mRNA-seq Traditional Time Course Analysis Across Meiotic Axis Genes

Group Members (1): Darryl Fung

Dr. Manpreet Katari's Applied Genomics Spring 2024 (Section 3)

PI: Dr. Andreas Hochwagen

Table of contents

Abstract & Background

I will be analyzing RNAseq data from the paper by Brar et al, 2012 (High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling) about meiotic yeast recombination and gene factors under different conditions. My PI Dr. Hochwagen and I are interested in this topic because these genes uniquely have lower binding with the meiotic axis proteins I am reviewing for my literature thesis, which reviews the function and discoveries in meiotic axis proteins Hop1, Red1, and Rec8 about their transcriptional and post-transcriptional regulation in yeast meiosis.

This project finds that the average gene expression for the genes in the pericentromeric, subtelomeric, and ribosomal regions is slightly decreased compared to the control group. A pattern appears of immense expression in the beginning stages of the meiotic timecourse (0.5 hours), followed by a drop leading to a second, smaller peak of expression at 24 hours in the time course.

The Brar paper has data from their mRNA-seq traditional time courses of genes in three regions of interest: the pericentromeres, the subtelomeres, and the regions surrounding ribosomal DNA. The time courses show the expressions of these regions at different times during meiosis, from 0 hours to 24 hours. (A to V) (Brar et al, 2011)

Hypothesis & Goal

The gene expressions should be independent of the meiotic axis proteins due to having low binding. They shouldn't be affected by the meiotic axis proteins' activity. We would find this through no specific pattern recognizable from the timecourse data.

I set out to complete the following objectives in this project:

1. Identify all the genes in the three regions of interest and calculate the average expression (exp) level for each time point (A through V).
2. As a control group, I will calculate the average exp level of all the genes in the dataset that do not fall within the regions of interest (pericentromeres, subtelomeres, ribosomal-surrounding).
3. Culminate the data in two graphs: one of the average exp levels in the 3 regions of interest and one of the control group, both as a function of time in meiosis.

The goal was to observe any special patterns across meiosis as these genes uniquely have lower binding to my literature thesis' proteins of interest Hop1, Red1, and Rec8.

Results & Discussion

Out of the 5480 genes present in this data set, only 572 were identified to be part of the three regions of interest. There were 415 found in the pericentromeric, 108 found in the subtelomeric, and 55 in the ribosomal. The lists of genes can be found in the attached zip file > Files > Results, under GenesPresent, Genes-in-PericentromericRegions.txt, Genes-in-RibosomalRegions.txt, and Genes-in-SubtelomericRegions.txt. For an aggregated list of the genes in any of the three regions, find Genesin3Regions.txt.

The gene expression results of this analysis are given in two figures of the average gene expression over time, one in the three regions of interest (Fig. 1), and one in the control group (Fig. 2)

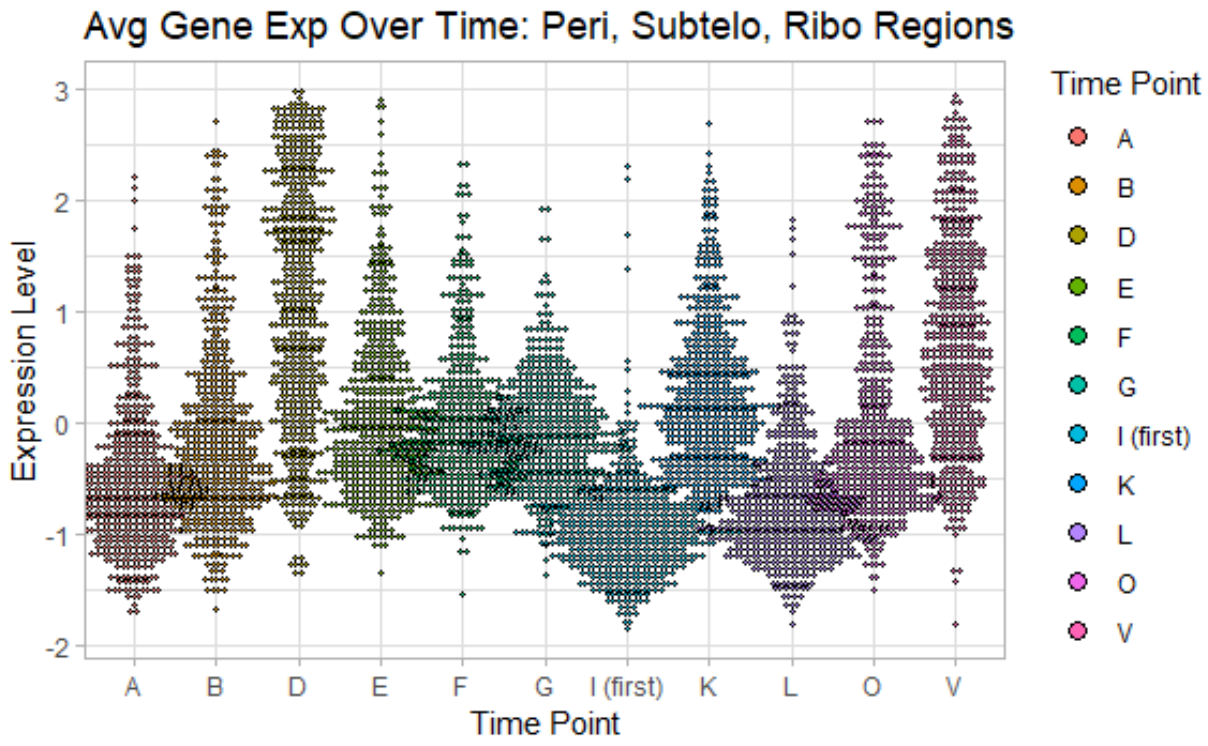

Fig. 1: The average gene expression over time for time points A-V in the three regions of interest. (pericentromeric, subtelomeric, and ribosomal)

Although not showing any specific genes (there would be too many to distinguish by eyesight-- there are over 570), we can see patterns from genes within the three special regions. Within the three special regions, there is a notably low expression at the start of the time course. It appears that none of these genes are master regulators, as none of them remain active at the start of meiosis to activate other genes. As early as the B time point, 0.033 hours in (two minutes), we can see the expression of these genes begin to rise. They reach a peak at 0.5 hours in, where there is much overexpression in these genes. The expression levels drop to a medium-range level

(near 0) and fluctuate until 24 hours in, where expression rises but not to the point of the overexpression peak early on.

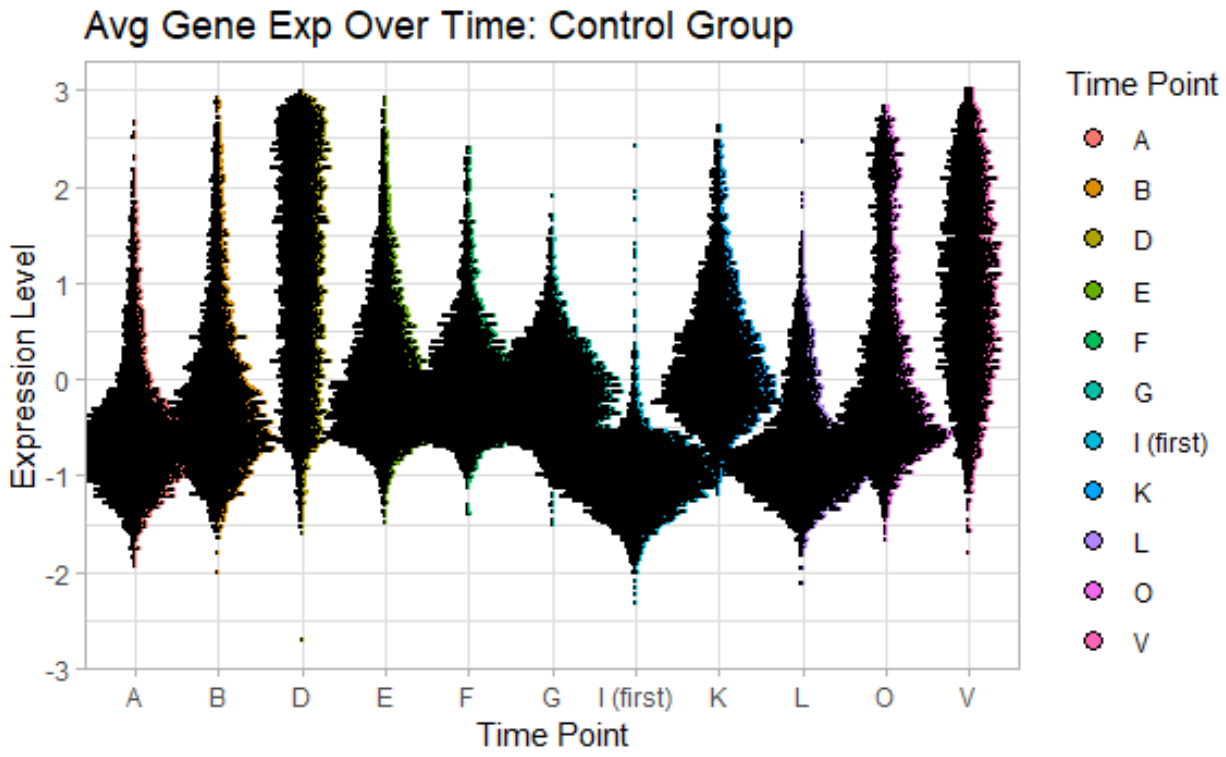

Fig. 2: The average gene expression over time for the control genes, i.e. the genes not residing in the three special regions.

The control genes follow the same general trend as those within the regions of interest, but there is a mild difference in expression levels at the two highest points. (Fig. 2) The biggest visible difference is in point D, where more of the control genes

exhibit the nearly-abnormally high expression. The 24-hour point (V) has genes with slightly higher expression than in the three regions.

Due to the small difference in expression between the regions of interest and the control group, there appears to be no discernable relationship between these and Hop1/Red1/Rec8.

Procedure

File Setup and Alignment

I was provided with GFF annotation and FASTA genome file for the SK1 strain of *Saccharomyces cerevisiae* by my supervisor, Dr. Andreas Hochwagen. In this project, I performed the following steps:

First, using fasterqdump from the sratools package, I obtained fastq files for the SRR values corresponding to the mRNA-seq traditional time course. I ultimately decided to focus on the traditional time course because it matches the method my PI's lab uses. I made this choice for the sake of familiarity and time. The files used included the initial header file and timepoints A, B, D, E, F, G, I (first file), K, L, O, and V.

STAR alignment was done the following way: Create indexes for and aligned all twelve .fastq files separately. This is Version 3 of the alignment methodology, details of which you can read about in “Limitations and Alternate Interpretations.” In RStudio, each count table was loaded and combined into a table with twelve columns, each holding each gene’s readcount for every timepoint.

In RStudio, after combining the count tables from DSS, I defined the genes residing in the pericentromeric, subtelomeric, and ribosomal regions of the genes.

The pericentromeric regions are 25,000 bases up and downstream of the centromere on each chromosome (16 in total). The subtelomeric regions are the regions 0 to 20,000 bases from each of the sixteen chromosomes’ ends. (32 in total) The ribosomal regions are only present on chromosome 12 and are 50,000 bases up and downstream of the rDNA loci. (RDN18-1 and RDN18-2).

My first operation in RStudio after combining columns was to extract the gene ID’s and names from the 9th column of the GFF annotation file. For context, the GFF file’s structure includes gene ID and either gene name (for genes) or gene parent (for subfeatures) information in the 9th column. The format looks like this:

ID=...,Name=... (or for subfeatures, ID=...,Parent=...) To extract the ID’s and names, I captured the text that read for ID, the text that read for gene name, and added them to two new columns gene_id and gene_name. This was done in order to find the proper gene id’s and names that corresponded to the combined readcounts table, and

to find which genes (by name) resided in the pericentromeric, subtelomeric, or ribosomal regions. With this added gene ID and gene name information, I was able to define where genes (by name) started and ended on the chromosomes, as the updated GFF table includes start and end positions for each element, the number chromosome they reside on (I to XVI), and – newly added – gene ID and gene name. I would use this information to line up which genes by name fell within each of the three special regions per chromosome.

For the subfeatures (Parent=...), note that the Parent section actually reads the gene ID of the parent that it belongs to. In the gene ID column, I gave each subfeature the same ID as the parent that they belonged to.

Defining the three regions of interest

To define the pericentromeric regions, the centromere location of each chromosome was required. This was included in the provided GFF file as CEN1 to CEN16. As the centromeres are not a singular point, they naturally have start and end positions of their own. I located the start of the pericentromeric region as 25,000 bases upstream of the centromere's start and the pericentromeric end as 25,000 bases downstream of the centromere's end.

Identifying genes present in the three regions

Before defining the genes present in the three regions of interest, I ran code to run the unique names of all gene names present in the data set (GFF table). This effectively defined all the genes present in the data set. I would filter out this vectors for those names who lined up in the three regions of interest later.

To define the genes within the three regions of interest, I filtered the GFF file (which had start and end positions for each gene ID) for the gene ID's that fit within each of the three regions, lined them up with the IDs' corresponding gene names, and saved them as vector objects "Genesin__", where the line could be Peri, Sub, or Ribo.

Calculating average gene expressions

In the readcounts table, the genes were lined up to ID's rather than gene names. I used the modified GFF table with gene ID's and gene names to line up each gene ID and gene subfeature and give them the same gene name in the readcount table under a new column. The result of this name assignment would lead to several rows with the same gene name – these rows were to be averaged into the average gene expression.

The average gene expression was used because each of the several elements per gene had their own read counts per time point in the whole timecourse. Although

there were still too many genes to individually manipulate, calculating the average made the data more manageable. Each gene would have one set of values rather than four or five. I also have exported as .csv files the read count data of the genes in the three regions and those outside of the three regions, as well as a read count of all the genes.

Limitations and Alternate Interpretations

The STAR alignment workflow went through many revisions. There were three versions of the alignment used.

Version 1: I initially attempted to sample all 70 fastq files from Brar et al's NCBI GEO sits, but the time-consuming nature of this process led to a re-evaluation and trim-down to 12 files.

Version 2: Under the 12 meiotic time course files, I initially ran them through the same loop (sbatch, which created only one BAM file for ht-seq to create a count table from. However, upon loading into RStudio, this method proved flawed, creating only one column of gene expression values. This version had counts in only one column, i.e. only one time point. However, the expected count table was to include multiple gene expression counts for each gene ID, one for each time point.

Version 3: Instead of aligning all 12 fastq files in the same operation, I finally opted to create indexes for and align all twelve separately. (Since each of the twelve had its own STAR operation, I gave Version 3 of my strategy an internal name “Operation Duodenary STAR System.” (DSS)) Upon running ht-seq count on each of the twelve resulting output files, twelve read count tables with one column each resulted. Upon loading in each in RStudio and sequentially binding them together into a dataframe, I now had a read count table with twelve columns, each holding each gene’s readcount for every timepoint.

Future Work

In the future, a follow-up analysis can be performed to analyze specifically the average gene expression in each of the three regions of interest. Comparing each of these to the control might give more detailed insights into gene patterns.

A follow-up analysis can also be performed using the Ndt time course, being able to draw conclusions on whether this set of genes within the regions is affected by the Ndt gene that is used to start the Ndt time course.

Bibliography

1. Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., & Weissman, J. S. (2011). High-Resolution View of the Yeast Meiotic Program Revealed by

Ribosome Profiling. *Science, 335(6068)*, 552–557. https://doi.org/10.1126/science.1215110

2. Supplementary Meiotic TIme Course Data from Brar et al (Also included as a PDF)
3. NCBI GEO link for Brar et al: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34082