BIO 724 Final Project

Kayla Wilhoit

Questions 1-3)

Infections by the fungal pathogen *Cryptococcus* are primarily caused by the species *C. neoformans,* with a small number of infections caused by the related *C. deneoformans*. The two species, previously described as a species complex, are estimated to have diverged 24.5 million years ago. Despite several genomic changes including a translocation between chromosomes 3 and 11, diploid hybrids between the two species are relatively common and are known to cause infection in human patients. While it is clear that these hybrids are an important factor in cryptococcal infections, little is known about the genomic interactions and gene expression changes that could potentially affect virulence.

The Magwene lab currently has access to several diploid hybrid samples collected from patients at Duke hospital, as well as hybrid samples collected from the environment around North Carolina and nonhybrid haploid samples. I plan to perform RNA-seq on both hybrid and nonhybrid samples to assess the impact of hybridization on gene expression levels and potentially identify cis and trans regulatory changes.

My main questions are:

1) How are hybrid cryptococcus samples behaving transcriptionally?

- Are parental haplotypes separately regulated?
- Cis/trans effects?
- Differences from synthetic hybrids?
- Isoforms?

2) How do ploidy changes affect transcriptional regulation in clinical cryptococcus samples?

- Linear increase in expression?
- Differences in hybrids/nonhybrids?

3) How do chromosomal rearrangements affect transcriptional regulation in hybrid samples?

- Cis/trans changes?
- Physical interactions?

The main computational challenge I anticipate facing is the structural rearrangements between the two species, coupled with the ploidy variation between the diploid hybrids and the haploid parental species. In addition to a known translocation between chromosomes 3 and 11, read depth analysis of the hybrid clinical isolates has suggested multiple additional deletion or translocation events, as well as an unequal distribution of ploidy between the two parental genomes. It may be useful to create synthetic hybrids of known parentage to analyze gene expression in a more controlled context, and I am particularly interested in how gene regulation may have evolved in the stable clinical and environmental samples compared to newly hybridized strains. These factors will all contribute to difficulties in accurately quantifying gene expression and normalizing RNA-seq results.

https://doi.org/10.1371/journal.pgen.1009409

The authors of the linked study addressed gene regulation and misexpression in hybrids between two *Caenorhabditis* nematode species, and have published their pipeline and tools in a github repository. The study focuses extensively on sex-related factors and potential contributions to reproductive isolation, which would be interesting to analyze in cryptococcal hybrids but is less directly applicable to the fungal mating type system.

There are several studies of gene expression and transcriptional regulation in yeast hybrids, but much of the relevant work has been done in species such as *Saccharomyces cerevisiae* or used older experimental and computational techniques. I hope to integrate yeast genetics with computational techniques developed to disentangle complex genome structure to perform a novel analysis of transcriptional regulation in a pathogenic yeast.

Question 4) – Data formats and metadata

I will primarily be using sequencing data in the form of paired-end short read RNA-seq output fastq files. During QC, alignment, and mapping, the original data will be likely represented in BAM/SAM and BED files. There will be different levels of relevant metadata for each sample. Of immediate importance for analysis will be information about the hybrid status and ploidy of each sample, which will determine factors in the pipeline such as needed reference genomes and normalization for differential expression. Of larger importance for the research question is metadata about the sample source, including patient location, duration of infection, and relevant comordibities.

Question 5) – Major software tools

A typical RNA-seq analysis pipeline would include quality control, read alignment, quantification, and differential expression analysis. As there are already several good-quality reference genomes and some annotations available for the relevant *Cryptococcus* species, I plan to primarily use a reference genome-based mapping approach. However, due to the extensive chromosomal rearrangements and ploidy differences in some hybrid samples, it may be of use to investigate creation of de novo transcript assemblies for comparison with the genome mapping results. The normalization and differential expression steps may need adaptable tools or specific tools that are able to deal with different ploidy and chromosomal duplications.

Potential tools and software would include:

- Quality control: **FastQC, Trimmomatic, Cutadapt, Skewer**
- Read mapping/alignment: **STAR, Stampy, Bowtie2**
- Normalization: **RSEM, edgeR**
- Differential expression: **DESeq2, edgeR**
- de novo assembly: **Trinity, Trans-ABySS, SOAPdenovo-Trans, BUSCO, DETONATE**

The majority of these tools are open source and are able to run on a linux system, which is what I will likely be using.

Question 6) – Computing needs

A challenge for RNA-seq analysis using paired-end short reads will be the large amount of raw data available. Despite the relatively small genome of Cryptococcus, there are many both hybrid and non-hybrid samples that would be available for sequencing. The small size of the genomes should make memory limitations less of a problem, but there will still be storage considerations. The samples would likely not be able to run on a laptop, but workstations in the lab would likely be the best place to run analysis. I anticipate many intermediate files being produced with all of the analysis workflow steps, but the most important final data files would need to be stored and backed up.

Question 7) – Workflow organization

I plan to create a workflow in Snakemake or similar system to make the steps and scripts for creating the final datasets and visualizations reproducible. This would include setting up built-in conda environments to be activated through snakemake, to ensure that software version changes would not necessarily affect the pipeline output. The code for analysis would be kept in a private GitHub repository available to lab members. I plan for the scripts, final outputs, and original sequencing data to be backed up on github and/or Dropbox or Duke Box online, as well as physical copies in at least two locations. The intermediate files and even many of the less important output files are likely to be easily regenerated due to the small genome size of *Cryptococcus*. This will allow more temporary storage in perhaps less robust frameworks such as lab computing clusters and hard drives.

Question 8) - Branch points and decision points

There are several major decision points in this project. Some decisions such as how to analyze isoforms and whether to create de novo assemblies will both require and inform decisions about sequencing platform, read length, and number of samples to analyze. Other decisions such as statistical significance cutoffs and differential expression goals can be made later in the process, but having a good understanding of the possibilities will inform the process.

Isoforms – include or not?

- May depend on sequencing platform and cost
- Long + short reads?
- Likely to be complicated in messy hybrids
- May need small preliminary dataset first

Alignment of reads – reference or de novo or both?

- Previous yeast analyses mostly reference-based
- May also create de novo genome assemblies

Differential expression – statistical significance

- Looking for specific virulence genes or general patterns?
- What are the significance cutoffs?

Question 9) - Visualization

The primary visualization output would be in the form of volcano plots for differential gene expression analysis. I also plan to integrate visualization steps into many of the pipeline steps, in order to keep track of changes to the data and catch errors early. Examples would include performing alignments and read depth analysis on all of the samples to confirm the expected ploidy and detect major chromosomal rearrangements. As reads are mapped and normalized, they will be visualized using genome browser tools for areas of interest or known controls, and the distribution of reads will be summarized in tables and simple plots to ensure consistent and expected results. I plan to perform most of the visualization in R using ggplot and related packages, but several steps in the pipeline may include their own preferred plotting software, which will require some research to keep consistency between.

Question 10) – Statistical challenges

Many of the statistical challenges of this project will relate to normalizing and analyzing the gene expression data from samples with variable ploidy and parentage. In a typical “triangle” structure, both the parental haploid strains and the hybrid offspring would be known and sequenced. While this will be possible using lab-generated hybrids, parentage will be unclear for the clinical hybrid isolates, which may have originally hybridized over 30 years before the present.

Normalization and differential expression

- How to deal with ploidy changes?

A paper comparing S. cerevisiae hybrids found small differences in differential expression between haploid and diploid samples, but that assumes an even distribution of ploidy across the chromosomes and parental haplotypes. Many of the clinical hybrids have uneven ploidy and/or chromosome biased toward one of the parental haplotypes, which is likely to affect differential expression metrics more significantly.

Synthetic hybrids vs. clinical hybrids

- Creating in vitro hybrids = better cis/trans calling, compare early/established hybrids
- Both parents known and available
- May not be stable or ‘realistic’

Clinical strains may not have clear parental strains available

- Could split haplotypes and treat as parents?
- Could use closest sequenced strain to each haplotype?

Future Plans

Testing on example cryptococcus samples and nematode “triangle” dataset

- Choosing software tools
- Testing with/without parents included
- Testing with imputed parents from offspring haplotypes

Finding more information on current software

- Isoform information
- Cis/trans calling
- Short + long read sequencing
- Hybrid organisms

Narrowing goals

- Looking for specific genes (drug resistance)
- Patterns temporally/geographically