Annotation of Fungal Transposable Elements

Jaydeep Srimani1, Igor Grigoriev1,2, Asaf Salamov1,2
1DOE Joint Genome Institute, 2Lawrence Berkeley National Laboratory

Methods & Validation

The transposon categories and structural characteristics are shown in Figure 1 [3]. The classification system uses RepeatScout [4] (RS) output files (consensus sequences and location coordinates of repeat candidates), and searches for both DNA/protein homology and the occurrence of these features.

BLASTn and BLASTp DNA/protein homology is used to isolate the best hit for each RS repeat sequence. The LTR/TIR search uses Smith-Waterman string matching to find nearly identical strings with minimal length of 70bp and 50bp for LTR and TIR, respectively. The Helitron search examines 10bp regions flanking transposon occurrences for secondary structure motif, and searches the trans.

The transposon categories and structural characteristics are shown in Figure 1 [3]. The classification system uses RepeatScout [4] (RS) output files (consensus sequences and location coordinates of repeat candidates), and searches for both DNA/protein homology and the occurrence of these features.

BLASTn and BLASTp DNA/protein homology is used to isolate the best hit for each RS repeat sequence. The LTR/TIR search uses Smith-Waterman string matching to find nearly identical strings with minimal length of 70bp and 50bp for LTR and TIR, respectively. The Helitron search examines 10bp regions flanking transposon occurrences for secondary structure motif, and searches the trans.

Abstract

Transposable elements (TE), also called transposons, in fungi have been shown to be regions of high DNA variability and subsequent gene evolution. This flexibility renders them important for fungal adaptation to changing hosts and environments. Several families of TEs have been identified; however, the methods and pipelines used to classify them are relatively new and rely heavily on BLAST output. In this study we augment traditional nucleotide and protein searches with statistics on notable features of both RNA and DNA transposons. Specifically, we examine long terminal repeats (LTRs), terminal inverse repeats (TIRs) and helitron motifs. These features are used to confirm the results of homology testing. This classification system was tested using previously annotated transposons from three genomes: Lactuca biicolor (73 sequences), Aspergillus nidulans (38), and Saccharomyces cerevisiae (18) and correctly annotated all of them. When run on a wide array of fungi, including several closely related members of the Ascomycete and Basidiomycetes families, the system revealed a relatively low proportion of transposons (~10%), with the exception of A. alcalophilum and L. bicolor but significant diversity between even closely related species.

Introduction

Fungal genomics represent a vast potential for ameliorating energy demands in the US. There are several established genetic characteristics/motifs for annotating fungal and plant genomes. One such characteristic is the transposable element (TE): segments of DNA that often replicate and insert themselves hundreds to thousands of times within a genome. Transposons increase genetic diversity and can affect gene expression via mutations and alteration of regulating mechanisms [1]. Various genomes have also demonstrated the ability to adapt to potentially disadvantageous transposon insertions by down-regulating affected genes or showing a tendency towards greater transposon presence in non-coding DNA segments [2]. TEs have been previously classified into families based on several auxiliary attributes and DNA homology to known repeat sequences. These methods rely heavily on BLAST searching, and can be difficult to transfer to different hardware systems [3].

We implement additional searches based on structural characteristics, including long terminal repeats (LTR), terminal inverse repeats (TIR), and helitrons. These additions greatly increase the number of transposons successfully classified, both in our validation study as well as in numerous full genomes.

Results

For both Ascomycetes and Basidiomycetes, TEs are dominated by LTRs, followed by non-LTR retrotransposons. However, LTR lengths varied greatly for all species.

For most species, repeat fraction was consistently low (<10%), regardless of genome size.

Homology accounted roughly half of all classifications in both groups (53.7% and 54.5%); a significant number of results were derived from the added feature classifications.

Septoria Case Study:

• Two closely related Septoria fungi with significantly different repeat fractions were studied. Table 2 shows S. populicola’s larger genome cannot account for the repeat fraction increase. We hypothesize there are undetected repeats in S. musiva. Combining the libraries yields marginal repeat fraction increase in both genomes.

• Decreasing the RS threshold from the standard 150x yields comparable increases for both. We conclude that the repeat content of S. populicola and musiva are significantly different.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Frequency Threshold</th>
<th>S. musiva</th>
<th>S. populicola</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>150x</td>
<td>4.2%</td>
<td>20.8%</td>
</tr>
<tr>
<td></td>
<td>100x</td>
<td>6.1%</td>
<td>20.8%</td>
</tr>
<tr>
<td></td>
<td>50x</td>
<td>7.7%</td>
<td>23.1%</td>
</tr>
<tr>
<td></td>
<td>10x</td>
<td>9.9%</td>
<td>25.6%</td>
</tr>
<tr>
<td></td>
<td>1x</td>
<td>10.4%</td>
<td>26%</td>
</tr>
</tbody>
</table>

Table 2-3. Septoria repeat fractions for separate and combined repeat libraries (Table 2) and decreasing RepeatScout threshold (Table 3).

Conclusions

• We developed a system for transposon classification, and validated it using curated sequences.

• For tested genomes, Class I retrotransposons dominate the classification, but even closely related genomes may show significantly different repeat content, as illustrated by Septoria.

References


Acknowledgements

The authors would like to thank the Department of Energy’s Advanced Scientific Computing Initiative (Science & Technology) for support. The work conducted at the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.