Summary

Introduction

_Candida tropicalis_ is an opportunistic pathogen which causes candidiasis in immune-compromised individuals. It is one of the members of the non-albicans group of _Candida_ that are known to be azole resistant and is frequently seen in individuals being treated for cancers, HIV-infection and bone-marrow transplant. The genome of _C. tropicalis_ was sequenced in 2009 and the annotated genome consists of 23 supercontigs and 6,258 predicted protein coding genes. However, the genome annotation has not been supported by experimental validation. To this end, we performed a high resolution, high accuracy mass spectrometry-based proteomic analysis of _C. tropicalis_. We also used the mass spectrometry derived data to refine the genome annotation using a proteogenomic approach.

Objectives

1. High resolution mass spectrometry-based proteomic analysis of _Candida tropicalis_
2. Genome annotation of _Candida tropicalis_ using a proteogenomic approach

Methods

_C. tropicalis_ was cultured in YNB media and the cells and supernatant were collected. Protein extracts from the cells and conditioned media was subjected to protein and peptide level fractionation using SDS-PAGE and basic pH reverse phase liquid chromatography respectively. The peptide fractions obtained were analyzed using Fourier transform mass spectrometry and the data was searched against the known protein database of _C. tropicalis_ using Mascot and SEQUEST search engines to confirm the predicted genes. Bioinformatics analysis was carried out using SignalP and TMHMM. The data was also searched against six-frame translated genome database to identify novel gene models and refinements to the current annotation of the genome. The Integrative Genomics Viewer (IGV) was utilized to perform proteogenomic analysis.
Results

A total of 23,173 unique peptides were identified that mapped to 2,743 proteins. This coverage experimentally validated ~44% of the predicted proteome of *C. tropicalis*. We identified 152 proteins in the conditioned media out of which 79 were potentially secreted. A large number of these proteins are known players in the pathogenesis of candidiasis.

Using a hypothetical six-frame translated genome database, we identified 627 peptides that did not correspond to any of the known proteins. These GSSPs, when mapped on to the annotated genome, led to the identification of novel protein coding regions and gene structure correction events. We identified a total of 86 novel genes, 12 novel exons, and refined the coordinates of 49 computationally predicted gene models.

Conclusions

High-resolution mass spectrometry based proteomic profile of *C. tropicalis* was obtained thus providing experimental evidence for the otherwise predicted proteome. Proteogenomic analysis using mass spectrometry generated proteomic data led to identification of 86 genes not annotated previously. Gene structure changes to a large number of genes were also proposed based on this analysis. This approach of using mass spectrometry-based proteomic data to annotate protein coding genes in the genome can prove to be an essential method complementary to computational methods for annotating both newly sequenced genomes as well as genome sequences which have been available for many years.