CHAPTER 1

Introduction

Candidiasis is one of the most common fungal infections in humans and is caused by species of fungi belonging to the Genus Candida. Although there are >20 species causing candidiasis, ~95% of the infections are caused by 5 species of Candida namely  
\( C. \text{ albicans}, \ C. \text{ tropicalis}, \ C. \text{ glabrata}, \ C. \text{ krusei} \) and \( C. \text{ parapsilosis} \). Species of Candida apart from albicans are classified as “Non-albicans clade” (NAC). The frequency of infections caused by different species of Candida varies with the geographical area under study. However, in general, \( C. \text{ albicans} \) is the most frequently encountered pathogen while \( C. \text{ tropicalis} \) is found to be one of the major players of NAC. The fact that \( C. \text{ tropicalis} \) is gradually turning azole resistant complicates the treatment regimen available for candidiasis. In spite of this being the case, there are very few molecular level studies on \( C. \text{ tropicalis} \) and most of the Candida biology is surrounding \( C. \text{ albicans} \).

Proteins are biological macromolecules that play major roles in physiology including 3D structure maintenance, cell cycle, metabolism, and signalling pathways. Further, proteins are also used as biomarkers and therapeutic targets for various diseases. Sequence information of proteins is the first step towards understanding their structure, functional regions such as domains and motifs. Sequence information of proteins cannot be complete if the genome of the organism is not well annotated for protein coding genes. Thus, a well annotated genome is a pre-requisite to elucidate mechanisms underlying virulence and infections. With the introduction of Next-Gen sequencing techniques, the number of novel organisms being sequenced has exponentially increased over the years (Figure 1.1). Owing to this, it is becoming increasingly difficult to follow classical approaches of gene calling and genome annotation. Although there are computational tools available for gene calling, the results are almost always never completely reliable. Over-prediction of genes and under-prediction of genes are commonly encountered.

The most reliable validation for the presence of a protein coding gene is to find the protein itself. In the past decade, there have been numerous improvements in the field of mass spectrometry-based proteomics and thus, it is now possible to profile hundreds of thousands of peptides in a single experiment. It is a high-throughput, high-accuracy technique and does
not need antibodies or other such “individual-protein-specific” reagents to detect the presence of proteins. This leads to an unbiased analysis during proteome profiling of any organism. In addition, modern mass spectrometers are highly sensitive and need minimal quantity of protein (in the range of µg to ng).

Figure 1.1: Progressive growth in the number of organisms being sequenced per year (Source: http://gregoryzynda.com/ncbi/genome/python/2014/03/31/ncbi-genome.html)

In the past few years, mass spectrometry-based proteomics has been used to refine the genome annotation of numerous organisms. In an integrated OMICS approach known as “proteogenomics”, spectra obtained from mass spectrometry experiments are searched against protein databases formed from predictive gene calling. This leads to identifying peptides that are valid evidence for the otherwise predicted genes. Further, spectra are also searched against various hypothetical databases such as 6-frame translated genome and 3-frame translated RNA databases (Figure 1.2). This results in identifying peptides that were not a part of the predicted genes. The so called Genome Search Specific Peptides (GSSPs) are then used to identify novel ORFs and alterations in annotated gene structures (Figure 1.3).

The genome of *C. tropicalis* was sequenced in 2009 and was annotated for protein coding regions using a predictive computational approach [1]. A total of 6,258 protein coding genes were annotated. However, there have been no studies to validate the presence of these genes. The authors write “The protocol used for Candida gene structure prediction..... this gene set
should be considered provisional, since it represents a first pass at gene calling, with comparative data used for only a subset of genes. We expect to refine this set of gene calls in the future….” [1]. To our knowledge, there have been no studies providing experimental validation for the predicted proteins of this important opportunistic pathogen. Therefore, we set out to study *C. tropicalis* at a molecular level with the following objectives:

1. High resolution mass spectrometry-based proteomic analysis of *Candida tropicalis*
2. Genome annotation of *Candida tropicalis* using a proteogenomic approach

---

**Figure 1.2:** Databases used for proteogenomic analysis [2]

**Figure 1.3:** Types of GSSPs and the alterations in genome annotation they bring about