtained from *Candida* Genome Database (CGD, last accessed on 17 of August 2017. In total, from publicly available data we have estimated gene expression levels of 1966 *C. albicans* samples, 123 *C. glabrata* samples, 129 *C. parapsilosis* samples and 46 *C. tropicalis* samples.

8.3.3 lncRNA data

Same data used to calculate expression profiles were used to predict long non-coding RNAs (lncRNA). After processing all samples with trimmomatic as described in the previous section. For more details about the process and the pipeline used please refer to (Hovhannisyan and Gabaldón 2020). This dataset includes lncRNA catalogues for *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis.*

8.3.4 Variant Data

Currently CandidaMine provides variant-calling information for all the paired-end whole-genome re-sequencing (WGS) datasets found in SRA (at 09/06/2020) related to major *Candida* pathogens. We keep only those with >40x coverage and >90% of the reference genome covered, after read alignment with bwa mem (v0.7.17,http://bio-bwa.sourceforge.net/bwa.shtml) and coverage calculation with mosdepth v0.2.6 (Pedersen and Quinlan 2018). This yielded 764, 652, 420, 26, 33, 51 and 4 datasets for *C. auris*, *C. albicans*, *C. glabrata*, *C. metapsilosis*, *C. orthopsilosis*, *C. parapsilosis* and *C. tropicalis*, respectively. Each sequencing dataset was processed with a custom pipeline (perSVade v0.4, available at https://github.com/Gabaldonlab/perSVade.). In brief, we use three different tools which are GATK Haplotype Caller (HC v4.1.2 (Poplin *et al.* 2018), freebayes (FB v1.3.1) (Garrison and Marth 2012) and bcftools (BT v1.9) (Danecek and McCarthy 2017) to call and filter Single Nucleotide Polymorphisms (SNP) and small insertions/deletions (INDELs) in haploid or diploid configuration, matching the ploidy of each organism. We define as high-confidence (PASS) variants those with read depth above 12, with extra filters for HC and FB. For HC, we keep as PASS variants those where 1) there are <4 additional variants within 20 bases; 2) the mapping quality is >40; 3) the confidence based on depth is >2; 4) the phred-scaled p-value is <60; 5) the MQRankSum is >-12.5 and 6) the ReadPosRankSum is > -8. For FB, we keep as PASS variants those where 1) quality is > 1 or alternate allele observation count is > 10; 2) strand balance probability of the alternate is > 0; 3) number of observations in the reverse is > 0; and 4) number of reads placed to the right/left of the allele are > 1. In addition, we discard variants where the fraction of reads covering the alternative allele is <90% (haploid) or <25% (diploid). We use a combination of bcftools and custom python code to normalise and merge the variants called by each software.

We keep only those variants called with high confidence by two or more tools. For diploid calls, we define the genotype with the strongest support (the one called by most programs). In addition, the quality of each variant is calculated from the mean of the three algorithms. VCF files are then processed with the SnpEff v5.0 (Cingolani *et al.* 2012) to predict and annotate the effects of all variants in the genome. Sequence variations and consequences are stored using Sequence ontology terms and Human Genome Variation Society (HGVS) recommended Nomenclature (den Dunnen *et al.* 2016).

8.4 CandidaMine Overview

The InterMine framework provides multiple ways to query or analyse the data stored in CandidaMine, as well as straightforward export options for the retrieval of results and sequence information. The framework stores every data item in the form of an object or entity, *e.g.* gene or protein, each object has specific type, *e.g.* gene, transcript , some attributes, a key identifier or more, and relationships to other objects in the mine such relationships could be cascaded or hierarchical; more details will be described in the query builder section. Additionally, the framework provides application program interface (API) and client libraries to mine data programmatically using HTTP API or client library in: Java, Perl, Ruby, Python, R and JavaScript.

8.4.1 Search

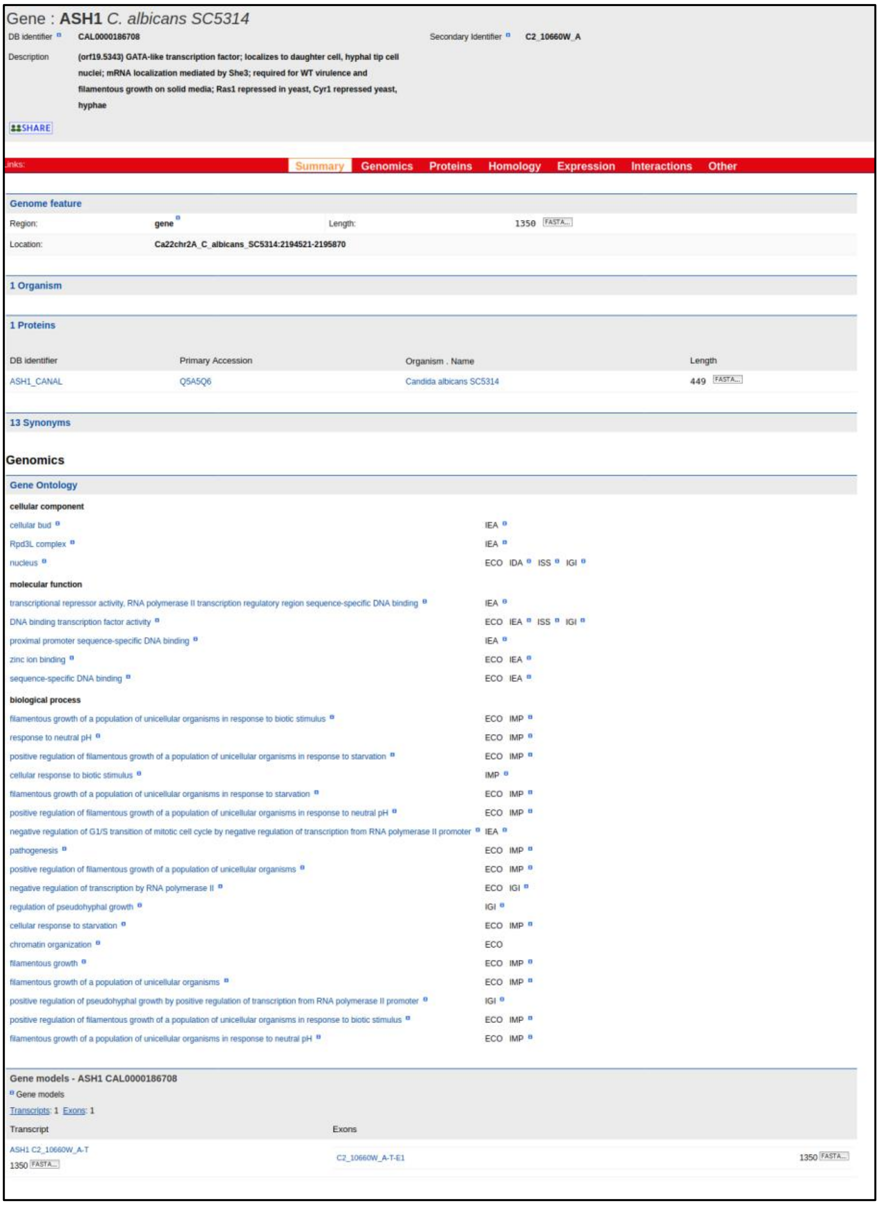
CanidaMine provides multiple ways to search or query the database including search by keyword, by genomic region, or by running custom or template queries. Keyword search allows searches for specific items by Ids or specific key annotations (*e.g.* domain names, gene ontology, etc). Genomic region search allows the user to fetch genomics features such as transcripts, exons, etc that are within a given set of genomic coordinates or are within a given number of bases flanking the coordinates. Custom and template queries allow the user run complex and constrained queries against CandidaMine to retrieve all matching items. Searches generally result in lists of objects associated with the given search Clicking an item brings up the respective detailed report page showing all available annotation information for this specific object.

8.4.2 Report page description

Every object (*e.g.*, Gene, Protein, Exon) in CandidaMine has a detailed report page to view and explore all associated information. The layout of the report page depends on the data available for that object. Genes as the main biological entity in the mine, have the most comprehensive report page to display genomic location, protein product, functional annotations, homologous genes, expression profiles, interactions, genomic variations, and other information as shown in Figure 8.1. The contents of the report page are divided into categories based on the type of information provided. A Genome Browser (jBrowse) (Buels *et al.* 2016) is integrated in the report page of all genomic features.

8.4.3 List Analysis and Enrichment analysis

A powerful feature of the InterMine framework is the analysis of features lists *e.g.* genes or proteins. Users can store gene lists for example and list of differentially expressed genes from a specific RNA-seq experiment then performing GO-term enrichment analysis on such lists. In addition, in such cases where the available annotations are scarce for the species under study, for example *C. tropicalis*, users can switch to an equivalent orthologues list in another species such as *C. albicans* and perform the enrichment analysis.



Gene : **ASH1** *C. albicans SC5314*

DB identifier	CAL0000186708	Secondary Identifier	C2_10660W_A
Description	(orf19.5343) GATA-like transcription factor; localizes to daughter cell, hyphal tip cell nuclei; mRNA localization mediated by She3; required for WT virulence and filamentous growth on solid media; Ras1 repressed in yeast, Cyr1 repressed yeast, hyphae		

Genome feature

Region:	gene	Length:	1350 FASTA...
Location:	Ca22chr2A_C_albicans_SC5314:2194521-2195870		

1 Organism

1 Proteins

DB identifier	Primary Accession	Organism . Name	Length
ASH1_CANAL	Q5A5Q6	Candida albicans SC5314	449 FASTA...

13 Synonyms

Genomics

Gene Ontology

cellular component	
cellular bud	IEA
Rpd3L complex	IEA
nucleus	ECO IDA ISS IGI
molecular function	
transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	IEA
DNA binding transcription factor activity	ECO IEA ISS IGI
proximal promoter sequence-specific DNA binding	IEA
zinc ion binding	ECO IEA
sequence-specific DNA binding	ECO IEA
biological process	
filamentous growth of a population of unicellular organisms in response to biotic stimulus	ECO IMP
response to neutral pH	ECO IMP
positive regulation of filamentous growth of a population of unicellular organisms in response to starvation	ECO IMP
cellular response to biotic stimulus	IMP
filamentous growth of a population of unicellular organisms in response to starvation	ECO IMP
positive regulation of filamentous growth of a population of unicellular organisms in response to neutral pH	ECO IMP
negative regulation of G1/S transition of mitotic cell cycle by negative regulation of transcription from RNA polymerase II promoter	IEA
pathogenesis	ECO IMP
positive regulation of filamentous growth of a population of unicellular organisms	ECO IMP
negative regulation of transcription by RNA polymerase II	ECO IGI
regulation of pseudohyphal growth	IGI
cellular response to starvation	ECO IMP
chromatin organization	ECO
filamentous growth	ECO IMP
filamentous growth of a population of unicellular organisms	ECO IMP
positive regulation of pseudohyphal growth by positive regulation of transcription from RNA polymerase II promoter	IGI
positive regulation of filamentous growth of a population of unicellular organisms in response to biotic stimulus	ECO IMP
filamentous growth of a population of unicellular organisms in response to neutral pH	ECO IMP

Gene models - ASH1 CAL0000186708

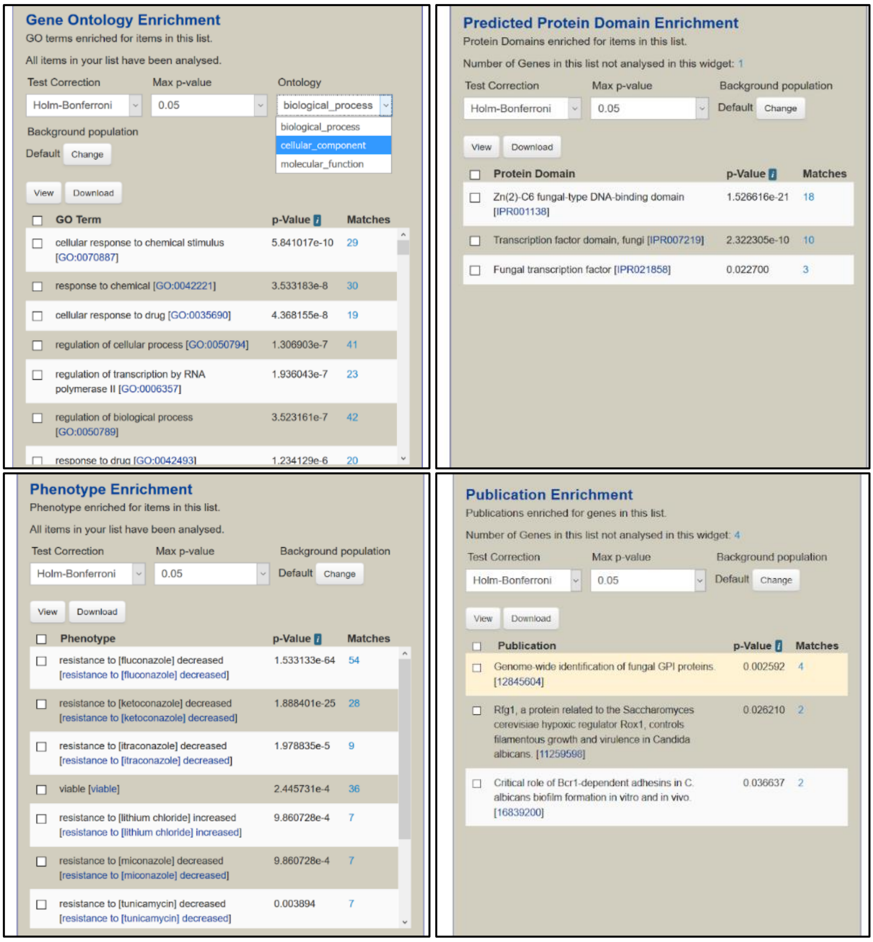
Gene models

Transcripts: 1 Exons: 1

Transcript	Exons	
ASH1 C2_10660W_A-T 1350 FASTA...	C2_10660W_A-T-E1	1350 FASTA...

Figure 8.1 A partial view of the report page for ASH1 gene in *C. albicans*. Full page view is available at http://candidamine.org/candidamine/gene:CAL0000186708.

Enrichment Analysis in CandidaMine include Gene Ontology (Go) enrichment, proteins domain Enrichment including proteins domain from Interpro database and predicted domain from CGD, functional pathway enrichment, Ascomycete Phenotype Ontology enrichments and publication enrichment. Figure 8.2 shows some examples of such enrichment tools.



Gene Ontology Enrichment

GO terms enriched for items in this list.

All items in your list have been analysed.

Test Correction: Holm-Bonferroni

Max p-value: 0.05

Ontology: biological_process

Background population: Default

□	GO Term	p-Value	Matches
□	cellular response to chemical stimulus [GO:0070887]	5.841017e-10	29
□	response to chemical [GO:0042221]	3.533183e-8	30
□	cellular response to drug [GO:0035690]	4.368155e-8	19
□	regulation of cellular process [GO:0050794]	1.306903e-7	41
□	regulation of transcription by RNA polymerase II [GO:0006357]	1.936043e-7	23
□	regulation of biological process [GO:0050789]	3.523161e-7	42
□	response to drug [GO:0042493]	1.234129e-6	20

Predicted Protein Domain Enrichment

Protein Domains enriched for items in this list.

Number of Genes in this list not analysed in this widget: 1

Test Correction: Holm-Bonferroni

Max p-value: 0.05

Background population: Default

□	Protein Domain	p-Value	Matches
□	Zn(2)-C6 fungal-type DNA-binding domain [IPR001138]	1.526616e-21	18
□	Transcription factor domain, fungi [IPR007219]	2.322305e-10	10
□	Fungal transcription factor [IPR021858]	0.022700	3

Phenotype Enrichment

Phenotype enriched for items in this list.

All items in your list have been analysed.

Test Correction: Holm-Bonferroni

Max p-value: 0.05

Background population: Default

□	Phenotype	p-Value	Matches
□	resistance to [fluconazole] decreased [resistance to [fluconazole] decreased]	1.533133e-64	54
□	resistance to [ketoconazole] decreased [resistance to [ketoconazole] decreased]	1.888401e-25	28
□	resistance to [itraconazole] decreased [resistance to [itraconazole] decreased]	1.978835e-5	9
□	viable [viable]	2.445731e-4	36
□	resistance to [lithium chloride] increased [resistance to [lithium chloride] increased]	9.860728e-4	7
□	resistance to [miconazole] decreased [resistance to [miconazole] decreased]	9.860728e-4	7
□	resistance to [tunicamycin] decreased [resistance to [tunicamycin] decreased]	0.003894	7

Publication Enrichment

Publications enriched for genes in this list.

Number of Genes in this list not analysed in this widget: 4

Test Correction: Holm-Bonferroni

Max p-value: 0.05

Background population: Default

□	Publication	p-Value	Matches
□	Genome-wide identification of fungal GPI proteins. [12845604]	0.002592	4
□	Rfg1, a protein related to the Saccharomyces cerevisiae hypoxic regulator Rox1, controls filamentous growth and virulence in Candida albicans. [11259598]	0.026210	2
□	Critical role of Bcr1-dependent adhesins in C. albicans biofilm formation in vitro and in vivo. [16839200]	0.036637	2

Figure 8.2 Examples of Gene List Enrichment & Analysis show some basic information about the genes and different Enrichment Analysis on the list with the ability to change background population, test correction and the adjusted p-value for selecting enrichment terms. A) Gene Ontology Enrichment. B) Proteins domain Enrichment (predicted domain from CGD). C) Phenotype (APO) Enrichment. D) Publication Enrichment.

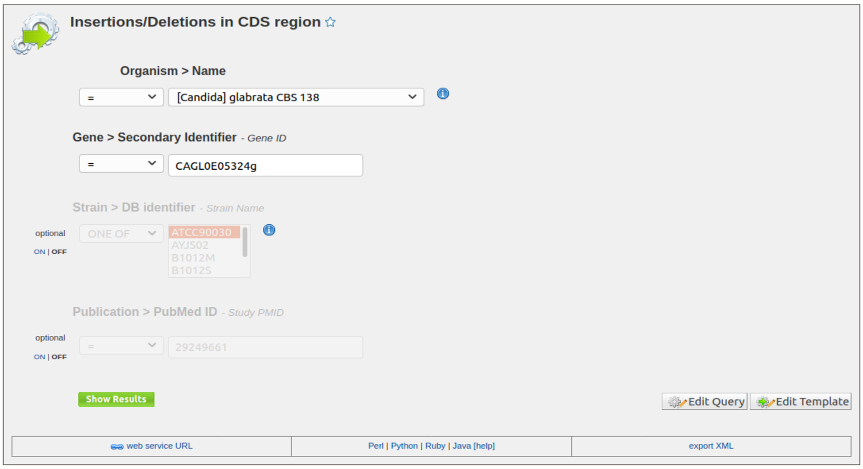
8.4.4 Template Query

Template Queries allow to mine the database without using Query builder with predefined search query which cover a full range of the stored data. Available Templates queries can be accessed from templates menu option or by popular templates sections which are displayed on the home page and grouped by category *e.g*. Genes, Protein, Homology, etc.

The following list mentions some examples of the available template in CandidaMine:

- **Genes to Proteins**: retrieve all genes and the corresponding protein product.
- **Genes to Go**: retrieve all genes with all associated gene ontology annotations.
- **Genes to Pathways**: retrieve all genes with all associated pathways annotations.
- **Genes to Domains**: retrieve all genes with all associated proteins domains.
- **Proteins -> Gene - Homologous <- Proteins**: retrieve proteins in one organism and the corresponding proteins in another organism by cross mapping homologous genes.
- **Gene to Phenotypes**: retrieve all genes with all phenotype annotations.
- **Gene to Sequence Alteration:** retrieve all genes with the corresponding sequence variations *e.g*. SNPs, insertions, and deletions in different strains.
- **INDELs in coding regions:** retrieve all insertions and deletions in coding regions.

Figure 8.3 shows an example of a template query to retrieve all INDELS located in coding regions. As shown in figure the template has required parameters to filters by organism and/or selected target gene. In addition, there are some optional filters that can be activated for a more constrained query.



Insertions/Deletions in CDS region

Organism > Name
= [Candida] glabrata CBS 138

Gene > Secondary Identifier - Gene ID
= CAGL0E05324g

Strain > DB identifier - Strain Name
optional ON | OFF
ONE OF ATCC90030 AYJS02 B1012M B1012S

Publication > PubMed ID - Study PMID
optional ON | OFF
= 29249661

Figure 8.3 Template Query Example

8.4.5 Custom Query

Query builder provides an easy way to create new search queries. Query builder has a fast learning curve and provides flexible tools to design complex queries that could target all stored information in CandidaMine. Building a new query starts by choosing a data type of interest *e.g.* gene or transcript based on the required result. After choosing a data type, the Model browser appears displaying the attributes for the selected feature class.

Figure 8.4 shows an example of building a new query to select all insertions and deletions with coding regions of a specific gene of interest filtered by some strains similar to template query shown in Figure 8.3. In this case Sequence Alteration data type (based on SO terms) was selected Figure 8.4A. Then desired attributes that would be retrieved in the result table are selected. To restrict the retrieved sequence alterations to be of Insertion or Deletion, a constraint is added to the query by selecting constrain button then configure the filter as shown in Figure 8.4B. Sequence Alteration data type is a sequences feature that overlap with other genomic sequence features, we can selected to retrieve all overlapping feature with the result Sequences alteration, however to select only those within coding region we constrain overlapping feature to be of only Exon data type as shown in Figure 8.4C. Once Overlapping features are constrained as Exons, more attributes are shown in the model browser under its node *e.g*. parent Gene. Accordingly, we can constrain the parent gene of the exons as shown in Figure 8.4D and constrain the strains as shown in Figure 8.4E.

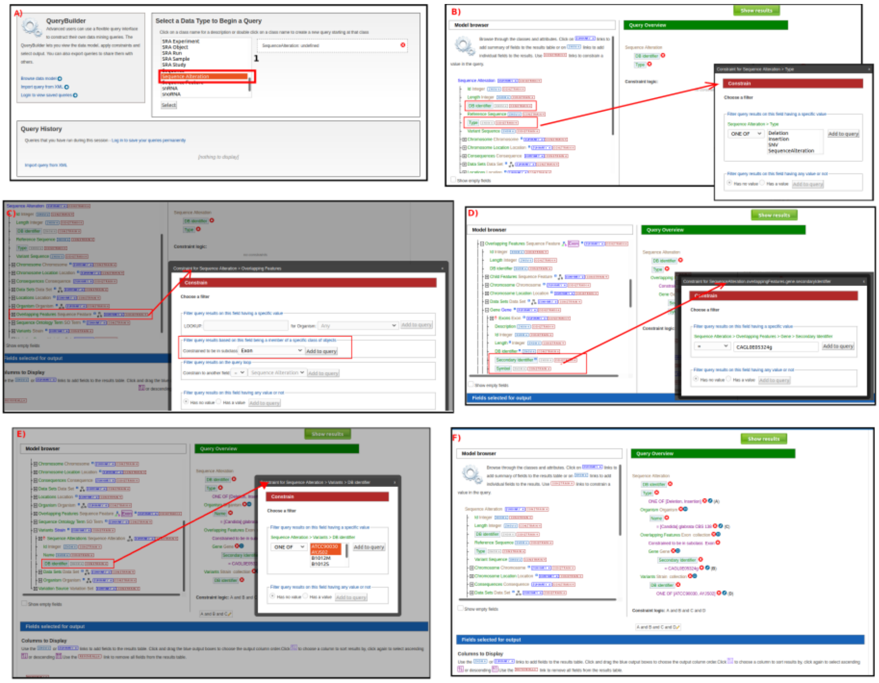


Figure 8.4 A step by step example on how to build a custom query to retrieve all insertion and deletions within the coding region of a target gene filter by some strains. A) Select Object of interest in this case is Sequences alteration to begin designing the query. B) add basic attribute to the query result and constraint type attribute to be Deletion and Insertion. C) Constrain overlapping features to be only of type Exons. D) Add basic attribute of the gene from the Exon object and constrain Secondary Identifier to specific gene of interest. E) Constrain Variant strain identifier. F) Final layout of the template after specifying all attributes to show in the result and the contains to control the final output.

9 Summarizing Discussion

Most of the described fungal infections in humans are caused by dermatophytes and *Candida* species (Richardson and Warnock 2012). Candidiasis is the general term used to designate fungal infections caused by *Candida* yeasts. Invasive candidiasis is a serious, progressive, and potentially fatal infection that can affect the blood and other tissues that are otherwise considered sterile. Invasive candidiasis accounts for up to 75% of all systemic fungal infections, and poses a serious threat to life, particularly in immunocompromised patients, for whom mortality rates that can exceed 50% (Brown *et al.* 2012) despite the use of currently available antifungal drugs. In particular, *Candida*-related bloodstream infections have a mortality rate between 30%-60% (Hirano *et al.* 2015). *Candida* infections are associated with a high economic burden - derived from longer hospital stays and the need for multiple analyses - with related costs accounting for over $7.2 billion in 2017 just in the USA (Benedict *et al.* 2019). For all these reasons there is a growing need for fast and effective diagnostics tools and/or new tools for finding reliable biomarkers that can be used for diagnostic purposes (Consortium Opathy and Gabaldón 2019).

Although still lagging behind the diagnosis of cancer or viral and bacterial infections, the field of diagnosis of fungal infections is starting to harness recent developments in areas such as proteomics and NGS. These developments are usually based on three major pillars; i) Analytical instrumentation: NGS, proteomic or other technology devices that process bio-samples and generate various data in a high throughput and resolution manner. ii) Computational tools: the software solutions that are responsible for processing the raw data generated by such devices and produces useful biological and clinical insights ready to be interpreted by researchers or clinicians; and, iii) Knowledge base. Existing knowledge or data generated from previous projects can be stored in a knowledge database that is consulted and utilized by the expert directly or via the software in order to retrieve and extract annotations, curated information or other relevant data required by the analysis. In the clinical context, fungal diagnosis is still in its infancy and part of the bottleneck is on the bioinformatics side because of the need of developing easy-to-use tools based on friendly to use solutions to translate high-throughput developments into ‘ready-to-use’ systems (also known as point-of-care testing) that could be routinely used by researchers and clinicians in their day-to-day practice.

This thesis has been developed in the framework of the EU-funded international training network OPATHY (www.opathy.org), an inter-disciplinary research network aiming to develop novel solutions to study, treat, and diagnose yeast pathogens. The thesis itself has been developed as a collaborative effort between an academic institution (CRG) and a bioinformatics company (Biotechvana) and has therefore a strong translational component. Due to the engagement in the form of a Marie Curie International Training Network, many collaborations with other partners in developing automated pipelines software tools and integrative database and analysis solutions have been established in this work. As a result, the developed tools have been used in different projects of the consortium, highlighting their practical utility. The overall objective of the presented thesis has been to develop integrative bioinformatics software tools and integrative databases for knowledge discovery which can be applied to fungal diagnostic applications. Aiming to provide better User Experience (UX), interactive GUI based software tools, pipeline automation and customization for maximizing productivity and reproducibility. This includes two major projects to develop and contribute to the software and database components of diagnostic application developments that can be used in research settings and hopefully will be the basis for further diagnostic applications used in clinical settings. Integrative software solutions were presented in Chapter 3 and Chapter 6 and an integrative database solution was presented in Chapter 8. However, to develop successful diagnostics applications, we need first to better understand the biology of infecting microbes, and the process of infection itself, to find useful biomarker candidates. In a research context, all the developed tools and solutions presented in this thesis have been developed with the aim to help other OPATHY researchers to achieve such objectives. In an industrial context it is worth to highlight that the deliverables of Chapter 3 and 6 (the applications called SeqEditor RNAseq, VariantSeq and DeNovoSeq) constitute together the basis of a bioinformatics product designated as GPRO suite, which after valorization, has been integrated in the business plan for cloud computing services of Biotechvana the industrial SME partner of OPATHY. This thesis is therefore of multidisciplinar character and its deliverables impact both in the industry and the academy. Next, I will provide an overall discussion of all solutions developed in the presented thesis.

9.1 Recurrent neural network as an efficient tool to enhance RNA secondary structure prediction.

With recent developments in next-generation DNA sequencing technologies, more and more diverse data types are produced, thus introducing many computational challenges that have the potential to be tackled by machine and deep learning applications. In particular, recurrent neural networks have shown promising results in learning complex patterns from sequence dependent input such as machine translation or speech recognition. One particular problem of interest is predicting the structural features of RNA molecules. RNA is a single-stranded sequence which folds into complex secondary and tertiary structures, these structures play a central role in their function and regulation which have important roles in virtually every cellular process. Although their role in the pathogenesis of *Candida* species is unknown, these organisms also encode non-coding transcripts (Sellam *et al.* 2010) and there is an interest within the OPATHY consortium to study their role. Determining the secondary structures of RNAs is central to understanding their function and evolution, as their functions are mediated through the adoption of specific structures that enable RNAs to interact with other molecules.

Recurrent neural network models are practical solutions for a wide class of regression and classification problems involving DNA and RNA sequences. Two datasets were constructed to assess and validate the performance of the RNN models presented in Chapter 5. To assess classification performance; a dataset was constructed from RNA secondary structures downloaded from the RNA STRAND database. And to assess the regression model a synthetic dataset was constructed from oligonucleotide DNA sequences and corresponding melting temperature. Performance results of RNN classification models suggest it was successfully able to learn highly accurate models to predict the preferred state of each nucleotide either being single-stranded or double-stranded in the RNA secondary structure. As well the result from the regression models on the synthetic melting temperature dataset show that RNN was successfully able to learn an accurate representation of the underlying problem. In an effort to provide non-experienced users an easy to use tool to train and deploy recurrent neural network models, we developed rnnXna, a general purpose tool for training and validating RNN models for a wide class of regression and classification problems involving DNA and RNA sequences.

The proposed models to predict the structural features of RNA were used to enhance the final prediction of nextPARS method (Saus *et al.* 2018). nextPARS is a rapid and easy in vitro protocol using Illumina sequencing technology to probe the secondary structure of RNAs experimentally and massively. nextPARS method is based on the parallel specific enzymatic digestion of single-stranded and double-stranded regions directly followed by Illumina library preparation and sequencing. nextPARS provides a computational procedure to go from the sequencing reads to a single score that can be used in downstream analyses. The proposed RNN models were used to build an ensemble classification used as a prior prediction score in nextPARS method to enhance the final prediction score provided by the computational method.

9.2 Optimizing sequencing strategies and available resources for Dual RNA-seq

Dual RNA-seq is an invaluable technique for studying complex transcriptomics of host-pathogen or cohabitant species interactions (Hovhannisyan and Gabaldón 2019), however it comes with additional cost. Dual RNA-seq among other sequencing studies such as sequencing of hybrid genomes, xenografts, mixed species systems, metagenomics, and meta-transcriptomics involve samples containing genetic material from divergent organisms. A common problem for downstream analysis is the possible cross-mapping when reads from one organism map to another organisms' reference genome. Thus, it is important for these studies to identify from which organism each sequencing read originated, and the experimental design should be directed to minimize biases caused by cross-mapping of reads to incorrect source genomes.

Taking this into account, CrossMapper presented in Chapter 4 was developed and designed as a computational tool able to assess cross-mapping prior to sequencing, therefore allowing optimization of experimental design to avoid possible cross-mapping. CrossMapper can be used to perform read simulation and back-mapping of reads originating from such experiments to a pool of any combination of reference genomes, quantifies and reports the cross-mapping rates for each organism.

It is also possible that genetic material from multiple divergent organisms (multiple-organisms samples) can be pooled together into a single sequencing library for sequencing, thus minimizing experimental cost, effort and resources of preparing multiple samples, however it requires a careful planning and assessment of the impact of cross-mapping on downstream analysis. This is important as the costs of sequencing are continuously dropping, the cost of library preparation could become a bottleneck for large projects. In that case, CrossMapper can be a potential application when dealing with large sequencing projects or in sequencing facilities, where many species and samples have to be analysed. In this situation to save resources on library preparation, sequencing facilities can combine the DNA or RNA samples from different species together and sequence them as one sample. With CrossMapper, that experimental design can be assessed, and the most optimal sequencing parameters can be decided in advance.

9.3 Integrative network analysis for biomarkers identification

The OPATHY project has also explored the potential of proteomics technologies to go beyond species identification and provide other relevant information such as potential resistance to antifungals, and the recognition of the disease stage. For example, antifungal therapy failure is a substantial clinical problem, due to the emergence of an increasing number of drug resistant isolates, in addition to the fact that there are relatively few classes of antifungal drugs available on the market. In this thesis cell wall proteome of *C. albicans* was investigated using proteomic data provided by the consortium and obtained by LC-MS/MS technique. The analysis was performed between two groups of *C. albicans* isolates where Echinocandin-resistant *C. albicans* clinical isolates were compared with sensitive isolates and their responses to echinocandin treatment analysed. The objective of the analysis was to find and identify candidate biomarkers of echinocandin resistance that can be utilized in clinical applications.

The data processing pipeline was customized for proteomic data, where data filtering and missing values imputation were performed prior to network construction. Network clustering has revealed that the cell wall proteins in the dataset are divided into seven major groups. A function name was assigned to the different groups, based on the most abundant gene ontology terms using CandidaMine Gene Ontology enrichment tool. The interaction network combined with differential expression analysis performed between two groups of isolates (caspofungin-resistant and caspofungin-susceptible) provided an efficient exploratory analysis tool to identify candidate biomarkers of echinocandin resistance. Our results suggest that instead of a single protein/gene biomarker, the specific echinocandin resistance apparently derives from a panel of different genes/proteins. Results also suggest that drug resistance might involve not only a different cell wall architecture, but also a different response to drugs. Therefore, the identified protein subsets (**Table 7.2**) can be potentially used for a rapid diagnosis of drug resistance in clinical settings albeit they should be validated on a broader panel of clinical isolates to ensure the identified changes in the cell wall glycoproteome are consistent and common traits in drug resistant isolates. A better characterization of the fungal cell wall will be also required for therapeutic purposes, in order to investigate new targets for the future development of antifungal drugs.

9.4 Sequence analysis and primer design tools

SeqEditor (presented in Chapter 3) can be easily managed by any researcher with bioinformatic skills at the user level in order to analyse both nucleotide and protein sequences. SeqEditor has been optimized for the analysis of large sequences such as scaffolds and chromosomes and can work either as a file manager or as a graphical sequence browser thus combining the graphical versatility of GUI applications with a high efficiency for data processing that is similar to that of command line based tools. In fact, SeqEditor is a valuable tool for biological researchers working with reference genomes and transcriptomes. This is granted to its GTF/GFF viewer, which allows to easily mine assemblies and extract data such as exons promoters, gene families that can be used furthermore in downstream analysis or in PCR primers design (Arastehfar *et al.* 2020, Megri *et al.* 2020). On the other hand, the set of tools for singleplex, multiplex primer design and primer pooling makes of SeqEditor a very easy to use application in order to meet growing needs for species-specific primer design to aid in clinical success, especially in developing countries (Arastehfar *et al.* 2019) and where PCR serves as an invaluable diagnostic and identification tool. Remarkably, this is because of the two search algorithms for search and design of species and target specific primers in multiplex experiments implemented in SeqEditor. These two algorithms are unique to SeqEditor and have been successfully validated using five human fungal pathogens. Such a validation makes of SeqEditor a very valuable tools in the field of fungal diagnosis, as species-specific primers are central for the identification of species in many microbiological processes such as yeast pathogens that are responsible for a given infection or for determining antimicrobial susceptibility as well as infection load (Consortium Opathy and Gabaldón 2019). Compared to other sequence analysis software SeqEditor has multiple advantages albeit still some work should be done in future updates. The current version of SeqEditor lacks some features such as graphical tools to read sanger chromatograms or for plotting restriction sites that we are committed to implementing these additional features in further updates of SeqEditor.

9.5 Software integration and GUI for an easy to use and a better user experience to run bioinformatics pipelines

A software application for NGS data analysis no longer consists of a single tool but more properly of a complex computational workflow where distinct applications are sequentially or simultaneously executed. The implementation of these protocols in routine application of NGS data analysis is still a challenge due to the technical complexity of the pipelines whose management requires bioinformatic experts with advanced skills on the command line interface or scripting languages. Even though in a research context where an expert bioinformatician is involved this is still a challenging task requiring custom methods to automate, deploy and manage such workflow to ensure better productivity and reproducibility of the workflow. In general, when it comes to a clinical application, a software tool should be characterized with interactive and easy to use GUI based implementation for a better user experience and better productivity. The pipeline tools introduced in Chapter 6 RNAseq, VariantSeq and DeNovoSeq, can be used in research settings and can be the basis for developing applications in clinical diagnosis. In particular; RNASeq and VariantSeq software tools are two examples of dedicated pipeline software tools for resequencing data that can be very useful in clinical settings for analysing transcriptomes, exome, genome or amplicon samples of patients, and/or pathogens in order to find biomarkers based on deregulated genes or in SNP or INDELS variants. On the other hand, DeNovoSeq manages analysis of data provided by *de novo* sequencing approaches and is more appropriate to be utilized in research settings to assemble genomes of newly sequenced species or to improve existing pathogens' genomes. Similarly, SeqEditor introduced in chapter 3 has a wide range of applications in research settings as a tool for exploratory sequences analysis, including primer design and/or mining of sequence patterns like ORFs genes etc.

Indeed, pipeline tools developed in the course of this thesis provide an alternative way to run bioinformatics analysis in the form of interactive GUI instead of the conventional CLI without sacrificing some level of extensibility and customization, thus enabling clinicians or wet lab personnel to have easy access to such analysis. Pipeline tools are built upon a pipeline framework that support extensibility; however current implementation only enables extension by an application developer. A more desirable feature would be to enable external bioinformaticians to extend the pipeline tools by adding more new tools or pipelines, a feature that we will implement in further releases of these tools. This can be achieved by designing Domain Specific Languages DSL with a visual design which can be utilized by anyone to write a description of a new tool, then the framework will translate and build the corresponding GUI, and handle running and manage the new addition. DSL/visual designer is an approach widely used to enable non-developers to extend existing applications without the need for developing it by themselves. This would enable bioinformaticians to design new tools/pipelines with little knowledge about software development in general and ship and deploy their custom application to be used by other bioinformaticians, clinical laboratory personals or wet lab personnel.

Another important feature is reproducibility, any analysis performed by the software tools should be reproducible. To ensure reproducibility, that the software tools' version should be the same while repeating any analysis, the current implementation of the pipeline tools enables versioning at the level of the whole software. However, in future updates a more reliable feature would be to support versions at the level of individual tools that are implemented within the pipeline.

While current implementation enables tracking and basic errors reporting for the users. It would be desirable that the tracking system is able to auto recover in case of simple errors or suggest possible solutions where more complex or severe errors occur. Error recovery of the tracking system would enable improved users' productivity especially when used by clinicians. This can be supported by enabling expert systems trained to resolve simple errors that normally happen when running an analysis, this can be trained to behave like an expert bioinformaticians. In that case of more severe or unknown errors, for which the expert system is not trained to resolve, the system will suggest possible solutions based on external knowledge databases. This can be achieved by training the system to mine for answers in online forums specialized in bioinformatics such as SEQAnswers (Li *et al.* 2012), harnessing the power of social sharing in which a large number of expert bioinformaticians share their knowledge and experiences. This would enable a greater level of knowledge and experience transfer.

9.6 Databases integration for better insights

Knowledge databases are essential because they provide the required information to generate the insights contributing to better understand the biology of pathogens. We believe that CandidaMine (presented in Chapter 8) is an invaluable dedicated resource for the broad research and clinical community working on *Candida* diseases.

Aiming for a complex and integrative data analysis, we developed CandidaMine to integrate datasets from various studies and resources to facilitate exploration, analysis, and interpretation of *Candida* related studies. CandidaMine currently integrates various datasets for 5 *Candida* yeast pathogens which focus on the most common species causing Candidiasis [*C. albicans, C. glabrata, C. parapsilosis* and *C. tropicalis*] (Gabaldón *et al.* 2016) and the emerging multidrug-resistant species *e.g. C. auris* (Rhodes and Fisher 2019). Furthermore, we plan to scale CandidaMine horizontally *i.e.* add more strains and vertically *i.e.* add more datasets for a given strain, such scaling will introduce many challenges ahead of us. With further scaling, we need to maintain a steady performance of CandidaMine and introduce a reliable and robust mechanism for updating and maintaining all stored dataset in CandidaMine.

CandidaMine provides easy to use tools to search and query all data types stored in the data warehouse. CandidaMine provides efficient tools to suit all types of users with different levels of expertise. For novice users, CandidaMine provides easy to use tools such as list analysis and ready to use template queries for a wide range of analysis. For intermediate and more experienced users, CandidaMine provides custom query builder for designing more complex queries and API access for programmatically performing more complex analysis. Major contribution of CandidaMine is the expression profiles and variant calling datasets. Where all SRA runs related to *Candida* pathogens stored in CandidaMine were downloaded and proceeded to generate the expression profiles and variant calling. Those datasets integrated with all metadata from the SRA database provide researchers with a variety of options to integrate this data and use it to get more insight in their analysis.

CandidaMine and the distinct software tools developed in this thesis complement each other. In the current implementation there is no technical integration between them, but users may easily enable a manual integration where they can retrieve data from CandidaMine and utilize it within the software tools. However, in future updates of the software tools, we will enable such integration, for example, the SeqEditor tool can be adapted to directly use genomes and annotations of CanidadaMine by utilizing Intermine API. Such integration will enable us to apply all implemented tools in SeqEditor to be performed on pathogen data from CandidaMine for example designing primers. Additionally, using variant calling dataset in CandidaMine, SeqEditor primers design tools implementation can be adjusted to design mutation-specific primers if possible, for any given strain, which have potential applications in identifying virulence or drug resistant strain. Also RNASeq or VariantSeq tools can access annotations provided by CandidaMine directly via API to download reference material or training sets or for functional annotation of the results; for example, downloading GOs and functional descriptions in the case of deregulated genes or annotations of phenotypic effects in the case of variants.

Developing diagnostic applications requires understanding host-pathogen interaction, which requires data from both sides - human and the pathogen under study (fungal in our case). While CandidaMine provides data for the pathogen side, we still need access to genomic data from the human side which can be provided by HumanMine (Smith *et al.* 2012) which is an integrated database of *Homo sapiens* genomic data. Both CandidaMine and HumanMine are powered by InterMine, thus enabling unified API access. This will enable further updates of our tools to use and integrate data from both mines to perform integrative host-pathogen analysis or even the development of newly designed tools dedicated for such analysis.

10 Conclusions

- Lack of appropriate, easy-to-use tools limits the possibilities of harnessing next generation sequencing, proteomics and other omics technologies in the field of study and diagnosis of yeast infections. This thesis has contributed to fill in this important gap.

- We have proposed and shown the utility of novel computational approaches (and build their corresponding tools) to i) develop recurrent neural network models for the prediction of sequence properties such as secondary RNA structure or melting temperature of a DNA duplex (rnnXna tool), and ii) to predict the level of cross-mapping in sequencing of complex samples or pools (CrossMapper tool).

- SeqEditor was developed and implemented as a GUI based tool for efficient sequence analysis including a set of efficient tools for primer design and tools for exploring, mining and extracting information for genomic annotations.

- Pipeline framework and tools were developed and designed to provide a fast and easy way to implement, design, deploy and run bioinformatics pipelines with GUI based tools for a better user experience. The pipeline framework was used to implement the following custom applications:

    - ➢ RNASeq was developed and implemented as a dedicated GUI based tool to run and manage pipelines and workflows for differential expression and enrichment analyses.

    - ➢ VariantSeq was developed and implemented as a dedicated GUI based tool to run and manage pipelines and workflows for variant calling and annotation of Single

  Nucleotide Polymorphisms (SNPs) and small insertions/deletions (Indels).

  - DeNovoSeq was developed and implemented as a dedicated GUI based tool to run and manage pipelines and workflows for assembly/scaffolding and annotation of new genomes and/or transcriptomes with no previous reference sequence.

- Network analysis is an important approach for analysing complex systems. Combining network analysis with other types of analysis such differential expression analysis can lead to robust methods for identifying key important genes/proteins in the network. An integrative network analysis was designed and applied to proteomics data obtained by LC-MS/MS. Using interaction networks combined with differential expression analysis, we were able to identify a set of proteins that can be useful as biomarkers in diagnosing drug resistance in *C. albicans* isolates.

- Integration of disparate types of data and information is a complex task that needs to cope with the volume of data and speed at which it is generated. CandidaMine was developed as an integrative omics data warehouse focusing on the most common *Candida* species causing Candidiasis and mutli-drug resistance strains.

Appendix A: Publication List

Hafez, Ahmed, Hrant Hovhannisyan, Miquel Àngel Schikora-Tamarit, Manuel Molina, Carlos Llorens and Toni Gabaldón. "CandidaMine, an integrative omics database for Candida yeast pathogens." *(In preparation)*.

Hafez, Ahmed, Ricardo Futami, Beatriz Soriano, Francisco J. Roig, Ana Miguel, Aya A. Elsayed, Ricardo Ramos-Ruiz, Miguel A. Torres-Font, Fernando Naya-Català, Josep Calduch-Giner, Lucia Trilla-Fuertes, Angelo Gamez-Pozo, Vicente Arnau, Jaume Perez-Sánchez, Jose M. Sempere, Toni Gabaldón, Carlos Llorens. "Applications and pipeline infrastructure of the GPRO suite for RNASeq, DeNovoSeq and VariantSeq analysis." *(In preparation)*.

Hafez, Ahmed, Essam H. Houssein, Carlos Llorens, Toni Gabaldón. "Recurrent neural networks for classification and regression problems in DNA and RNA sequences with rnnXna." *(In preparation)*.

Buda De Cesare, Giuseppe, **Ahmed Hafez**, David Stead , Carlos Llorens and Carol A. Munro "Biomarkers of caspofungin resistance in C. albicans isolates: a proteomic approach." *(In preparation)*.

Hafez, Ahmed, Ricardo Futami, Amir Arastehfar, Farnaz Daneshnia, Ana Miguel, Francisco J. Roig, Beatriz Soriano, Jaume Perez-Sánchez, Teun Boekhout, Toni Gabaldón, and Carlos Llorens. 2020. "SeqEditor: an application for primer design and sequence analysis with or without GTF/GFF files." *Bioinformatics*. doi: 10.1093/bioinformatics/btaa903.

Hovhannisyan, Hrant*, **Ahmed Hafez***, Carlos Llorens, and Toni Gabaldón. 2020. "CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies." *Bioinformatics* 36 (3):925-927. doi: 10.1093/bioinformatics/btz626.

Arastehfar, Amir, Farnaz Daneshnia, **Ahmed Hafez**, Sadegh Khodavaisy, Mohammad-Javad Najafzadeh, Arezoo Charsizadeh, Hossein Zarrinfar, Mohammadreza Salehi, Zahra Zare Shahrabadi, Elahe Sasani, Kamiar Zomorodian, Weihua Pan, Ferry Hagen, Macit Ilkit, Markus Kostrzewa, and Teun Boekhout. 2020. "Antifungal susceptibility, genotyping, resistance mechanism, and clinical

profile of Candida tropicalis blood isolates." *Med. Mycol.* 58 (6):766-773. doi: 10.1093/mmy/myz124.

Arastehfar, Amir, Süleyha Hilmioğlu-Polat, Farnaz Daneshnia, **Ahmed Hafez**, Mohammadreza Salehi, Furkan Polat, Melike Yaşar, Nazlı Arslan, Tuğrul Hoşbul, Nevzat Ünal, Dilek Yeşim Metin, Şaban Gürcan, Asuman Birinci, Ayşe Nedret Koç, Weihua Pan, Macit Ilkit, David S. Perlin, and Cornelia Lass-Flörl. 2020. "Recent Increase in the Prevalence of Fluconazole-Non-susceptible Candida tropicalis Blood Isolates in Turkey: Clinical Implication of Azole-Non-susceptible and Fluconazole Tolerant Phenotypes and Genotyping." *Frontiers in Microbiology* 11. doi: 10.3389/fmicb.2020.587278.

Megri, Youcef, Amir Arastehfar, Teun Boekhout, Farnaz Daneshnia, Caroline Hörtnagl, Bettina Sartori, **Ahmed Hafez**, Weihua Pan, Cornelia Lass-Flörl, and Boussad Hamrioui. 2020. "Candida tropicalis is the most prevalent yeast species causing candidemia in Algeria: the urgent need for antifungal stewardship and infection control measures." *Antimicrob. Resist. Infect. Control* 9 (1):50. doi: 10.1186/s13756-020-00710-z.

Consortium Opathy (Inluding **Ahmed Hafez**), and Toni Gabaldón. 2019. "Recent trends in molecular diagnostics of yeast infections: from PCR to NGS." *FEMS Microbiol. Rev.* 43 (5):517-547. doi: 10.1093/femsre/fuz015.

Saus, Ester, Jesse R. Willis, Leszek P. Pryszcz, **Ahmed Hafez**, Carlos Llorens, Heinz Himmelbauer, and Toni Gabaldón. 2018. "nextPARS: parallel probing of RNA structures in Illumina." *RNA* 24 (4):609-619. doi: 10.1261/rna.063073.117.

Intellectual property:

Title: GPRO Suite. Request NO: V-959-20, ENTRY NUMBER: pending, Priority date: 23-11-2020, Entity: BIOTECH VANA S.L, Country: SPAIN, Companies exploiting IT: BIOTECH VANA S.L (software).

*equal contribution

Appendix B: Online Resources

- **RNASeq**
  - ➢ https://gpro.biotechvana.com/download/RNAseq
  - ➢ https://gpro.biotechvana.com/tool/rnaseq/manual

- **DeNovoSeq**
  - ➢ https://gpro.biotechvana.com/download/DeNovoSeq
  - ➢ https://gpro.biotechvana.com/tool/denovoseq/manual

- **VariantSeq**
  - ➢ https://gpro.biotechvana.com/download/VariantSeq
  - ➢ https://gpro.biotechvana.com/tool/variantseq/manual

- **GPRO-Server**
  - ➢ https://gpro.biotechvana.com/tool/gpro-server/manual

- **SeqEditor**
  - https://gpro.biotechvana.com/download/SeqEditor
  - https://gpro.biotechvana.com/tool/seqeditor/manual

- **CrossMapper**
  - https://github.com/Gabaldonlab/crossmapper

- **rnnXna**
  - https://github.com/Gabaldonlab/rnnXna

- **CandidaMine**
  - http://candidamine.org/candidamine/begin.do
  - https://candidamine.readthedocs.io

References

Achkar, Jacqueline M., and Bettina C. Fries. 2010. "Candida infections of the genitourinary tract." *Clin. Microbiol. Rev.* 23 (2):253-273. doi: 10.1128/CMR.00076-09.

Afgan, Enis, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Cech, John Chilton, Dave Clements, Nate Coraor, Björn A. Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. 2018. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update." *Nucleic Acids Res.* 46 (W1):W537-W544. doi: 10.1093/nar/gky379.

Ahmed, Abdalla O. A., G. Sybren De Hoog, and Wendy W. J. van de Sande. 2015. "Fungi Causing Eumycotic Mycetoma." In *Manual of Clinical Microbiology*, edited by James H. Jorgensen, Karen C. Carroll, Guido Funke, Michael A. Pfaller, Marie Louise Landry, Sandra S. Richter and David W. Warnock, 2173-2187. Washington, DC, USA: ASM Press.

Alexander, M. D. Barbara. 2017. *Reference Method for Broth Dilution Antifungal Susceptibility Testing of Filamentous Fungi, 3rd Edition*: Clin Lab Stand Inst.

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning." *Nat. Biotechnol.* 33 (8):831-838. doi: 10.1038/nbt.3300.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic local alignment search tool." *J. Mol. Biol.* 215 (3):403-410. doi: 10.1016/S0022-2836(05)80360-2.

Ammari, Mais G., Cathy R. Gresham, Fiona M. McCarthy, and Bindu Nanduri. 2016. "HPIDB 2.0: a curated database for host–pathogen interactions." *Database* 2016:baw103. doi: 10.1093/database/baw103.

Amorim-Vaz, Sara, Van Du T. Tran, Sylvain Pradervand, Marco Pagni, Alix T. Coste, and Dominique Sanglard. 2015. "RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens In

Vivo Applied to the Fungus Candida albicans." *MBio* 6 (5):e00942-15. doi: 10.1128/mBio.00942-15.

Anders, S., P. T. Pyl, and W. Huber. 2015. "HTSeq--a Python framework to work with high-throughput sequencing data." *Bioinformatics* 31 (2):166-169. doi: 10.1093/bioinformatics/btu638.

Andrews, Simon, and Others. 2010. "FastQC: a quality control tool for high throughput sequence data."

Andronescu, Mirela, Vera Bereg, Holger H. Hoos, and Anne Condon. 2008. "RNA STRAND: the RNA secondary structure and statistical analysis database." *BMC Bioinformatics* 9:340. doi: 10.1186/1471-2105-9-340.

Arastehfar, Amir, Farnaz Daneshnia, Ahmed Hafez, Sadegh Khodavaisy, Mohammad-Javad Najafzadeh, Arezoo Charsizadeh, Hossein Zarrinfar, Mohammadreza Salehi, Zahra Zare Shahrabadi, Elahe Sasani, Kamiar Zomorodian, Weihua Pan, Ferry Hagen, Macit Ilkit, Markus Kostrzewa, and Teun Boekhout. 2020. "Antifungal susceptibility, genotyping, resistance mechanism, and clinical profile of Candida tropicalis blood isolates." *Med. Mycol.* 58 (6):766-773. doi: 10.1093/mmy/myz124.

Arastehfar, Amir, Brian L. Wickes, Macit Ilkit, David H. Pincus, Farnaz Daneshnia, Weihua Pan, Wenjie Fang, and Teun Boekhout. 2019. "Identification of Mycoses in Developing Countries." *J Fungi (Basel)* 5 (4). doi: 10.3390/jof5040090.

Araujo, Ricardo. 2014. "Towards the Genotyping of Fungi: Methods, Benefits and Challenges." *Current Fungal Infection Reports* 8 (3):203-210. doi: 10.1007/s12281-014-0190-1.

Arendrup, M. C., T. Boekhout, M. Akova, J. F. Meis, O. A. Cornely, O. Lortholary, Microbiology European Society of Clinical, Group Infectious Diseases Fungal Infection Study, and Mycology European Confederation of Medical. 2014. "ESCMID and ECMM joint clinical guidelines for the diagnosis and management of rare invasive yeast infections." *Clin. Microbiol. Infect.* 20 Suppl 3:76-98. doi: 10.1111/1469-0691.12360.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000.

"Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat. Genet.* 25 (1):25-29. doi: 10.1038/75556.

Attwood, Lucy O., Michelle J. Francis, John Hamblin, Tony M. Korman, Julian Druce, and Maryza Graham. 2020. "Clinical evaluation of AusDiagnostics SARS-CoV-2 multiplex tandem PCR assay." *J. Clin. Virol.* 128:104448. doi: 10.1016/j.jcv.2020.104448.

Aznar-Marin, Pilar, Fátima Galan-Sanchez, Pilar Marin-Casanova, Pedro García-Martos, and Manuel Rodríguez-Iglesias. 2016. "Candida nivariensis as a New Emergent Agent of Vulvovaginal Candidiasis: Description of Cases and Review of Published Studies." *Mycopathologia* 181 (5-6):445-449. doi: 10.1007/s11046-015-9978-y.

Badiee, Parisa, and Zahra Hashemizadeh. 2014. "Opportunistic invasive fungal infections: diagnosis & clinical management." *Indian J. Med. Res.* 139 (2):195-204.

Bairoch, A. 2000. "The ENZYME database in 2000." *Nucleic Acids Res.* 28 (1):304-305. doi: 10.1093/nar/28.1.304.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. 2012. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *J. Comput. Biol.* 19 (5):455-477. doi: 10.1089/cmb.2012.0021.

Basenko, Evelina Y., Jane A. Pulman, Achchuthan Shanmugasundram, Omar S. Harb, Kathryn Crouch, David Starns, Susanne Warrenfeltz, Cristina Aurrecoechea, Christian J. Stoeckert, Jr., Jessica C. Kissinger, David S. Roos, and Christiane Hertz-Fowler. 2018. "FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes." *J Fungi (Basel)* 4 (1). doi: 10.3390/jof4010039.

Beaton, Wayne, and Jeff McAffer. 2007. "Eclipse Rich Client Platform."

Benedict, Kaitlin, Brendan R. Jackson, Tom Chiller, and Karlyn D. Beer. 2019. "Estimation of Direct Healthcare Costs of Fungal Diseases in the United States." *Clin. Infect. Dis.* 68 (11):1791-1797. doi: 10.1093/cid/ciy776.

Berkman, Paul J., Kaitao Lai, Michal T. Lorenc, and David Edwards. 2012. "Next-generation sequencing applications for wheat crop improvement." *Am. J. Bot.* 99 (2):365-371. doi: 10.3732/ajb.1100309.

Berkow, Elizabeth L., and Shawn R. Lockhart. 2017. "Fluconazole resistance in Candida species: a current perspective." *Infect. Drug Resist.* 10:237-245. doi: 10.2147/IDR.S118892.

Bersanelli, Matteo, Ettore Mosca, Daniel Remondini, Gastone Castellani, and Luciano Milanesi. 2016. "Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules." *Sci. Rep.* 6:34841. doi: 10.1038/srep34841.

Binnicker, M. J., D. J. Jespersen, J. E. Bestrom, and L. O. Rollins. 2012. "Comparison of four assays for the detection of cryptococcal antigen." *Clin. Vaccine Immunol.* 19 (12):1988-1990. doi: 10.1128/CVI.00446-12.

Biswas, Chayanika, Sharon C. A. Chen, Catriona Halliday, Elena Martinez, Rebecca J. Rockett, Qinning Wang, Verlaine J. Timms, Rajat Dhakal, Rosemarie Sadsad, Karina J. Kennedy, Geoffrey Playford, Deborah J. Marriott, Monica A. Slavin, Tania C. Sorrell, and Vitali Sintchenko. 2017. "Whole Genome Sequencing of Candida glabrata for Detection of Markers of Antifungal Drug Resistance." *J. Vis. Exp.* (130). doi: 10.3791/56714.

Bittinger, Kyle, 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. 2014. "Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing." *Genome Biology* 15 (10). doi: 10.1186/s13059-014-0487-y.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15):2114-2120. doi: 10.1093/bioinformatics/btu170.

Bond, Stephen R., Karl E. Keat, Sofia N. Barreira, and Andreas D. Baxevanis. 2017. "BuddySuite: Command-Line Toolkits for Manipulating Sequences, Alignments, and Phylogenetic Trees." *Mol. Biol. Evol.* 34 (6):1543-1546. doi: 10.1093/molbev/msx089.

Børsting, Claus, and Niels Morling. 2015. "Next generation sequencing and its applications in forensic genetics." *Forensic Sci. Int. Genet.* 18:78-89. doi: 10.1016/j.fsigen.2015.02.002.

Botschuijver, Sara, Guus Roeselers, Evgeni Levin, Daisy M. Jonkers, Olaf Welting, Sigrid E. M. Heinsbroek, Heleen H. de Weerd, Teun Boekhout, Matteo Fornai, Ad A. Masclee, Frank H. J. Schuren, Wouter J. de Jonge, Jurgen Seppen, and René M. van den Wijngaard. 2017. "Intestinal Fungal Dysbiosis Is Associated With Visceral Hypersensitivity in Patients With Irritable Bowel Syndrome and Rats." *Gastroenterology* 153 (4):1026-1039. doi: 10.1053/j.gastro.2017.06.004.

Brown, Gordon D., David W. Denning, Neil A. R. Gow, Stuart M. Levitz, Mihai G. Netea, and Theodore C. White. 2012. "Hidden killers: human fungal infections." *Sci. Transl. Med.* 4 (165):165rv13. doi: 10.1126/scitranslmed.3004404.

Brown, Silas S., Yun-Wen Chen, Ming Wang, Alexandra Clipson, Eguzkine Ochoa, and Ming-Qing Du. 2017. "PrimerPooler: automated primer pooling to prepare library for targeted sequencing." *Biol Methods Protoc* 2 (1):bpx006. doi: 10.1093/biomethods/bpx006.

Buels, Robert, Eric Yao, Colin M. Diesh, Richard D. Hayes, Monica Munoz-Torres, Gregg Helt, David M. Goodstein, Christine G. Elsik, Suzanna E. Lewis, Lincoln Stein, and Ian H. Holmes. 2016. "JBrowse: a dynamic web platform for genome visualization and analysis." *Genome Biol.* 17:66. doi: 10.1186/s13059-016-0924-1.

Burns, John A., Huanjia Zhang, Elizabeth Hill, Eunsoo Kim, and Ryan Kerney. 2017. "Transcriptome analysis illuminates the nature of the intracellular interaction in a vertebrate-algal symbiosis." doi: 10.7554/eLife.22054.

Byrne, Kevin P., and Kenneth H. Wolfe. 2005. "The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species." *Genome Res.* 15 (10):1456-1461. doi: 10.1101/gr.3672305.

Cherry, J. Michael, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J.

Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng, and Edith D. Wong. 2012. "Saccharomyces Genome Database: the genomics resource of budding yeast." *Nucleic Acids Res.* 40 (Database issue):D700-5. doi: 10.1093/nar/gkr1029.

Chiquet, Julien, Stephane Robin, and Mahendra Mariadassou. 2019. "Variational Inference for sparse network reconstruction from count data." 97:1162-1171.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi: 10.3115/v1/d14-1179.

Choi, Edward, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2017. "Using recurrent neural network models for early detection of heart failure onset." *J. Am. Med. Inform. Assoc.* 24 (2):361-370. doi: 10.1093/jamia/ocw112.

Chollet, François. 2015. "Keras." Last Modified 2020. https://keras.io/.

Chumpitazi, Bernabé F. F., Bernadette Lebeau, Odile Faure-Cognet, Rebecca Hamidfar-Roy, Jean-François Timsit, Patricia Pavese, Anne Thiebaut-Bertrand, Jean-Louis Quesada, Hervé Pelloux, and Claudine Pinel. 2014. "Characteristic and clinical relevance of Candida mannan test in the diagnosis of probable invasive candidiasis." *Med. Mycol.* 52 (5):462-471. doi: 10.1093/mmy/myu018.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." *Fly* 6 (2):80-92. doi: 10.4161/fly.19695.

Clauset, Aaron, M. E. J. Newman, and Cristopher Moore. 2004. "Finding community structure in very large networks." *arXiv [cond-mat.stat-mech]*.

Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. "Ribosomal Database Project: data and tools for high throughput rRNA analysis." *Nucleic Acids Res.* 42 (Database issue):D633-42. doi: 10.1093/nar/gkt1244.

Consortium Opathy, and Toni Gabaldón. 2019. "Recent trends in molecular diagnostics of yeast infections: from PCR to NGS." *FEMS Microbiol. Rev.* 43 (5):517-547. doi: 10.1093/femsre/fuz015.

Contrino, Sergio, Richard N. Smith, Daniela Butano, Adrian Carr, Fengyuan Hu, Rachel Lyne, Kim Rutherford, Alex Kalderimis, Julie Sullivan, Seth Carbon, Ellen T. Kephart, Paul Lloyd, E. O. Stinson, Nicole L. Washington, Marc D. Perry, Peter Ruzanov, Zheng Zha, Suzanna E. Lewis, Lincoln D. Stein, and Gos Micklem. 2012. "modMine: flexible access to modENCODE data." *Nucleic Acids Res.* 40 (Database issue):D1082-8. doi: 10.1093/nar/gkr921.

Cornely, O. A., M. Bassetti, T. Calandra, J. Garbino, B. J. Kullberg, O. Lortholary, W. Meersseman, M. Akova, M. C. Arendrup, S. Arikan-Akdagli, J. Bille, E. Castagnola, M. Cuenca-Estrella, J. P. Donnelly, A. H. Groll, R. Herbrecht, W. W. Hope, H. E. Jensen, C. Lass-Flörl, G. Petrikkos, M. D. Richardson, E. Roilides, P. E. Verweij, C. Viscoli, A. J. Ullmann, and Escmid Fungal Infection Study Group. 2012. "ESCMID* guideline for the diagnosis and management of Candida diseases 2012: non-neutropenic adult patients." *Clin. Microbiol. Infect.* 18 Suppl 7:19-37. doi: 10.1111/1469-0691.12039.

Cornish, Adam, and Chittibabu Guda. 2015. "A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference." *BioMed Research International* 2015:1-11. doi: 10.1155/2015/456479.

Cristianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*: Cambridge University Press.

Crump, J. A., and P. J. Collignon. 2000. "Intravascular catheter-associated infections." *Eur. J. Clin. Microbiol. Infect. Dis.* 19 (1):1-8. doi: 10.1007/s100960050001.

Cruz, José Almeida, and Eric Westhof. 2009. "The dynamic landscapes of RNA architecture." *Cell* 136 (4):604-609. doi: 10.1016/j.cell.2009.02.003.

Cuenca-Estrella, M., P. E. Verweij, M. C. Arendrup, S. Arikan-Akdagli, J. Bille, J. P. Donnelly, H. E. Jensen, C. Lass-Flörl, M. D. Richardson, M. Akova, M. Bassetti, T. Calandra, E. Castagnola, O. A. Cornely, J. Garbino, A. H. Groll, R. Herbrecht, W. W. Hope, B. J. Kullberg, O. Lortholary, W. Meersseman, G. Petrikkos, E. Roilides, C. Viscoli, A. J. Ullmann, and Escmid Fungal Infection Study Group. 2012. "ESCMID* guideline for the diagnosis and management of Candida diseases 2012: diagnostic procedures." *Clin. Microbiol. Infect.* 18 Suppl 7:9-18. doi: 10.1111/1469-0691.12038.

Culibrk, Luka, Carys A. Croft, and Scott J. Tebbutt. 2016. "Systems Biology Approaches for Host-Fungal Interactions: An Expanding Multi-Omics Frontier." *OMICS* 20 (3):127-138. doi: 10.1089/omi.2015.0185.

Cunningham, Fiona, Premanand Achuthan, Wasiu Akanni, James Allen, M. Ridwan Amode, Irina M. Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Carla Cummins, Claire Davidson, Kamalkumar Jayantilal Dodiya, Astrid Gall, Carlos García Girón, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Mike Kay, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, José C. Marugán, Thomas Maurel, Aoife C. McMahon, Benjamin Moore, Joannella Morales, Jonathan M. Mudge, Michael Nuhn, Denye Ogeh, Anne Parker, Andrew Parton, Mateus Patricio, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Eloise Stapleton, Marek Szuba, Kieron Taylor, Glen Threadgold, Anja Thormann, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nick Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Andrew D. Yates, Daniel R. Zerbino, and Paul Flicek. 2019. "Ensembl 2019." *Nucleic Acids Res.* 47 (D1):D745-D751. doi: 10.1093/nar/gky1113.

Cuomo, Christina A. 2017. "Harnessing Whole Genome Sequencing in Medical Mycology." *Curr. Fungal Infect. Rep.* 11 (2):52-59. doi: 10.1007/s12281-017-0276-7.

da Matta, Daniel Archimedes, Ana Carolina Remondi Souza, and Arnaldo Lopes Colombo. 2017. "Revisiting Species Distribution and Antifungal Susceptibility of Candida Bloodstream Isolates from

Latin American Medical Centers." *J Fungi (Basel)* 3 (2). doi: 10.3390/jof3020024.

Danecek, Petr, and Shane A. McCarthy. 2017. "BCFtools/csq: haplotype-aware variant consequences." *Bioinformatics* 33 (13):2037-2039. doi: 10.1093/bioinformatics/btx100.

Dannemiller, Karen C., Darryl Reeves, Kyle Bibby, Naomichi Yamamoto, and Jordan Peccia. 2014. "Fungal High-throughput Taxonomic Identification tool for use with Next-Generation Sequencing (FHiTINGS)." *Journal of Basic Microbiology* 54 (4):315-321. doi: 10.1002/jobm.201200507.

Darty, Kévin, Alain Denise, and Yann Ponty. 2009. "VARNA: Interactive drawing and editing of the RNA secondary structure." *Bioinformatics* 25 (15):1974-1975. doi: 10.1093/bioinformatics/btp250.

Davidson, Nadia M., and Alicia Oshlack. 2014. "Corset: enabling differential gene expression analysis for de novoassembled transcriptomes." *Genome Biology* 15 (7). doi: 10.1186/s13059-014-0410-6.

Dawyndt, P., M. Vancanneyt, H. De Meyer, and J. Swings. 2005. "Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources." *IEEE Transactions on Knowledge and Data Engineering* 17 (8):1111-1126. doi: 10.1109/tkde.2005.131.

De Cremer, Kaat, Janick Mathys, Christine Vos, Lutz Froenicke, Richard W. Michelmore, Bruno P. A. Cammue, and Barbara De Coninck. 2013. "RNAseq-based transcriptome analysis of Lactuca sativa infected by the fungal necrotroph Botrytis cinerea." *Plant Cell Environ.* 36 (11):1992-2007. doi: 10.1111/pce.12106.

den Dunnen, Johan T., Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, and Peter E. M. Taschner. 2016. "HGVS Recommendations for the Description of Sequence Variants: 2016 Update." *Hum. Mutat.* 37 (6):564-569. doi: 10.1002/humu.22981.

Desnos-Ollivier, Marie, Stéphane Bretagne, Dorothée Raoux, Damien Hoinard, Françoise Dromer, Eric Dannaoui, and Testing European Committee on Antibiotic Susceptibility. 2008. "Mutations in the fks1 gene in Candida albicans, C. tropicalis, and C. krusei correlate with elevated caspofungin MICs uncovered in AM3 medium using the method of the European Committee on Antibiotic Susceptibility Testing." *Antimicrob. Agents Chemother.* 52 (9):3092-3098. doi: 10.1128/AAC.00088-08.

Deurenberg, Ruud H., Erik Bathoorn, Monika A. Chlebowicz, Natacha Couto, Mithila Ferdous, Silvia García-Cobos, Anna M. D. Kooistra-Smid, Erwin C. Raangs, Sigrid Rosema, Alida C. M. Veloo, Kai Zhou, Alexander W. Friedrich, and John W. A. Rossen. 2017. "Application of next generation sequencing in clinical microbiology and infection prevention." *J. Biotechnol.* 243:16-24. doi: 10.1016/j.jbiotec.2016.12.022.

Di Resta, Chiara, Silvia Galbiati, Paola Carrera, and Maurizio Ferrari. 2018. "Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities." *Ejifcc* 29 (1):4.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29 (1):15-21. doi: 10.1093/bioinformatics/bts635.

Douglas, C. M. 2001. "Fungal beta(1,3)-D-glucan synthesis." *Med. Mycol.* 39 Suppl 1:55-66. doi: 10.1080/mmy.39.1.55.66.

Durmuş, Saliha, Tunahan Çakır, Reinhard Guthke, Emrah Nikerel, and Arzucan Özgür. 2016. *Computational Systems Biology of Pathogen-Host Interactions*: Frontiers Media SA.

Dutton, L. C., K. H. Paszkiewicz, R. J. Silverman, P. R. Splatt, S. Shaw, A. H. Nobbs, R. J. Lamont, H. F. Jenkinson, and M. Ramsdale. 2016. "Transcriptional landscape of trans-kingdom communication betweenCandida albicansandStreptococcus gordonii." *Molecular Oral Microbiology* 31 (2):136-161. doi: 10.1111/omi.12111.

Eilbeck, Karen, Suzanna E. Lewis, Christopher J. Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. 2005. "The Sequence Ontology: a tool for the unification of genome annotations." *Genome Biol.* 6 (5):R44. doi: 10.1186/gb-2005-6-5-r44.

Ellepola, Arjuna N. B., and Christine J. Morrison. 2005. "Laboratory diagnosis of invasive candidiasis." *J. Microbiol.* 43 Spec No:65-84.

Enguita, Francisco, Marina Costa, Ana Fusco-Almeida, Maria Mendes-Giannini, and Ana Leitão. 2016. "Transcriptomic Crosstalk between Fungal Invasive Pathogens and Their Host Cells: Opportunities and Challenges for Next-Generation Sequencing Methods." *Journal of Fungi* 2 (1):7. doi: 10.3390/jof2010007.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* 32 (19):3047-3048. doi: 10.1093/bioinformatics/btw354.

Farmakiotis, Dimitrios, and Dimitrios P. Kontoyiannis. 2017. "Epidemiology of antifungal resistance in human pathogenic yeasts: current viewpoint and practical recommendations for management." *Int. J. Antimicrob. Agents* 50 (3):318-324. doi: 10.1016/j.ijantimicag.2017.05.019.

Finn, Robert D., Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin-Yu Chang, Zsuzsanna Dosztányi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L. Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A. Natale, Marco Necci, Gift Nuka, Christine A. Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C. Potter, Neil D. Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D. Thomas, Silvio C. E. Tosatto, Cathy H. Wu, Ioannis Xenarios, Lai-Su Yeh, Siew-Yit Young, and Alex L. Mitchell. 2017. "InterPro in 2017-beyond protein family and domain annotations." *Nucleic Acids Res.* 45 (D1):D190-D199. doi: 10.1093/nar/gkw1107.

Fitzpatrick, David A., Peadar O'Gaora, Kevin P. Byrne, and Geraldine Butler. 2010. "Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser." *BMC Genomics* 11:290. doi: 10.1186/1471-2164-11-290.

Fonzi, W. A. 1999. "PHR1 and PHR2 of Candida albicans encode putative glycosidases required for proper cross-linking of beta-1,3- and beta-1,6-glucans." *J. Bacteriol.* 181 (22):7070-7079. doi: 10.1128/JB.181.22.7070-7079.1999.

Fonzi, W. A., and M. Y. Irwin. 1993. "Isogenic strain construction and gene mapping in Candida albicans." *Genetics* 134 (3):717-728.

Forman, George, and Martin Scholz. 2010. "Apples-to-apples in cross-validation studies." *ACM SIGKDD Explorations Newsletter* 12 (1):49-57. doi: 10.1145/1882471.1882479.

Frade, João Pedro, and Beth A. Arthington-Skaggs. 2011. "Effect of serum and surface characteristics on Candida albicans biofilm formation." *Mycoses* 54 (4):e154-62. doi: 10.1111/j.1439-0507.2010.01862.x.

Fuchs, Stefan, Cornelia Lass-Flörl, and Wilfried Posch. 2019. "Diagnostic Performance of a Novel Multiplex PCR Assay for Candidemia among ICU Patients." *J Fungi (Basel)* 5 (3). doi: 10.3390/jof5030086.

Fürnkranz, Johannes. 2010. "Decision Tree." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 263-267. Boston, MA: Springer US.

Futami, R., L. Muñoz-Pomer, J. M. Viu, and others. 2011. "GPRO: the professional tool for management, functional analysis and annotation of omic sequences and databases." *Biotechvana*.

Gabaldón, Toni, Miguel A. Naranjo-Ortíz, and Marina Marcet-Houben. 2016. "Evolutionary genomics of yeast pathogens in the Saccharomycotina." *FEMS Yeast Res.* 16 (6). doi: 10.1093/femsyr/fow064.

Gao, Song, Wing-Kin Sung, and Niranjan Nagarajan. 2011. "Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences." *J. Comput. Biol.* 18 (11):1681-1691. doi: 10.1089/cmb.2011.0170.

Garcia-Effron, Guillermo, Samuel Lee, Steven Park, John D. Cleary, and David S. Perlin. 2009. "Effect of Candida glabrata FKS1 and FKS2 mutations on echinocandin sensitivity and kinetics of 1,3-beta-D-glucan synthase: implication for the existing susceptibility breakpoint." *Antimicrob. Agents Chemother.* 53 (9):3690-3699. doi: 10.1128/AAC.00443-09.

Garcia-Effron, Guillermo, Steven Park, and David S. Perlin. 2009. "Correlating echinocandin MIC and kinetic inhibition of fks1 mutant glucan synthases for Candida albicans: implications for interpretive breakpoints." *Antimicrob. Agents Chemother.* 53 (1):112-122. doi: 10.1128/AAC.01162-08.

Garrison, Erik, and Gabor Marth. 2012. "Haplotype-based variant detection from short-read sequencing." *arXiv [q-bio.GN]*.

Geddes-McAlister, Jennifer, and Rebecca S. Shapiro. 2019. "New pathogens, new tricks: emerging, drug-resistant fungal pathogens and future prospects for antifungal therapeutics." *Ann. N. Y. Acad. Sci.* 1435 (1):57-78. doi: 10.1111/nyas.13739.

Gene Ontology Consortium. 2015. "Gene Ontology Consortium: going forward." *Nucleic Acids Res.* 43 (Database issue):D1049-56. doi: 10.1093/nar/gku1179.

Gillum, Amanda M., Emma Y. H. Tsay, and Donald R. Kirsch. 1984. "Isolation of the Candida albicans gene for orotidine-5′-phosphate decarboxylase by complementation of S. cerevisiae ura3 and E. coli pyrF mutations." *Molecular and General Genetics MGG* 198 (1):179-182. doi: 10.1007/bf00328721.

Gogoshin, Grigoriy, Eric Boerwinkle, and Andrei S. Rodin. 2017. "New Algorithm and Software (BNOmics) for Inferring and Visualizing Bayesian Networks from Heterogeneous Big Biological and Genetic Data." *J. Comput. Biol.* 24 (4):340-356. doi: 10.1089/cmb.2016.0100.

González-Torres, Pedro, Leszek P. Pryszcz, Fernando Santos, Manuel Martínez-García, Toni Gabaldón, and Josefa Antón. 2015. "Interactions between closely related bacterial strains are revealed by deep transcriptome sequencing." *Appl. Environ. Microbiol.* 81 (24):8445-8456. doi: 10.1128/AEM.02690-15.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of age: ten years of next-generation sequencing technologies." *Nat. Rev. Genet.* 17 (6):333-351. doi: 10.1038/nrg.2016.49.

Gordon, A., and G. Hannon. 2017. "Fastx-toolkit. FASTQ/A short-reads pre-processing tools. 2010." *Unpublished available online at: http://hannonlab.cshl.edu/fastx_toolkit*.

Gow, Neil A. R., and Bernhard Hube. 2012. "Importance of the Candida albicans cell wall during commensalism and infection." *Curr. Opin. Microbiol.* 15 (4):406-412. doi: 10.1016/j.mib.2012.04.005.

Grigoriev, Igor V., Roman Nikitin, Sajeet Haridas, Alan Kuo, Robin Ohm, Robert Otillar, Robert Riley, Asaf Salamov, Xueling Zhao, Frank Korzeniewski, Tatyana Smirnova, Henrik Nordberg, Inna Dubchak, and Igor Shabalov. 2014. "MycoCosm portal: gearing up for 1000 fungal genomes." *Nucleic Acids Res.* 42 (Database issue):D699-704. doi: 10.1093/nar/gkt1183.

Gu, Wei, Steve Miller, and Charles Y. Chiu. 2019. "Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection." *Annual Review of Pathology: Mechanisms of Disease* 14 (1):319-338. doi: 10.1146/annurev-pathmechdis-012418-012751.

Hagen, Ferry, Kantarawee Khayhan, Bart Theelen, Anna Kolecka, Itzhack Polacheck, Edward Sionov, Rama Falk, Sittiporn Parnmen, H. Thorsten Lumbsch, and Teun Boekhout. 2015. "Recognition of seven species in the Cryptococcus gattii/Cryptococcus neoformans species complex." *Fungal Genet. Biol.* 78:16-48. doi: 10.1016/j.fgb.2015.02.009.

Hartl, Barbara, Iris Zeller, Angelika Manhart, Brigitte Selitsch, Cornelia Lass-Flörl, and Birgit Willinger. 2018. "A Retrospective Assessment of Four Antigen Assays for the Detection of Invasive Candidiasis Among High-Risk Hospitalized Patients." *Mycopathologia* 183 (3):513-519. doi: 10.1007/s11046-017-0238-1.

Hendling, Michaela, Stephan Pabinger, Konrad Peters, Noa Wolff, Rick Conzemius, and Ivan Barišic. 2018. "Oli2go: an automated multiplex oligonucleotide design tool." *Nucleic Acids Res.* 46 (W1):W252-W256. doi: 10.1093/nar/gky319.

Hernández-Chávez, Marco J., Luis A. Pérez-García, Gustavo A. Niño-Vega, and Héctor M. Mora-Montes. 2017. "Fungal Strategies to Evade the Host Immune Recognition." *J Fungi (Basel)* 3 (4). doi: 10.3390/jof3040051.

Hirano, Ryuichi, Yuichi Sakamoto, Kumiko Kudo, and Motoki Ohnishi. 2015. "Retrospective analysis of mortality and Candida isolates of 75 patients with candidemia: a single hospital experience." *Infect. Drug Resist.* 8:199-205. doi: 10.2147/IDR.S80677.

Hoang, Minh Thuy Vi, Laszlo Irinyi, Sharon C. A. Chen, Tania C. Sorrell, Isham Barcoding of Medical Fungi Working Group, and Wieland Meyer. 2019. "Dual DNA Barcoding for the Molecular

Identification of the Agents of Invasive Fungal Infections." *Front. Microbiol.* 10:1647. doi: 10.3389/fmicb.2019.01647.

Hochreiter, S., and J. Schmidhuber. 1997. "Long short-term memory." *Neural Comput.* 9 (8):1735-1780. doi: 10.1162/neco.1997.9.8.1735.

Holzinger, Andreas. 2016. *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*: Springer.

Holzinger, Andreas, and Igor Jurisica. 2014. "Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions." *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*:1-18. doi: 10.1007/978-3-662-43968-5_1.

Hope, W. W., E. Castagnola, A. H. Groll, E. Roilides, M. Akova, M. C. Arendrup, S. Arikan-Akdagli, M. Bassetti, J. Bille, O. A. Cornely, M. Cuenca-Estrella, J. P. Donnelly, J. Garbino, R. Herbrecht, H. E. Jensen, B. J. Kullberg, C. Lass-Flörl, O. Lortholary, W. Meersseman, G. Petrikkos, M. D. Richardson, P. E. Verweij, C. Viscoli, A. J. Ullmann, and Escmid Fungal Infection Study Group. 2012. "ESCMID* guideline for the diagnosis and management of Candida diseases 2012: prevention and management of invasive infections in neonates and children caused by Candida spp." *Clin. Microbiol. Infect.* 18 Suppl 7:38-52. doi: 10.1111/1469-0691.12040.

Hou, Tsung-Yun, Chuan Chiang-Ni, and Shih-Hua Teng. 2019. "Current status of MALDI-TOF mass spectrometry in clinical microbiology." *J. Food Drug Anal.* 27 (2):404-414. doi: 10.1016/j.jfda.2019.01.001.

Hovhannisyan, Hrant, and Toni Gabaldón. 2019. "Transcriptome Sequencing Approaches to Elucidate Host-Microbe Interactions in Opportunistic Human Fungal Pathogens." *Curr. Top. Microbiol. Immunol.* 422:193-235. doi: 10.1007/82_2018_122.

Hovhannisyan, Hrant, and Toni Gabaldón. 2020. "The lncRNA landscape of Candida pathogens." *in preparation*.

Huseyin, Chloe E., Paul W. O'Toole, Paul D. Cotter, and Pauline D. Scanlan. 2017. "Forgotten fungi-the gut mycobiome in human health and disease." *FEMS Microbiol. Rev.* 41 (4):479-511. doi: 10.1093/femsre/fuw047.

Hynes, Seán O., Brendan Pang, Jacqueline A. James, Perry Maxwell, and Manuel Salto-Tellez. 2017. "Tissue-based next generation sequencing: application in a universal healthcare system." *British Journal of Cancer* 116 (5):553-560. doi: 10.1038/bjc.2016.452.

Irinyi, Laszlo, Yiheng Hu, Minh Thuy Vi Hoang, Lana Pasic, Catriona Halliday, Menuk Jayawardena, Indira Basu, Wendy McKinney, Arthur J. Morris, John Rathjen, Eric Stone, Sharon Chen, Tania C. Sorrell, Benjamin Schwessinger, and Wieland Meyer. 2020. "Long-read sequencing based clinical metagenomics for the detection and confirmation of Pneumocystis jirovecii directly from clinical specimens: A paradigm shift in mycological diagnostics." *Med. Mycol.* 58 (5):650-660. doi: 10.1093/mmy/myz109.

Irinyi, Laszlo, Michaela Lackner, G. Sybren de Hoog, and Wieland Meyer. 2016. "DNA barcoding of fungi causing infections in humans and animals." *Fungal Biol.* 120 (2):125-136. doi: 10.1016/j.funbio.2015.04.007.

Irinyi, Laszlo, Carolina Serena, Dea Garcia-Hermoso, Michael Arabatzis, Marie Desnos-Ollivier, Duong Vu, Gianluigi Cardinali, Ian Arthur, Anne-Cécile Normand, Alejandra Giraldo, Keith Cassia da Cunha, Marcelo Sandoval-Denis, Marijke Hendrickx, Angela Satie Nishikaku, Analy Salles de Azevedo Melo, Karina Bellinghausen Merseguel, Aziza Khan, Juliana Alves Parente Rocha, Paula Sampaio, Marcelo Ribeiro da Silva Briones, Renata Carmona e Ferreira, Mauro de Medeiros Muniz, Laura Rosio Castañón-Olivares, Daniel Estrada-Barcenas, Carole Cassagne, Charles Mary, Shu Yao Duan, Fanrong Kong, Annie Ying Sun, Xianyu Zeng, Zuotao Zhao, Nausicaa Gantois, Françoise Botterel, Barbara Robbertse, Conrad Schoch, Walter Gams, David Ellis, Catriona Halliday, Sharon Chen, Tania C. Sorrell, Renaud Piarroux, Arnaldo L. Colombo, Célia Pais, Sybren de Hoog, Rosely Maria Zancopé-Oliveira, Maria Lucia Taylor, Conchita Toriello, Célia Maria de Almeida Soares, Laurence Delhaes, Dirk Stubbe, Françoise Dromer, Stéphane Ranque, Josep Guarro, Jose F. Cano-Lira, Vincent Robert, Aristea Velegraki, and Wieland Meyer. 2015. "International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database--the quality controlled standard tool for routine identification of human and animal pathogenic fungi." *Med. Mycol.* 53 (4):313-337. doi: 10.1093/mmy/myv008.

Ji, Boyang, and Jens Nielsen. 2015. "From next-generation sequencing to systematic modeling of the gut microbiome." *Front. Genet.* 6:219. doi: 10.3389/fgene.2015.00219.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. 2014. "InterProScan 5: genome-scale protein function classification." *Bioinformatics* 30 (9):1236-1240. doi: 10.1093/bioinformatics/btu031.

Kalderimis, Alex, Rachel Lyne, Daniela Butano, Sergio Contrino, Mike Lyne, Joshua Heimbach, Fengyuan Hu, Richard Smith, Radek Štěpán, Julie Sullivan, and Gos Micklem. 2014. "InterMine: extensive web services for modern biology." *Nucleic Acids Research* 42 (W1):W468-W472. doi: 10.1093/nar/gku301.

Kalderimis, Alexis, Radek Stepan, Julie Sullivan, Rachel Lyne, Michael Lyne, and Gos Micklem. 2014. "BioJS InterMine List Analysis: A BioJS component for displaying graphical or statistical analysis of collections of items from InterMine endpoints." *F1000Research* 3:45. doi: 10.12688/f1000research.3-45.v1.

Kanehisa, M. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1):27-30. doi: 10.1093/nar/28.1.27.

Kapteyn, J. C., L. L. Hoyer, J. E. Hecht, W. H. Müller, A. Andel, A. J. Verkleij, M. Makarow, H. Van Den Ende, and F. M. Klis. 2000. "The cell wall architecture of Candida albicans wild-type cells and cell wall-defective mutants." *Mol. Microbiol.* 35 (3):601-611. doi: 10.1046/j.1365-2958.2000.01729.x.

Kelly, Judy, and Kevin Kavanagh. 2010. "Proteomic analysis of proteins released from growth-arrested Candida albicans following exposure to caspofungin." *Med. Mycol.* 48 (4):598-605. doi: 10.3109/13693780903405782.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. "The human genome browser at UCSC." *Genome Res* 12 (6):996-1006. doi: 10.1101/gr.229102.

Kersey, Paul Julian, James E. Allen, Irina Armean, Sanjay Boddu, Bruce J. Bolt, Denise Carvalho-Silva, Mikkel Christensen, Paul Davis, Lee J. Falin, Christoph Grabmueller, Jay Humphrey, Arnaud

Kerhornou, Julia Khobova, Naveen K. Aranganathan, Nicholas Langridge, Ernesto Lowy, Mark D. McDowall, Uma Maheswari, Michael Nuhn, Chuang Kee Ong, Bert Overduin, Michael Paulini, Helder Pedro, Emily Perry, Giulietta Spudich, Electra Tapanari, Brandon Walts, Gareth Williams, Marcela Tello-Ruiz, Joshua Stein, Sharon Wei, Doreen Ware, Daniel M. Bolser, Kevin L. Howe, Eugene Kulesha, Daniel Lawson, Gareth Maslen, and Daniel M. Staines. 2016. "Ensembl Genomes 2016: more genomes, more complexity." *Nucleic Acids Res.* 44 (D1):D574-80. doi: 10.1093/nar/gkv1209.

Kertesz, Michael, Yue Wan, Elad Mazor, John L. Rinn, Robert C. Nutter, Howard Y. Chang, and Eran Segal. 2010. "Genome-wide measurement of RNA secondary structure in yeast." *Nature* 467 (7311):103-107. doi: 10.1038/nature09322.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: a fast spliced aligner with low memory requirements." *Nat. Methods* 12 (4):357-360. doi: 10.1038/nmeth.3317.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." *Genome Biol.* 14 (4):R36. doi: 10.1186/gb-2013-14-4-r36.

Kingma, Diederik P., and Jimmy Ba. 2015. "Adam: A Method for Stochastic Optimization." 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing." *Genome Res.* 22 (3):568-576. doi: 10.1101/gr.129684.111.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation." *Genome Res.* 27 (5):722-736. doi: 10.1101/gr.215087.116.

Koumakis, Lefteris. 2020. "Deep learning models in genomics; are we there yet?" *Comput. Struct. Biotechnol. J.* 18:1466-1473. doi: 10.1016/j.csbj.2020.06.017.

Kristensen, Vessela N., Ole Christian Lingjærde, Hege G. Russnes, Hans Kristian M. Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. 2014. "Principles and methods of integrative genomic analyses in cancer." *Nat. Rev. Cancer* 14 (5):299-313. doi: 10.1038/nrc3721.

Ksiezopolska, Ewa, and Toni Gabaldón. 2018. "Evolutionary Emergence of Drug Resistance in Candida Opportunistic Pathogens." *Genes* 9 (9). doi: 10.3390/genes9090461.

Kullberg, Bart Jan, and Maiken C. Arendrup. 2016. "Invasive Candidiasis." *N. Engl. J. Med.* 374 (8):794-795. doi: 10.1056/NEJMc1514201.

Kumamoto, Carol A. 2011. "Inflammation and gastrointestinal Candida colonization." *Current Opinion in Microbiology* 14 (4):386-391. doi: 10.1016/j.mib.2011.07.015.

Kumar, Ranjit, and Bindu Nanduri. 2010. "HPIDB--a unified resource for host-pathogen interactions." *BMC Bioinformatics* 11 Suppl 6:S16. doi: 10.1186/1471-2105-11-S6-S16.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast gapped-read alignment with Bowtie 2." *Nat. Methods* 9 (4):357-359. doi: 10.1038/nmeth.1923.

Larrañaga, Pedro, Hossein Karshenas, Concha Bielza, and Roberto Santana. 2013. "A review on evolutionary algorithms in Bayesian network learning and inference tasks." *Information Sciences* 233:109-125. doi: 10.1016/j.ins.2012.12.051.

Larsen, Peter E., Avinash Sreedasyam, Geetika Trivedi, Shalaka Desai, Yang Dai, Leland J. Cseke, and Frank R. Collart. 2015. "Multi-Omics Approach Identifies Molecular Mechanisms of Plant-Fungus Mycorrhizal Interaction." *Front. Plant Sci.* 6:1061. doi: 10.3389/fpls.2015.01061.

Lawrence, Travis J., Kyle T. Kauffman, Katherine C. H. Amrine, Dana L. Carper, Raymond S. Lee, Peter J. Becich, Claudia J. Canales, and David H. Ardell. 2015. "FAST: FAST Analysis of Sequences Toolbox." *Front. Genet.* 6:172. doi: 10.3389/fgene.2015.00172.

Leamy, Kathleen A., Sarah M. Assmann, David H. Mathews, and Philip C. Bevilacqua. 2016. "Bridging the gap between in vitro and in vivo RNA folding." *Q. Rev. Biophys.* 49:e10. doi: 10.1017/S003358351600007X.

Lecuit, Marc, and Marc Eloit. 2014. "The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening." *Frontiers in Cellular and Infection Microbiology* 4. doi: 10.3389/fcimb.2014.00025.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521 (7553):436-444. doi: 10.1038/nature14539.

Lee, Keunsook K., Donna M. Maccallum, Mette D. Jacobsen, Louise A. Walker, Frank C. Odds, Neil A. R. Gow, and Carol A. Munro. 2012. "Elevated cell wall chitin in Candida albicans confers echinocandin resistance in vivo." *Antimicrob. Agents Chemother.* 56 (1):208-217. doi: 10.1128/AAC.00683-11.

Lefterova, Martina I., Carlos J. Suarez, Niaz Banaei, and Benjamin A. Pinsky. 2015. "Next-Generation Sequencing for Infectious Disease Diagnosis and Management: A Report of the Association for Molecular Pathology." *J. Mol. Diagn.* 17 (6):623-634. doi: 10.1016/j.jmoldx.2015.07.004.

Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and Collaboration International Nucleotide Sequence Database. 2011. "The sequence read archive." *Nucleic Acids Res.* 39 (Database issue):D19-21. doi: 10.1093/nar/gkq1019.

Li, Bo, and Colin N. Dewey. 2011. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." *BMC Bioinformatics* 12 (1). doi: 10.1186/1471-2105-12-323.

Li, Heng, and Richard Durbin. 2009. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25 (14):1754-1760. doi: 10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and Subgroup Genome Project Data Processing. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16):2078-2079. doi: 10.1093/bioinformatics/btp352.

Li, Heng, and Nils Homer. 2010. "A survey of sequence alignment algorithms for next-generation sequencing." *Brief. Bioinform.* 11 (5):473-483. doi: 10.1093/bib/bbq015.

Li, Jing-Woei, Robert Schmieder, R. Matthew Ward, Joann Delenick, Eric C. Olivares, and David Mittelman. 2012. "SEQanswers: an open access community for collaboratively decoding genomes." *Bioinformatics* 28 (9):1272-1273. doi: 10.1093/bioinformatics/bts128.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol.* 15 (12):550. doi: 10.1186/s13059-014-0550-8.

Lu, Jennifer, Andrew Johnston, Philippe Berichon, Ke-Lin Ru, Darren Korbie, and Matt Trau. 2017. "PrimerSuite: A High-Throughput Web-Based Primer Design Program for Multiplex Bisulfite PCR." *Sci. Rep.* 7:41328. doi: 10.1038/srep41328.

Lu, Zhipeng, Jing Gong, and Qiangfeng Cliff Zhang. 2018. "PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution." *Methods Mol. Biol.* 1649:59-84. doi: 10.1007/978-1-4939-7213-5_4.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W. Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. 2012. "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." *Gigascience* 1 (1):18. doi: 10.1186/2047-217X-1-18.

Luthra, Rajyalakshmi, Hui Chen, Sinchita Roy-Chowdhuri, and R. Rajesh Singh. 2015. "Next-Generation Sequencing in Clinical Molecular Diagnostics of Cancer: Advantages and Challenges." *Cancers* 7 (4):2023-2036. doi: 10.3390/cancers7040874.

Lyne, Mike, Richard N. Smith, Rachel Lyne, Jelena Aleksic, Fengyuan Hu, Alex Kalderimis, Radek Stepan, and Gos Micklem. 2013. "metabolicMine: an integrated genomics, genetics and proteomics data warehouse for common metabolic disease research." *Database* 2013:bat060. doi: 10.1093/database/bat060.

Maguire, Sarah L., Seán S. ÓhÉigeartaigh, Kevin P. Byrne, Markus S. Schröder, Peadar O'Gaora, Kenneth H. Wolfe, and Geraldine Butler. 2013. "Comparative genome analysis and gene finding in Candida species using CGOB." *Mol. Biol. Evol.* 30 (6):1281-1291. doi: 10.1093/molbev/mst042.

Mar Rodríguez, M., Daniel Pérez, Felipe Javier Chaves, Eduardo Esteve, Pablo Marin-Garcia, Gemma Xifra, Joan Vendrell, Mariona Jové, Reinald Pamplona, Wifredo Ricart, Manuel Portero-Otin, Matilde R. Chacón, and José Manuel Fernández Real. 2016. "Erratum: Obesity changes the human gut mycobiome." *Sci. Rep.* 6:21679. doi: 10.1038/srep21679.

Martín, Abadi, Agarwal Ashish, Barham Paul, Brevdo Eugene, Chen Zhifeng, Citro Craig, S. Corrado Greg, Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Goodfellow Ian, Harp Andrew, Irving Geoffrey, Isard Michael, Yangqing Jia, Jozefowicz Rafal, Kaiser Lukasz, Kudlur Manjunath, Levenberg Josh, Mané Dandelion, Monga Rajat, Moore Sherry, Murray Derek, Olah Chris, Schuster Mike, Shlens Jonathon, Steiner Benoit, Sutskever Ilya, Talwar Kunal, Tucker Paul, Vanhoucke Vincent, Vasudevan Vijay, Viégas Fernanda, Vinyals Oriol, Warden Pete, Wattenberg Martin, Wicke Martin, Yu Yuan, and Zheng Xiaoqiang. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems."

Martin, Marcel. 2011. "Cutadapt removes adapter sequences from high-throughput sequencing reads." *EMBnet.journal* 17 (1):10. doi: 10.14806/ej.17.1.200.

Martin, Ronny, Daniela Albrecht-Eckardt, Sascha Brunke, Bernhard Hube, Kerstin Hünniger, and Oliver Kurzai. 2013. "A core filamentation response network in Candida albicans is restricted to eight genes." *PLoS One* 8 (3):e58613. doi: 10.1371/journal.pone.0058613.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. 2010. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome Res.* 20 (9):1297-1303. doi: 10.1101/gr.107524.110.

McLachlan, Geoffrey J. 1992. "Discriminant Analysis and Statistical Pattern Recognition." *Wiley Series in Probability and Statistics*. doi: 10.1002/0471725293.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biol.* 17 (1):122. doi: 10.1186/s13059-016-0974-4.

McTaggart, Lisa R., Julia K. Copeland, Anuradha Surendra, Pauline W. Wang, Shahid Husain, Bryan Coburn, David S. Guttman, and Julianne V. Kus. 2019. "Mycobiome Sequencing and Analysis Applied to Fungal Community Profiling of the Lower Respiratory Tract During Fungal Pathogenesis." *Front. Microbiol.* 10:512. doi: 10.3389/fmicb.2019.00512.

Megri, Youcef, Amir Arastehfar, Teun Boekhout, Farnaz Daneshnia, Caroline Hörtnagl, Bettina Sartori, Ahmed Hafez, Weihua Pan, Cornelia Lass-Flörl, and Boussad Hamrioui. 2020. "Candida tropicalis is the most prevalent yeast species causing candidemia in Algeria: the urgent need for antifungal stewardship and infection control measures." *Antimicrob. Resist. Infect. Control* 9 (1):50. doi: 10.1186/s13756-020-00710-z.

Mellmann, A., F. Bimet, C. Bizet, A. D. Borovskaya, R. R. Drake, U. Eigner, A. M. Fahr, Y. He, E. N. Ilina, M. Kostrzewa, T. Maier, L. Mancinelli, W. Moussaoui, G. Prévost, L. Putignani, C. L. Seachord, Y. W. Tang, and D. Harmsen. 2009. "High interlaboratory reproducibility of matrix-assisted laser desorption ionization-time of flight mass spectrometry-based species identification of nonfermenting bacteria." *J. Clin. Microbiol.* 47 (11):3732-3734. doi: 10.1128/JCM.00921-09.

Merkel, Dirk. 2014. "Docker: lightweight linux containers for consistent development and deployment." *Linux J.* 2014 (239):2.

Metz, Charles E. 1978. "Basic principles of ROC analysis." *Seminars in Nuclear Medicine* 8 (4):283-298. doi: 10.1016/s0001-2998(78)80014-2.

Metzger, Brian P. H., Patricia J. Wittkopp, and Joseph D. Coolon. 2017. "Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species." *Genome Biol. Evol.* 9 (4):843-854. doi: 10.1093/gbe/evx035.

Meyer, Wieland, Laszlo Irinyi, Minh Thuy Vi Hoang, Vincent Robert, Dea Garcia-Hermoso, Marie Desnos-Ollivier, Chompoonek Yurayart, Chi-Ching Tsang, Chun-Yi Lee, Patrick C. Y. Woo, Ivan Mikhailovich Pchelin, Silke Uhrlaß, Pietro Nenoff, Ariya Chindamporn, Sharon Chen, Paul D. N. Hebert, Tania C. Sorrell, and Isham barcoding of pathogenic fungi working group. 2019. "Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α)." *Genome* 62 (3):160-169. doi: 10.1139/gen-2018-0083.

Miller, Jason R., Sergey Koren, and Granger Sutton. 2010. "Assembly algorithms for next-generation sequencing data." *Genomics* 95 (6):315-327. doi: 10.1016/j.ygeno.2010.03.001.

Mistry, Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. 2013. "Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions." *Nucleic Acids Res.* 41 (12):e121. doi: 10.1093/nar/gkt263.

Mixão, Verónica, and Toni Gabaldón. 2018. "Hybridization and emergence of virulence in opportunistic human yeast pathogens." *Yeast* 35 (1):5-20. doi: 10.1002/yea.3242.

Morrison, C. J., S. F. Hurst, S. L. Bragg, R. J. Kuykendall, H. Diaz, J. Pohl, and E. Reiss. 1993. "Heterogeneity of the purified extracellular aspartyl proteinase from Candida albicans: characterization with monoclonal antibodies and N-terminal amino acid sequence analysis." *Infect. Immun.* 61 (5):2030-2036. doi: 10.1128/IAI.61.5.2030-2036.1993.

Mortimer, Stefanie A., Mary Anne Kidwell, and Jennifer A. Doudna. 2014. "Insights into RNA structure and function from genome-wide studies." *Nat. Rev. Genet.* 15 (7):469-479. doi: 10.1038/nrg3681.

Mullis, K., F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. 1986. "Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction." *Cold Spring Harbor Symposia on Quantitative Biology* 51 (0):263-273. doi: 10.1101/sqb.1986.051.01.032.

Mundodi, V., A. S. Kucknoor, and J. F. Alderete. 2008. "Immunogenic and plasminogen-binding surface-associated alpha-enolase of Trichomonas vaginalis." *Infect. Immun.* 76 (2):523-531. doi: 10.1128/IAI.01352-07.

Munoz-Pomer, A., R. Futami, L. Covelli, L. Dominguez-Escriba, G. P. Bernet, J. M. Sempere, A. Moya, and C. Llorens. 2011. "TIME: a sequence editor for the molecular analysis of large DNA and protein sequence samples." *Biotechvana Bioinformatics* 437.

Nather, Kerstin, and Carol A. Munro. 2008. "Generating cell surface diversity in Candida albicans and other fungal pathogens." *FEMS Microbiol. Lett.* 285 (2):137-145. doi: 10.1111/j.1574-6968.2008.01263.x.

Nguyen, Linh D. N., Eric Viscogliosi, and Laurence Delhaes. 2015. "The lung mycobiome: an emerging field of the human respiratory microbiome." *Front. Microbiol.* 6:89. doi: 10.3389/fmicb.2015.00089.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." *N. Engl. J. Med.* 375 (13):1216-1219. doi: 10.1056/NEJMp1606181.

Oughtred, Rose, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-Aryamontri, Kara Dolinski, and Mike Tyers. 2019. "The BioGRID interaction database: 2019 update." *Nucleic Acids Res.* 47 (D1):D529-D541. doi: 10.1093/nar/gky1079.

Ouyang, Zhengqing, Michael P. Snyder, and Howard Y. Chang. 2013. "SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data." *Genome Res.* 23 (2):377-387. doi: 10.1101/gr.138545.112.

Papon, Nicolas, Vincent Courdavault, Marc Clastre, and Richard J. Bennett. 2013. "Emerging and emerged pathogenic Candida species: beyond the Candida albicans paradigm." *PLoS Pathog.* 9 (9):e1003550. doi: 10.1371/journal.ppat.1003550.

Pardini, Giacomo, Piet W. J. De Groot, Alix T. Coste, Mahir Karababa, Frans M. Klis, Chris G. de Koster, and Dominique Sanglard. 2006. "The CRH family coding for cell wall glycosylphosphatidylinositol proteins with a predicted transglycosidase domain affects cell wall organization and virulence of Candida albicans." *J. Biol. Chem.* 281 (52):40399-40411. doi: 10.1074/jbc.M606361200.

Park, S., R. Kelly, J. Nielsen Kahn, J. Robles, M. J. Hsu, E. Register, W. Li, V. Vyas, H. Fan, G. Abruzzo, A. Flattery, C. Gill, G. Chrebet, S. A. Parent, M. Kurtz, H. Teppler, C. M. Douglas, and D. S. Perlin. 2005. "Specific substitutions in the echinocandin target Fks1p account for reduced susceptibility of rare laboratory and clinical Candida sp. isolates." *Antimicrob. Agents Chemother.* 49 (8):3264-3273. doi: 10.1128/AAC.49.8.3264-3273.2005.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon provides fast and bias-aware quantification of transcript expression." *Nat. Methods* 14 (4):417-419. doi: 10.1038/nmeth.4197.

Pedersen, Brent S., and Aaron R. Quinlan. 2018. "Mosdepth: quick coverage calculation for genomes and exomes." *Bioinformatics* 34 (5):867-868. doi: 10.1093/bioinformatics/btx699.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *J. Mach. Learn. Res.* 12 (85):2825-2830.

Pérez-Sánchez, Jaume, Fernando Naya-Català, Beatriz Soriano, M. Carla Piazzon, Ahmed Hafez, Toni Gabaldón, Carlos Llorens, Ariadna Sitjà-Bobadilla, and Josep A. Calduch-Giner. 2019. "Genome Sequencing and Transcriptome Analysis Reveal Recent Species-Specific Gene Duplications in the Plastic Gilthead Sea Bream (Sparus aurata)." *Frontiers in Marine Science* 6. doi: 10.3389/fmars.2019.00760.

Perlin, David S. 2007. "Resistance to echinocandin-class antifungal drugs." *Drug Resist. Updat.* 10 (3):121-130. doi: 10.1016/j.drup.2007.04.002.

Petersen, Thomas Nordahl, Oksana Lukjancenko, Martin Christen Frølund Thomsen, Maria Maddalena Sperotto, Ole Lund, Frank Møller Aarestrup, and Thomas Sicheritz-Pontén. 2017. "MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads." *PLoS One* 12 (5):e0176469. doi: 10.1371/journal.pone.0176469.

Plaine, Armêl, Louise Walker, Gregory Da Costa, Héctor M. Mora-Montes, Alastair McKinnon, Neil A. R. Gow, Claude Gaillardin, Carol A.

Munro, and Mathias L. Richard. 2008. "Functional analysis of Candida albicans GPI-anchored proteins: roles in cell wall integrity and caspofungin sensitivity." *Fungal Genet. Biol.* 45 (10):1404-1414. doi: 10.1016/j.fgb.2008.08.003.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. 2018. "Scaling accurate genetic variant discovery to tens of thousands of samples." *bioRxiv*:201178. doi: 10.1101/201178.

Prella, Maura, Jacques Bille, Mauro Pugnale, Bertrand Duvoisin, Matthias Cavassini, Thierry Calandra, and Oscar Marchetti. 2005. "Early diagnosis of invasive candidiasis with mannan antigenemia and antimannan antibodies." *Diagn. Microbiol. Infect. Dis.* 51 (2):95-101. doi: 10.1016/j.diagmicrobio.2004.08.015.

Pryszcz, Leszek P., Tibor Németh, Attila Gácser, and Toni Gabaldón. 2014. "Genome Comparison of Candida orthopsilosis Clinical Strains Reveals the Existence of Hybrids between Two Distinct Subspecies." *Genome Biology and Evolution* 6 (5):1069-1078. doi: 10.1093/gbe/evu082.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, H. I. T. Consortium Meta, Peer Bork, S. Dusko Ehrlich, and Jun Wang. 2010. "A human gut microbial gene catalogue established by metagenomic sequencing." *Nature* 464 (7285):59-65. doi: 10.1038/nature08821.

Quang, Daniel, and Xiaohui Xie. 2016. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic Acids Res.* 44 (11):e107. doi: 10.1093/nar/gkw226.

Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. "Corrigendum: Shotgun metagenomics, from sampling to analysis." *Nat. Biotechnol.* 35 (12):1211. doi: 10.1038/nbt1217-1211b.

Rajendran, R., L. Sherry, C. J. Nile, A. Sherriff, E. M. Johnson, M. F. Hanson, C. Williams, C. A. Munro, B. J. Jones, and G. Ramage. 2016. "Biofilm formation is a risk factor for mortality in patients with Candida albicans bloodstream infection-Scotland, 2012-2013." *Clin. Microbiol. Infect.* 22 (1):87-93. doi: 10.1016/j.cmi.2015.09.018.

Ramage, G., K. Vande Walle, B. L. Wickes, and J. L. López-Ribot. 2001. "Standardized method for in vitro antifungal susceptibility testing of Candida albicans biofilms." *Antimicrob. Agents Chemother.* 45 (9):2475-2479. doi: 10.1128/aac.45.9.2475-2479.2001.

Rehm, Heidi L., Genetics for the Working Group of the American College of Medical, Committee Genomics Laboratory Quality Assurance, Sherri J. Bale, Pinar Bayrak-Toydemir, Jonathan S. Berg, Kerry K. Brown, Joshua L. Deignan, Michael J. Friez, Birgit H. Funke, Madhuri R. Hegde, and Elaine Lyon. 2013. "ACMG clinical laboratory standards for next-generation sequencing." *Genetics in Medicine* 15 (9):733-747. doi: 10.1038/gim.2013.92.

Remmele, Christian W., Christian H. Luther, Johannes Balkenhol, Thomas Dandekar, Tobias Müller, and Marcus T. Dittrich. 2015. "Integrated inference and evaluation of host-fungi interaction networks." *Front. Microbiol.* 6:764. doi: 10.3389/fmicb.2015.00764.

Ren, Ye, Le Zhang, and P. N. Suganthan. 2016. "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]." *IEEE Computational Intelligence Magazine* 11 (1):41-53. doi: 10.1109/mci.2015.2471235.

Rhodes, Johanna, and Matthew C. Fisher. 2019. "Global epidemiology of emerging Candida auris." *Curr. Opin. Microbiol.* 52:84-89. doi: 10.1016/j.mib.2019.05.008.

Richardson, Malcolm D., and David W. Warnock. 2012. *Fungal Infection: Diagnosis and Management*: John Wiley & Sons.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Res.* 43 (7):e47. doi: 10.1093/nar/gkv007.

Robert, Vincent, Duong Vu, Ammar Ben Hadj Amor, Nathalie van de Wiele, Carlo Brouwer, Bernard Jabas, Szaniszlo Szoke, Ahmed Dridi, Maher Triki, Samy Ben Daoud, Oussema Chouchen, Lea Vaas, Arthur de Cock, Joost A. Stalpers, Dora Stalpers, Gerard J. M. Verkley, Marizeth Groenewald, Felipe Borges Dos Santos, Gerrit Stegehuis, Wei Li, Linhuan Wu, Run Zhang, Juncai Ma, Miaomiao Zhou, Sergio Pérez Gorjón, Lily Eurwilaichitr, Supawadee Ingsriswang, Karen Hansen, Conrad Schoch, Barbara Robbertse, Laszlo Irinyi, Wieland Meyer, Gianluigi Cardinali, David L. Hawksworth, John W. Taylor, and Pedro W. Crous. 2013. "MycoBank gearing up for new horizons." *IMA Fungus* 4 (2):371-379. doi: 10.5598/imafungus.2013.04.02.16.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26 (1):139-140. doi: 10.1093/bioinformatics/btp616.

Rodríguez, Pau, Miguel A. Bautista, Jordi Gonzàlez, and Sergio Escalera. 2018. "Beyond one-hot encoding: Lower dimensional target embedding." *Image and Vision Computing* 75:21-31. doi: 10.1016/j.imavis.2018.04.004.

Rotroff, Daniel M., and Alison A. Motsinger-Reif. 2016. "Embracing Integrative Multiomics Approaches." *Int. J. Genomics Proteomics* 2016:1715985. doi: 10.1155/2016/1715985.

Rybowicz, Joseph, and Cheryle Gurk-Turner. 2002. "Caspofungin: the first agent available in the echinocandin class of antifungals." *Proc.* 15 (1):97-99. doi: 10.1080/08998280.2002.11927822.

Sahlin, Kristoffer, Francesco Vezzi, Björn Nystedt, Joakim Lundeberg, and Lars Arvestad. 2014. "BESST--efficient scaffolding of large fragmented assemblies." *BMC Bioinformatics* 15:281. doi: 10.1186/1471-2105-15-281.

Sammut, Claude, and Geoffrey I. Webb. 2010a. "Adaboost." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 19-19. Boston, MA: Springer US.

Sammut, Claude, and Geoffrey I. Webb. 2010b. "Nearest Neighbor Methods." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 715-715. Boston, MA: Springer US.

Sammut, Claude, and Geoffrey I. Webb. 2010c. "Random Forests." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 828-828. Boston, MA: Springer US.

SantaLucia, J., Jr. 1998. "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." *Proc. Natl. Acad. Sci. U. S. A.* 95 (4):1460-1465. doi: 10.1073/pnas.95.4.1460.

Sardi, J. C. O., L. Scorzoni, T. Bernardi, A. M. Fusco-Almeida, and M. J. S. Mendes Giannini. 2013. "Candida species: current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options." *J. Med. Microbiol.* 62 (Pt 1):10-24. doi: 10.1099/jmm.0.045054-0.

Saus, Ester, Jesse R. Willis, Leszek P. Pryszcz, Ahmed Hafez, Carlos Llorens, Heinz Himmelbauer, and Toni Gabaldón. 2018. "nextPARS: parallel probing of RNA structures in Illumina." *RNA* 24 (4):609-619. doi: 10.1261/rna.063073.117.

Sayers, Eric W., Jeff Beck, J. Rodney Brister, Evan E. Bolton, Kathi Canese, Donald C. Comeau, Kathryn Funk, Anne Ketter, Sunghwan Kim, Avi Kimchi, Paul A. Kitts, Anatoliy Kuznetsov, Stacy Lathrop, Zhiyong Lu, Kelly McGarvey, Thomas L. Madden, Terence D. Murphy, Nuala O'Leary, Lon Phan, Valerie A. Schneider, Françoise Thibaud-Nissen, Bart W. Trawick, Kim D. Pruitt, and James Ostell. 2020. "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Res.* 48 (D1):D9-D16. doi: 10.1093/nar/gkz899.

Schmalreck, A. F., M. Lackner, K. Becker, W. Fegeler, V. Czaika, H. Ulmer, and C. Lass-Flörl. 2014. "Phylogenetic relationships matter: antifungal susceptibility among clinically relevant yeasts." *Antimicrob. Agents Chemother.* 58 (3):1575-1585. doi: 10.1128/AAC.01799-13.

Schmieder, Robert, and Robert Edwards. 2011. "Quality control and preprocessing of metagenomic datasets." *Bioinformatics* 27 (6):863-864. doi: 10.1093/bioinformatics/btr026.

Schneider, Michel, Lydie Lane, Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Lydie Bougueleret, and Amos Bairoch. 2009. "The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program." *J. Proteomics* 72 (3):567-573. doi: 10.1016/j.jprot.2008.11.010.

Schoch, Conrad L., Keith A. Seifert, Sabine Huhndorf, Vincent Robert, John L. Spouge, C. André Levesque, Wen Chen, Consortium Fungal Barcoding, and List Fungal Barcoding Consortium Author. 2012. "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi." *Proc. Natl. Acad. Sci. U. S. A.* 109 (16):6241-6246. doi: 10.1073/pnas.1117018109.

Schröder, Markus S., Kontxi Martinez de San Vicente, Tâmara H. R. Prandini, Stephen Hammel, Desmond G. Higgins, Eduardo Bagagli, Kenneth H. Wolfe, and Geraldine Butler. 2016. "Multiple Origins of the Pathogenic Yeast Candida orthopsilosis by Separate Hybridizations between Two Parental Species." *PLoS Genet.* 12 (11):e1006404. doi: 10.1371/journal.pgen.1006404.

Schulz, Marcel H., Daniel R. Zerbino, Martin Vingron, and Ewan Birney. 2012. "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels." *Bioinformatics* 28 (8):1086-1092. doi: 10.1093/bioinformatics/bts094.

Sekyere, John Osei, and Jonathan Asante. 2018. "Emerging mechanisms of antimicrobial resistance in bacteria and fungi: advances in the era of genomics." *Future Microbiol.* 13:241-262. doi: 10.2217/fmb-2017-0172.

Sellam, Adnane, Hervé Hogues, Christopher Askew, Faiza Tebbji, Marco van Het Hoog, Hugo Lavoie, Carol A. Kumamoto, Malcolm Whiteway, and André Nantel. 2010. "Experimental annotation of the human pathogen Candida albicans coding and noncoding transcribed regions using high-resolution tiling arrays." *Genome Biol.* 11 (7):R71. doi: 10.1186/gb-2010-11-7-r71.

Shen, Zhen, Wenzheng Bao, and De-Shuang Huang. 2018. "Recurrent Neural Network for Predicting Transcription Factor Binding Sites." *Sci. Rep.* 8 (1):15270. doi: 10.1038/s41598-018-33321-1.

Sherstinsky, Alex. 2020. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network." *Physica D: Nonlinear Phenomena* 404:132306. doi: 10.1016/j.physd.2019.132306.

Skrzypek, Marek S., Jonathan Binkley, Gail Binkley, Stuart R. Miyasato, Matt Simison, and Gavin Sherlock. 2017. "The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data." *Nucleic Acids Res.* 45 (D1):D592-D596. doi: 10.1093/nar/gkw924.

Smith, David Roy. 2015. "Buying in to bioinformatics: an introduction to commercial sequence analysis software." *Brief. Bioinform.* 16 (4):700-709. doi: 10.1093/bib/bbu030.

Smith, Richard N., Jelena Aleksic, Daniela Butano, Adrian Carr, Sergio Contrino, Fengyuan Hu, Mike Lyne, Rachel Lyne, Alex Kalderimis, Kim Rutherford, Radek Stepan, Julie Sullivan, Matthew Wakeling, Xavier Watkins, and Gos Micklem. 2012. "InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data." *Bioinformatics* 28 (23):3163-3165. doi: 10.1093/bioinformatics/bts577.

Stanke, Mario, Mark Diekhans, Robert Baertsch, and David Haussler. 2008. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." *Bioinformatics* 24 (5):637-644. doi: 10.1093/bioinformatics/btn013.

Stavrou, Aimilia A., Verónica Mixão, Teun Boekhout, and Toni Gabaldón. 2018. "Misidentification of genome assemblies in public databases: The case of Naumovozyma dairenensis and proposal of a protocol to correct misidentifications." *Yeast* 35 (6):425-429. doi: 10.1002/yea.3303.

Stielow, J. B., C. A. Lévesque, K. A. Seifert, W. Meyer, L. Iriny, D. Smits, R. Renfurm, G. J. M. Verkley, M. Groenewald, D. Chaduli, A. Lomascolo, S. Welti, L. Lesage-Meessen, A. Favel, A. M. S. Al-Hatmi, U. Damm, N. Yilmaz, J. Houbraken, L. Lombard, W. Quaedvlieg, M. Binder, L. A. I. Vaas, D. Vu, A. Yurkov, D. Begerow, O. Roehl, M. Guerreiro, A. Fonseca, K. Samerpitak, A. D. van Diepeningen, S. Dolatabadi, L. F. Moreno, S. Casaregola, S. Mallet, N. Jacques, L. Roscini, E. Egidi, C. Bizet, D. Garcia-Hermoso, M. P. Martín, S. Deng, J. Z. Groenewald, T. Boekhout, Z. W. de Beer, I. Barnes, T. A. Duong, M. J. Wingfield, G. S. de Hoog, P. W. Crous, C. T. Lewis, S. Hambleton, T. A. A. Moussa,

H. S. Al-Zahrani, O. A. Almaghrabi, G. Louis-Seize, R. Assabgui, W. McCormick, G. Omer, K. Dukik, G. Cardinali, U. Eberhardt, M. de Vries, and V. Robert. 2015. "One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes." *Persoonia* 35:242-263. doi: 10.3767/003158515X689135.

Strobel, Eric J., Angela M. Yu, and Julius B. Lucks. 2018. "High-throughput determination of RNA structures." *Nat. Rev. Genet.* 19 (10):615-634. doi: 10.1038/s41576-018-0034-x.

Strommenger, Birgit, Christiane Kettlitz, Guido Werner, and Wolfgang Witte. 2003. "Multiplex PCR assay for simultaneous detection of nine clinically relevant antibiotic resistance genes in Staphylococcus aureus." *J. Clin. Microbiol.* 41 (9):4089-4094. doi: 10.1128/jcm.41.9.4089-4094.2003.

Sullivan, Julie, Kalpana Karra, Sierra A. T. Moxon, Andrew Vallejos, Howie Motenko, J. D. Wong, Jelena Aleksic, Rama Balakrishnan, Gail Binkley, Todd Harris, Benjamin Hitz, Pushkala Jayaraman, Rachel Lyne, Steven Neuhauser, Christian Pich, Richard N. Smith, Quang Trinh, J. Michael Cherry, Joel Richardson, Lincoln Stein, Simon Twigger, Monte Westerfield, Elizabeth Worthey, and Gos Micklem. 2013. "InterMOD: integrated data and tools for the unification of model organism research." *Sci. Rep.* 3:1802. doi: 10.1038/srep01802.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to sequence learning with neural networks." Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, 2014/12/8.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering. 2019. "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." *Nucleic Acids Res.* 47 (D1):D607-D613. doi: 10.1093/nar/gky1131.

Thanos, M., G. Schonian, W. Meyer, C. Schweynoch, Y. Graser, T. G. Mitchell, W. Presber, and H. J. Tietz. 1996. "Rapid identification of Candida species by DNA fingerprinting with PCR." *J. Clin. Microbiol.* 34 (3):615-621. doi: 10.1128/JCM.34.3.615-621.1996.

The Gene Ontology Consortium. 2017. "Expansion of the Gene Ontology knowledgebase and resources." *Nucleic Acids Res.* 45 (D1):D331-D338. doi: 10.1093/nar/gkw1108.

Tieleman, T., and G. Hinton. 2012. "Lecture 6.5 - RmsProp: Divide the gradient by a running average of its recent magnitude."

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." *Nat. Protoc.* 7 (3):562-578. doi: 10.1038/nprot.2012.016.

Turabelidze, George, Steven J. Lawrence, Hongyu Gao, Erica Sodergren, George M. Weinstock, Sahar Abubucker, Todd Wylie, Makedonka Mitreva, Nurmohammad Shaikh, Romesh Gautom, and Phillip I. Tarr. 2013. "Precise dissection of an Escherichia coli O157:H7 outbreak by single nucleotide polymorphism analysis." *J. Clin. Microbiol.* 51 (12):3950-3954. doi: 10.1128/JCM.01930-13.

Turland, Nicholas J., Werner Greuter, Nick J. Turland, John H. Wiersema, Wolf-Henning Kusber, Fred R. Barrie, David L. Hawksworth, Patrick Stephen Herendeen, Sandra Knapp, Karol Marhold, De-Zhu Li, John McNeill, Tom William May, Anna M. Munro, Jefferson Prado, Michelle J. Price, and Gideon Smith. 2018. *International Code of Nomenclature for Algae, Fungi, and Plants (Shenzhen Code): Adopted by the Nineteenth International Botanical Congress, Shenzhen, China, July, 2017*.

UniProt Consortium, The. 2018. "UniProt: the universal protein knowledgebase." *Nucleic Acids Res.* 46 (5):2699. doi: 10.1093/nar/gky092.

Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. 2012. "Primer3—new capabilities and interfaces." *Nucleic Acids Research* 40 (15):e115-e115. doi: 10.1093/nar/gks596.

Untergasser, Andreas, Harm Nijveen, Xiangyu Rao, Ton Bisseling, René Geurts, and Jack A. M. Leunissen. 2007. "Primer3Plus, an enhanced web interface to Primer3." *Nucleic Acids Res.* 35 (Web Server issue):W71-4. doi: 10.1093/nar/gkm306.

Urban, Martin, Alayne Cuzick, Kim Rutherford, Alistair Irvine, Helder Pedro, Rashmi Pant, Vidyendra Sadanadan, Lokanath Khamari, Santoshkumar Billal, Sagar Mohanty, and Kim E. Hammond-Kosack. 2017. "PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database." *Nucleic Acids Res.* 45 (D1):D604-D610. doi: 10.1093/nar/gkw1089.

van Dijk, Erwin L., Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. 2014. "Ten years of next-generation sequencing technology." *Trends Genet.* 30 (9):418-426. doi: 10.1016/j.tig.2014.07.001.

Van Keuren-Jensen, Kendall, Jonathan J. Keats, and David W. Craig. 2014. "Bringing RNA-seq closer to the clinic." *Nat. Biotechnol.* 32 (9):884-885. doi: 10.1038/nbt.3017.

Vaske, Charles J., Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. 2010. "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM." *Bioinformatics* 26 (12):i237-45. doi: 10.1093/bioinformatics/btq182.

Verslyppe, Bert, Wim De Smet, Bernard De Baets, Paul De Vos, and Peter Dawyndt. 2014. "StrainInfo introduces electronic passports for microorganisms." *Syst. Appl. Microbiol.* 37 (1):42-50. doi: 10.1016/j.syapm.2013.11.002.

Vlek, Anneloes, Anna Kolecka, Kantarawee Khayhan, Bart Theelen, Marizeth Groenewald, Edwin Boel, Group Multicenter Study, and Teun Boekhout. 2014. "Interlaboratory comparison of sample preparation methods, database expansions, and cutoff values for identification of yeasts by matrix-assisted laser desorption ionization-time of flight mass spectrometry using a yeast test panel." *J. Clin. Microbiol.* 52 (8):3023-3029. doi: 10.1128/JCM.00563-14.

Walker, Louise A., and Carol A. Munro. 2020. "Caspofungin Induced Cell Wall Changes of Candida Species Influences Macrophage Interactions." *Front. Cell. Infect. Microbiol.* 10:164. doi: 10.3389/fcimb.2020.00164.

Walker, Louise A., Carol A. Munro, Irene de Bruijn, Megan D. Lenardon, Alastair McKinnon, and Neil A. R. Gow. 2008. "Stimulation of chitin synthesis rescues Candida albicans from echinocandins." *PLoS Pathog.* 4 (4):e1000040. doi: 10.1371/journal.ppat.1000040.

Wang, Can, Markus S. Schröder, Stephen Hammel, and Geraldine Butler. 2016. "Using RNA-seq for Analysis of Differential Gene Expression in Fungal Species." *Methods Mol. Biol.* 1361:1-40. doi: 10.1007/978-1-4939-3079-1_1.

Webb, Geoffrey I. 2010. "Naïve Bayes." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 713-714. Boston, MA: Springer US.

Westermann, Alexander J., Lars Barquist, and Jörg Vogel. 2017. "Resolving host–pathogen interactions by dual RNA-seq." *PLoS Pathog.* 13 (2):e1006033. doi: 10.1371/journal.ppat.1006033.

Westermann, Alexander J., Stanislaw A. Gorski, and Jörg Vogel. 2012. "Dual RNA-seq of pathogen and host." *Nature Reviews Microbiology* 10 (9):618-630. doi: 10.1038/nrmicro2852.

Westermann, Alexander J., and Jörg Vogel. 2018. "Host-Pathogen Transcriptomics by Dual RNA-Seq." *Methods Mol. Biol.* 1737:59-75. doi: 10.1007/978-1-4939-7634-8_4.

White, James Robert, Cynthia Maddox, Owen White, Samuel V. Angiuoli, and W. Florian Fricke. 2013. "CloVR-ITS: Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota." *Microbiome* 1 (1). doi: 10.1186/2049-2618-1-6.

Wilkening, Stefan, Manu M. Tekkedil, Gen Lin, Emilie S. Fritsch, Wu Wei, Julien Gagneur, David W. Lazinski, Andrew Camilli, and Lars M. Steinmetz. 2013. "Genotyping 1000 yeast strains by next-generation sequencing." *BMC Genomics* 14 (1):90. doi: 10.1186/1471-2164-14-90.

Wilson, M. R., S. N. Naccache, E. Samayoa, M. Biagtan, H. Bashir, G. Yu, S. M. Salamat, S. Somasekar, S. Federman, S. Miller, R. Sokolic, E. Garabedian, F. Candotti, R. H. Buckley, K. D. Reed, T. L. Meyer, C. M. Seroogy, R. Galloway, S. L. Henderson, J. E. Gern, J. L. DeRisi, and C. Y. Chiu. 2014. "Actionable diagnosis of neuroleptospirosis by next-generation sequencing." *N Engl J Med* 370 (25):2408-17. doi: 10.1056/NEJMoa1401268.

Wolf, Thomas, Philipp Kämmer, Sascha Brunke, and Jörg Linde. 2018. "Two's company: studying interspecies relationships with dual RNA-seq." *Curr. Opin. Microbiol.* 42:7-12. doi: 10.1016/j.mib.2017.09.001.

Wu, Yang, Binbin Shi, Xinqiang Ding, Tong Liu, Xihao Hu, Kevin Y. Yip, Zheng Rong Yang, David H. Mathews, and Zhi John Lu. 2015. "Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data." *Nucleic Acids Res.* 43 (15):7247-7259. doi: 10.1093/nar/gkv706.

Wysoker, Alec, Kathleen Tibbetts, and Tim Fennell. 2013. "Picard tools." Last Modified 2020. http://broadinstitute.github.io/picard/.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, Xin Zhou, Tak-Wah Lam, Yingrui Li, Xun Xu, Gane Ka-Shu Wong, and Jun Wang. 2014. "SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads." *Bioinformatics* 30 (12):1660-1666. doi: 10.1093/bioinformatics/btu077.

Ye, Jian, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, and Thomas L. Madden. 2012. "Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction." *BMC Bioinformatics* 13:134. doi: 10.1186/1471-2105-13-134.

Young, Matthew D., Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. 2010. "Gene ontology analysis for RNA-seq: accounting for selection bias." *Genome Biol.* 11 (2):R14. doi: 10.1186/gb-2010-11-2-r14.

Yu, Z. W., and P. J. Quinn. 1998. "Solvation effects of dimethyl sulphoxide on the structure of phospholipid bilayers." *Biophys. Chem.* 70 (1):35-39. doi: 10.1016/s0301-4622(97)00100-2.

Yuan, Shuai, and Zhaohui Qin. 2012. "Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression." *IEEE Int Conf Bioinform Biomed Workshops* 2012:718-724. doi: 10.1109/BIBMW.2012.6470225.

Zerbino, D. R., and E. Birney. 2008. "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." *Genome Research* 18 (5):821-829. doi: 10.1101/gr.074492.107.

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. 2018. "Ensembl 2018." *Nucleic Acids Res.* 46 (D1):D754-D761. doi: 10.1093/nar/gkx1098.

Zhang, Xinhua. 2010. "Support Vector Machines." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 941-946. Boston, MA: Springer US.

Zhao, Yanan, and David S. Perlin. 2014. "Use of Novel Tools to Probe Drug Resistance in Fungi." In *Handbook of Antimicrobial Resistance*, edited by Matthias Gotte, Albert Berghuis, Greg Matlashewski, Mark Wainberg and Donald Sheppard, 1-15. New York, NY: Springer New York.

Zoll, Jan, Eveline Snelders, Paul E. Verweij, and Willem J. G. Melchers. 2016. "Next-Generation Sequencing in the Mycology Lab." *Curr. Fungal Infect. Rep.* 10:37-42. doi: 10.1007/s12281-016-0253-6.