Project 1. Transcriptome Assembly (Extra Tasks)

Indronil Bhattacharjee, Erica Flores

Task E1: Using STAR

E1.1 The STAR aligner was used to map the RNA-seq reads to the reference genome. The comparison of the results from STAR and HISAT2, in terms of the numbers or percentages of reads uniquely aligned to the reference genome is in the following -

Sample	STAR Reads	STAR Percentage	HISAT2 Reads	HISAT2 Percentage
Testis_Replicate_1	11013738	96%	10343346	89%
Testis_Replicate_2	9018430	97%	8421960	90%
Ovary_Replicate_1	16200742	98%	15183182	91%
Ovary_Replicate_2	11354930	98%	10601218	91%

E1.2 STAR by default generated a tab-separated file ("SJ.out.tab") that contains all unique splicing junctions and the number of reads that cross each splicing junction.

Here is the R code to map the junctions to known genes if the start and end sites of a junction are entirely contained within a gene on the chromosome.

```
library(dplyr)
```

```
gtf_file <- "/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-junctions/genes_only.g

# Create an empty data frame to store gene information
genes <- data.frame()

# Read the GTF file line by line
```

```
con <- file(gtf_file, "r")
while (length(line <- readLines(con, n = 1)) > 0) {
  if (!grepl("^#", line)) {
    fields <- unlist(strsplit(line, "\t"))

    if (fields[3] == "gene") {
      # Extract gene ID from the attributes field
      gene_id <- gsub(".*gene_id \"(.*?)\";.*", "\\1", fields[9])

      # Create a data frame with gene information
      gene_info <- data.frame(
        chromosome = fields[1],
        start = as.numeric(fields[4]),
        end = as.numeric(fields[5]),
        gene_id = gene_id
      )

      # Append the gene_info to the genes data frame
      genes <- rbind(genes, gene_info)
    }
  }
}
close(con)
```

Splice-gene mapping function

::: {.cell _kg_hide-output='true' execution='{"iopub.execute_input":"2023-10-18T22:16:35.553520Z","iopub.st
10-18T22:16:35.552121Z","iopub.status.idle":"2023-10-18T22:16:35.565042Z"}' trusted='true'
vscode='{"languageId":"r"}' execution_count=166}

```
splice.gene.mapping <- function(sj_data){
    # Create an empty data frame to store the mapping of splice junctions to genes
    sj_to_gene_mapping <- data.frame(chromosome = sj_data$chromosome, start = sj_data$star

    # Add a column to store the mapped gene IDs
    sj_to_gene_mapping$gene_id <- NA

    suppressWarnings({
    # Iterate through each splice junction
    for (i in 1:nrow(sj_to_gene_mapping)) {
      # Find the gene that contains the splice junction
      gene_id <- genes$gene_id[genes$chromosome == sj_to_gene_mapping$chromosome[i] &
```

```
                            genes$start <= sj_to_gene_mapping$start[i] &
                            genes$end >= sj_to_gene_mapping$end[i]]

      # If a gene is found, add its ID to the sj_to_gene_mapping data frame
      if (length(gene_id) > 0) {
        sj_to_gene_mapping$gene_id[i] <- gene_id
      } else {
        sj_to_gene_mapping$gene_id[i] <- NA
      }
    }})
    sj_to_gene_mapping_out <- sj_to_gene_mapping[!is.na(sj_to_gene_mapping$gene_id), ]
    return (sj_to_gene_mapping_out)
}
```

:::

```
sj_file <- "/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-junctions/testis_replic

# Read the SJ.out.tab file into a data frame
sj_data <- read.table(sj_file, header = FALSE, sep = "\t")
colnames(sj_data) <- c("chromosome", "start", "end", "strand", "intron_motif", "intron_ann

sj_to_gene_mapping <- splice.gene.mapping(sj_data)
head(sj_to_gene_mapping,10)
```

A data.frame: 10 × 5

	chromosome <chr>	start <int>	end <int>	total_reads <int>	gene_id <chr>
1	2L	8117	8192	343	FBgn0031208
2	2L	8117	8228	29	FBgn0031208
3	2L	11345	11409	1	FBgn0002121
4	2L	11519	11778	1	FBgn0002121
5	2L	17213	19879	1	FBgn0002121
6	2L	22942	22997	5	FBgn0031209
7	2L	26965	27052	5	FBgn0051973
8	2L	34289	34557	11	FBgn0051973
9	2L	70550	70606	2	FBgn0067779
10	2L	72978	74902	4	FBgn0031213

```r
sj_file <- "/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-junctions/ovary_replica

# Read the SJ.out.tab file into a data frame
sj_data <- read.table(sj_file, header = FALSE, sep = "\t")
colnames(sj_data) <- c("chromosome", "start", "end", "strand", "intron_motif", "intron_ann
sj_to_gene_mapping <- splice.gene.mapping(sj_data)
head(sj_to_gene_mapping,10)
```

A data.frame: 10 × 5

	chromosome <chr>	start <int>	end <int>	total_reads <int>	gene_id <chr>
1	2L	11345	11409	129	FBgn0002121
2	2L	11519	11778	50	FBgn0002121
3	2L	12222	12285	82	FBgn0002121
4	2L	12929	13519	27	FBgn0002121
5	2L	13493	13559	1	FBgn0002121
6	2L	13626	13682	52	FBgn0002121
7	2L	14875	14932	63	FBgn0002121
8	2L	15712	17052	19	FBgn0002121
9	2L	17208	18260	1	FBgn0002121
10	2L	17208	21065	1	FBgn0002121

E1.3 Top 5 splice junctions of high fold-change and low fold-change using the formula from Task 3.

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
```

Between Replicates 1

```r
# Read the SJ.out.tab file into a data frame
testis_replicate1 <- read.table("/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-ju
testis_replicate2 <- read.table("/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-ju
ovary_replicate1 <- read.table("/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-jun
ovary_replicate2 <- read.table("/kaggle/input/star-fruit-fly-ovary-and-testis-splicing-jun

testis_merged <- testis_replicate1 %>% semi_join(ovary_replicate1, by = c("V2", "V3"))
ovary_merged <- ovary_replicate1 %>% semi_join(testis_replicate1, by = c("V2", "V3"))
```

```
# Calculate the average transcript abundance for ovary and testis
avg_abundance_ovary <- ovary_merged[,7]+ovary_merged[,8]
avg_abundance_testis <- testis_merged[,7]+testis_merged[,8]

# Calculate the log fold change (r_g)
log_fold_change <- abs(log2((1 + avg_abundance_ovary) / (1 + avg_abundance_testis)))
```

```
testis_merged$log_fold_change <- log_fold_change[1:nrow(testis_merged)]
top_transcripts_testis_highest <- testis_merged %>% arrange(desc(log_fold_change)) %>% hea
top_transcripts_testis_lowest <- testis_merged %>% arrange(log_fold_change) %>% head(5)
```

```
top_transcripts_testis_highest
top_transcripts_testis_lowest
```

A data.frame: 5 × 10

	V1 <chr>	V2 <int>	V3 <int>	V4 <int>	V5 <int>	V6 <int>	V7 <int>	V8 <int>	V9 <int>	log_fold_change <dbl>
1	3L	10639857	10639919	1	3	1	21889	14	50	11.248982
2	2L	19183339	19183391	2	2	1	4678	0	50	11.191985
3	3R	11240813	11240865	2	2	1	18	2623	50	9.782452
4	2L	21628210	21629666	2	2	1	1	0	25	9.523562
5	2L	18687127	18687186	2	2	1	1	0	7	9.321928

A data.frame: 5 × 10

	V1 <chr>	V2 <int>	V3 <int>	V4 <int>	V5 <int>	V6 <int>	V7 <int>	V8 <int>	V9 <int>	log_fold_change <dbl>
1	2L	123795	123855	2	2	1	4	0	31	0
2	2L	203892	203992	1	1	1	2	0	45	0
3	2L	547901	547977	2	2	1	3	0	29	0
4	2L	815422	815474	2	2	1	6	0	33	0
5	2L	852443	852498	2	2	1	14	0	45	0

```
ovary_rows_h <- ovary_merged[ovary_merged$V2 %in% top_transcripts_testis_highest$V2, ]
ovary_rows_l <- ovary_merged[ovary_merged$V2 %in% top_transcripts_testis_lowest$V2, ]
```

```
colnames(top_transcripts_testis_highest) <- c("chromosome", "start", "end", "strand", "int
data_highest <- splice.gene.mapping(top_transcripts_testis_highest)
data_highest
```

A data.frame: 5 × 5

	chromosome <chr>	start <int>	end <int>	total_reads <int>	gene_id <chr>
1	3L	10639857	10639919	21903	FBgn0036091
2	2L	19183339	19183391	4678	FBgn0267726
3	3R	11240813	11240865	2641	FBgn0037888
4	2L	21628210	21629666	1	FBgn0022893
5	2L	18687127	18687186	1	FBgn0005617

```
colnames(top_transcripts_testis_lowest) <- c("chromosome", "start", "end", "strand", "intr
data_lowest <- splice.gene.mapping(top_transcripts_testis_lowest)
data_lowest
```

A data.frame: 5 × 5

	chromosome <chr>	start <int>	end <int>	total_reads <int>	gene_id <chr>
1	2L	123795	123855	4	FBgn0031228
2	2L	203892	203992	2	FBgn0031231
3	2L	547901	547977	3	FBgn0005660
4	2L	815422	815474	6	FBgn0031281
5	2L	852443	852498	14	FBgn0031287

```
merged_lowest <- data.frame(gene_id = data_lowest$gene_id, chromosome = data_lowest$chromo

    suppressWarnings({
    # Iterate through each splice junction
    for (i in 1:nrow(merged_lowest)) {
      # Find the gene that contains the splice junction
      ovary_reads <- (ovary_rows_l$V7[ovary_rows_l$V1 == merged_lowest$chromosome[i] &
                            ovary_rows_l$V2 == merged_lowest$start[i] &
                            ovary_rows_l$V3 == merged_lowest$end[i]] +
                            ovary_rows_l$V8[ovary_rows_l$V1 == merged_lowest$chromosome[i]
                            ovary_rows_l$V2 == merged_lowest$start[i] &
                            ovary_rows_l$V3 == merged_lowest$end[i]])
```

```
     logfold_change <- (top_transcripts_testis_lowest$logfold_change[top_transcripts_testi
                           top_transcripts_testis_lowest$start == merged_lowest$start[i]
                           top_transcripts_testis_lowest$end == merged_lowest$end[i]])

        merged_lowest$ovary_reads[i] <- ovary_reads
        merged_lowest$logfold_change[i] <- logfold_change
    }})

merged_highest <- data.frame(gene_id = data_highest$gene_id, chromosome = data_highest$chr

    suppressWarnings({
    # Iterate through each splice junction
    for (i in 1:nrow(merged_highest)) {
      # Find the gene that contains the splice junction
      ovary_reads <- (ovary_rows_h$V7[ovary_rows_h$V1 == merged_highest$chromosome[i] &
                           ovary_rows_h$V2 == merged_highest$start[i] &
                           ovary_rows_h$V3 == merged_highest$end[i]] +
                           ovary_rows_h$V8[ovary_rows_h$V1 == merged_highest$chromosome[i
                           ovary_rows_h$V2 == merged_highest$start[i] &
                           ovary_rows_h$V3 == merged_highest$end[i]])
     logfold_change <- (top_transcripts_testis_highest$logfold_change[top_transcripts_test
                           top_transcripts_testis_highest$start == merged_highest$start[i
                           top_transcripts_testis_highest$end == merged_highest$end[i]])

        merged_highest$ovary_reads[i] <- ovary_reads
        merged_highest$logfold_change[i] <- logfold_change
    }})

merged_highest
merged_lowest
```

A data.frame: 5 × 7

gene_id <chr>	chromosome <chr>	start <int>	end <int>	testis_reads <int>	ovary_reads <dbl>	logfold_change <dbl>
FBgn00360913	L	10639857	10639919	21903	8	11.248982
FBgn02677262	L	19183339	19183391	4678	1	11.191985
FBgn00378883	R	11240813	11240865	2641	2	9.782452
FBgn00228932	L	21628210	21629666	1	1471	9.523562
FBgn00056172	L	18687127	18687186	1	1279	9.321928

A data.frame: 5 × 7

gene_id <chr>	chromosome <chr>	start <int>	end <int>	testis_reads <int>	ovary_reads <dbl>	logfold_change <dbl>
FBgn0031228	2L	123795	123855	4	4	0
FBgn0031231	2L	203892	203992	2	2	0
FBgn0005660	2L	547901	547977	3	3	0
FBgn0031281	2L	815422	815474	6	6	0
FBgn0031287	2L	852443	852498	14	14	0

Between Replicates 2

```
testis_merged <- testis_replicate2 %>% semi_join(ovary_replicate2, by = c("V2", "V3"))
ovary_merged <- ovary_replicate2 %>% semi_join(testis_replicate2, by = c("V2", "V3"))

# Calculate the average transcript abundance for ovary and testis
avg_abundance_ovary <- ovary_merged[,7]+ovary_merged[,8]
avg_abundance_testis <- testis_merged[,7]+testis_merged[,8]

# Calculate the log fold change (r_g)
log_fold_change <- abs(log2((1 + avg_abundance_ovary) / (1 + avg_abundance_testis)))
```

```
testis_merged$log_fold_change <- log_fold_change[1:nrow(testis_merged)]
top_transcripts_testis_highest <- testis_merged %>% arrange(desc(log_fold_change)) %>% hea
top_transcripts_testis_lowest <- testis_merged %>% arrange(log_fold_change) %>% head(5)
```

```
top_transcripts_testis_highest
top_transcripts_testis_lowest
```

A data.frame: 5 × 10

	V1 <chr>	V2 <int>	V3 <int>	V4 <int>	V5 <int>	V6 <int>	V7 <int>	V8 <int>	V9 <int>	log_fold_change <dbl>
1	3L	10639857	10639919	1	3	1	15126	5	50	10.184875
2	3L	21353014	21353108	1	1	1	1996	3	50	9.965784
3	2L	19183339	19183391	2	2	1	4758	2	50	9.632086
4	3L	9107376	9107618	2	2	1	2165	0	34	9.495855
5	3R	11225650	11225702	1	1	1	455	1364	50	9.244760

A data.frame: 5 × 10

	V1 <chr>	V2 <int>	V3 <int>	V4 <int>	V5 <int>	V6 <int>	V7 <int>	V8 <int>	V9 <int>	log_fold_change <dbl>
1	2L	119077	119133	2	2	1	1	0	49	0
2	2L	141341	141395	2	2	1	1	0	29	0
3	2L	186856	186909	1	1	1	1	0	21	0
4	2L	542347	542429	2	2	1	8	0	50	0
5	2L	542380	542429	2	2	1	1	0	17	0

```
ovary_rows_h <- ovary_merged[ovary_merged$V2 %in% top_transcripts_testis_highest$V2, ]
ovary_rows_l <- ovary_merged[ovary_merged$V2 %in% top_transcripts_testis_lowest$V2, ]
```

```
colnames(top_transcripts_testis_highest) <- c("chromosome", "start", "end", "strand", "int
data_highest <- splice.gene.mapping(top_transcripts_testis_highest)
data_highest
```

A data.frame: 5 × 5

	chromosome <chr>	start <int>	end <int>	total_reads <int>	gene_id <chr>
1	3L	10639857	10639919	15131	FBgn0036091
2	3L	21353014	21353108	1999	FBgn0052436
3	2L	19183339	19183391	4760	FBgn0267726
4	3L	9107376	9107618	2165	FBgn0011206
5	3R	11225650	11225702	1819	FBgn0037879

```
colnames(top_transcripts_testis_lowest) <- c("chromosome", "start", "end", "strand", "intr
data_lowest <- splice.gene.mapping(top_transcripts_testis_lowest)
data_lowest
```

A data.frame: 5 × 5

	chromosome <chr>	start <int>	end <int>	total_reads <int>	gene_id <chr>
1	2L	119077	119133	1	FBgn0031228
2	2L	141341	141395	1	FBgn0031228
3	2L	186856	186909	1	FBgn0016977
4	2L	542347	542429	8	FBgn0010602
5	2L	542380	542429	1	FBgn0010602

```
merged_lowest_2 <- data.frame(gene_id = data_lowest$gene_id, chromosome = data_lowest$chro

    suppressWarnings({
    # Iterate through each splice junction
    for (i in 1:nrow(merged_lowest_2)) {
      # Find the gene that contains the splice junction
      ovary_reads <- (ovary_rows_l$V7[ovary_rows_l$V1 == merged_lowest_2$chromosome[i] &
                            ovary_rows_l$V2 == merged_lowest_2$start[i] &
                            ovary_rows_l$V3 == merged_lowest_2$end[i]] +
                            ovary_rows_l$V8[ovary_rows_l$V1 == merged_lowest_2$chromosome[
                            ovary_rows_l$V2 == merged_lowest_2$start[i] &
                            ovary_rows_l$V3 == merged_lowest_2$end[i]])
     logfold_change <- (top_transcripts_testis_lowest$logfold_change[top_transcripts_testi
                            top_transcripts_testis_lowest$start == merged_lowest_2$start[i
                            top_transcripts_testis_lowest$end == merged_lowest_2$end[i]])

        merged_lowest_2$ovary_reads[i] <- ovary_reads
        merged_lowest_2$logfold_change[i] <- logfold_change
    }})

merged_highest <- data.frame(gene_id = data_highest$gene_id, chromosome = data_highest$chr

    suppressWarnings({
    # Iterate through each splice junction
    for (i in 1:nrow(merged_highest)) {
      # Find the gene that contains the splice junction
      ovary_reads <- (ovary_rows_h$V7[ovary_rows_h$V1 == merged_highest$chromosome[i] &
                            ovary_rows_h$V2 == merged_highest$start[i] &
                            ovary_rows_h$V3 == merged_highest$end[i]] +
                            ovary_rows_h$V8[ovary_rows_h$V1 == merged_highest$chromosome[i
                            ovary_rows_h$V2 == merged_highest$start[i] &
                            ovary_rows_h$V3 == merged_highest$end[i]])
     logfold_change <- (top_transcripts_testis_highest$logfold_change[top_transcripts_test
                            top_transcripts_testis_highest$start == merged_highest$start[i
                            top_transcripts_testis_highest$end == merged_highest$end[i]])

        merged_highest$ovary_reads[i] <- ovary_reads
        merged_highest$logfold_change[i] <- logfold_change
    }})

merged_highest
```

```
merged_lowest_2
```

A data.frame: 5 × 7

gene_id <chr>	chromosome <chr>	start <int>	end <int>	testis_reads <int>	ovary_reads <dbl>	logfold_change <dbl>
FBgn0036091	3L	10639857	10639919	15131	12	10.184875
FBgn0052436	3L	21353014	21353108	1999	1	9.965784
FBgn0267726	2L	19183339	19183391	4760	5	9.632086
FBgn0011206	3L	9107376	9107618	2165	2	9.495855
FBgn0037879	3R	11225650	11225702	1819	2	9.244760

A data.frame: 5 × 7

gene_id <chr>	chromosome <chr>	start <int>	end <int>	testis_reads <int>	ovary_reads <dbl>	logfold_change <dbl>
FBgn0031228	2L	119077	119133	1	1	0
FBgn0031228	2L	141341	141395	1	1	0
FBgn0016977	2L	186856	186909	1	1	0
FBgn0010602	2L	542347	542429	8	8	0
FBgn0010602	2L	542380	542429	1	1	0

It's not logically correct to consider splicing junctions as being alternatively spliced solely based on log fold change values. Log fold change measures the difference in gene expression between two conditions, in this case, between ovary and testis. While a high log fold change indicates a substantial difference in expression, it doesn't directly imply alternative splicing of the same gene.

Alternative splicing refers to the process by which different exons of a gene can be included or excluded in the final mRNA transcript, leading to the generation of multiple transcript isoforms from a single gene. To identify alternative splicing events, we have to look further at differences in the exon composition of transcripts or the presence of different splice junctions within the same gene.

CS509 - Extra Credit Task E2 Polyadenylation (50%)
(Modified to Visualize Splice Junctions)

Indronil Bhattacharjee & Erica Flores

Introduction

Integrated Genomics Viewer (IGV) was used to visualize read coverage and splice junctions of RNA-seq reads from fruit fly ovary and testits. The bam and index files were uploaded into IGV, along with the reference genome for *Drosophila melanogaster*.
Below are the outputs from the R code from task E1, a summary of the highest fold-change and lowest-fold genes in terms of reads across splice junctions (Table 1 and Table 2).

Results

Table 1 - Highest-fold change

gene_id	chromosome	start	end	testis_reads	ovary_reads	logfold_change
<chr>	<chr>	<int>	<int>	<int>	<dbl>	<dbl>
FBgn0036091	3L	10639857	10639919	21903	8	11.248982
FBgn0267726	2L	19183339	19183391	4678	1	11.191985
FBgn0037888	3R	11240813	11240865	2641	2	9.782452
FBgn0022893	2L	21628210	21629666	1	1471	9.523562
FBgn0005617	2L	18687127	18687186	1	1279	9.321928

Table 2 - Lowest-fold change

gene_id	chromosome	start	end	testis_reads	ovary_reads	logfold_change
<chr>	<chr>	<int>	<int>	<int>	<dbl>	<dbl>
FBgn0031228	2L	123795	123855	4	4	0
FBgn0031231	2L	203892	203992	2	2	0
FBgn0005660	2L	547901	547977	3	3	0
FBgn0031281	2L	815422	815474	6	6	0
FBgn0031287	2L	852443	852498	14	14	0

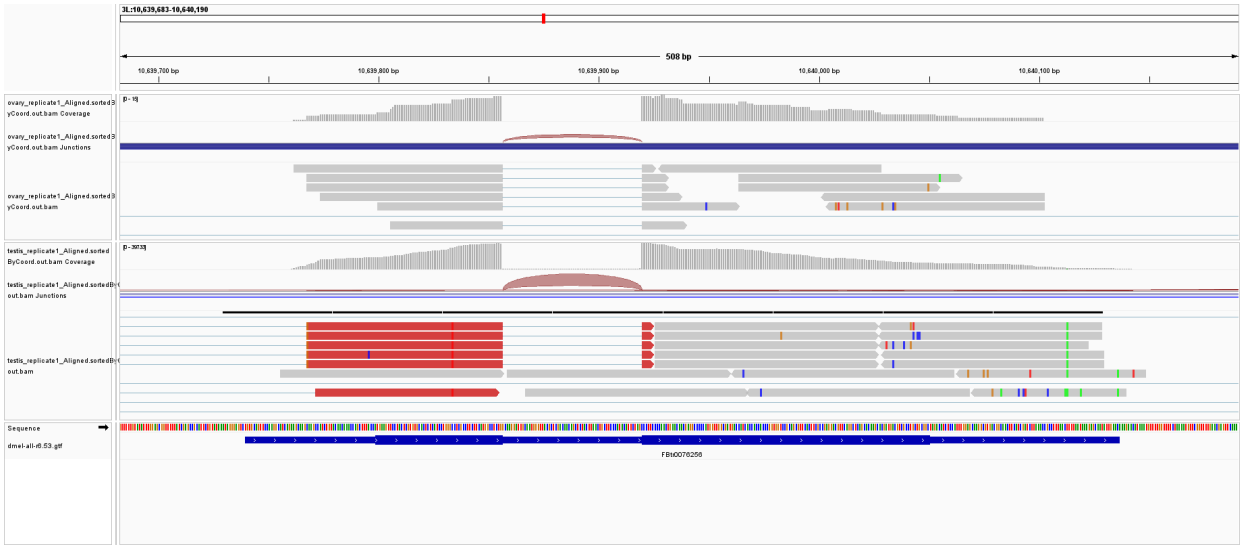
From here, we decided to focus on those with the highest fold change, and a table was created to summarize the gene ID with the corresponding transcript ID, as well as the biological function (Table 3).

Table 3. Transcripts Corresponding to Top Genes and Biological Function

Highest		
Gene	**Transcript**	**Function**
FBgn0036091	FBtr0076256	Involved in sexual reproduction. Located in extracellular space. Is expressed in spermatheca and testis. (Alliance, FBgn0036091)
FBgn0267726	FBtr0347276	The **biological processes** in which it is involved are not known BLAST Hit for *Drosophila mauritiana* male-specific sperm protein Mst84Dd
FBgn0037888	FBtr0082362	**scpr-B** involved_in multicellular organism reproduction Predicted to be located in extracellular region. Orthologous to human R3HDML (R3H domain containing like) and PI15 (peptidase inhibitor 15). (Alliance, FBgn0037888)
FBgn0022893	FBtr0100293	**Decondensation factor 31 (Df31)** encodes a histone binding protein involved in nucleosome assembly. [Date last reviewed: 2019-09-12] (FlyBase Gene Snapshot)
FBgn0005617	FBtr0081130	**male-specific lethal 1** male-specific lethal 1 (msl-1) encodes a protein that is thought to form a scaffold to organize the full male-specific-lethal dosage compensation complex, which increases male X chromosome transcription approximately two-fold. msl-1 homozygous mutant males die as larvae, while females are viable. [Date last reviewed: 2019-03-14] (FlyBase Gene Snapshot)

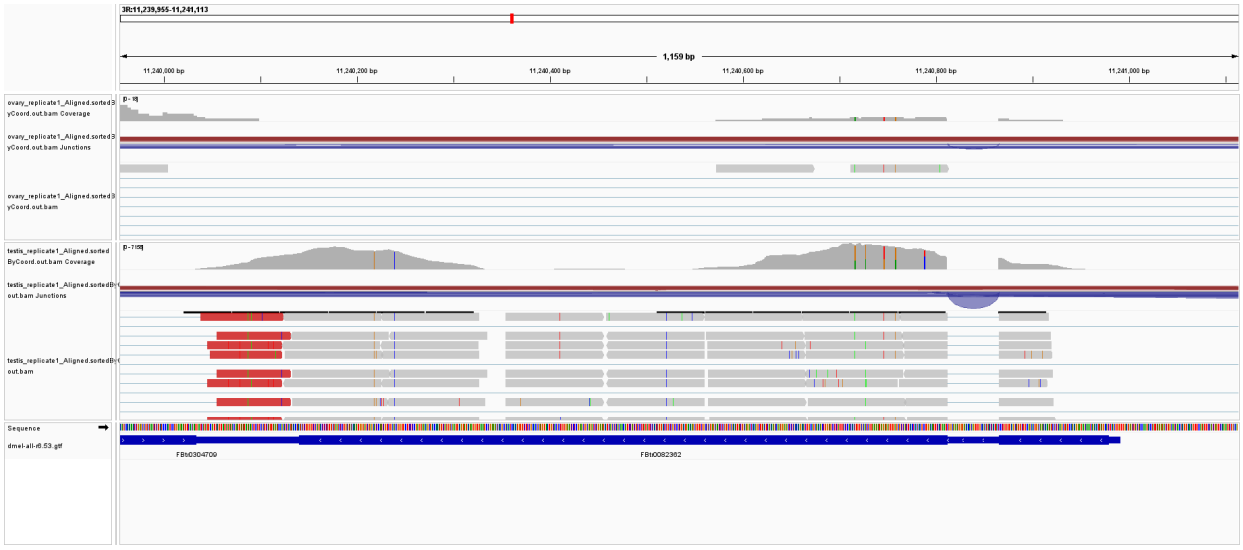
For the visualization, only transcript replicate 1 was used for the testis and ovary. Images exported from the IGV viewer are only captured from what is visible in the window at the time, so only 1 replicate from each tissue was used. In the following images, the ovary transcript is on top, with the testis sample below that, and the reference genome at the bottom. Gray histogram bars indicate the read coverage and the arcs of color (red for the positive strand and blue for the negative strand) show splice junctions. The height and width of the arcs indicate the number of reads across the junction. For each sample below the splice junction arcs, there is a small proportion of the splice variants found visible, however, this is a large scrolling window and we were not able to capture all of the variants.

FBgn0036091-FBtr0076256

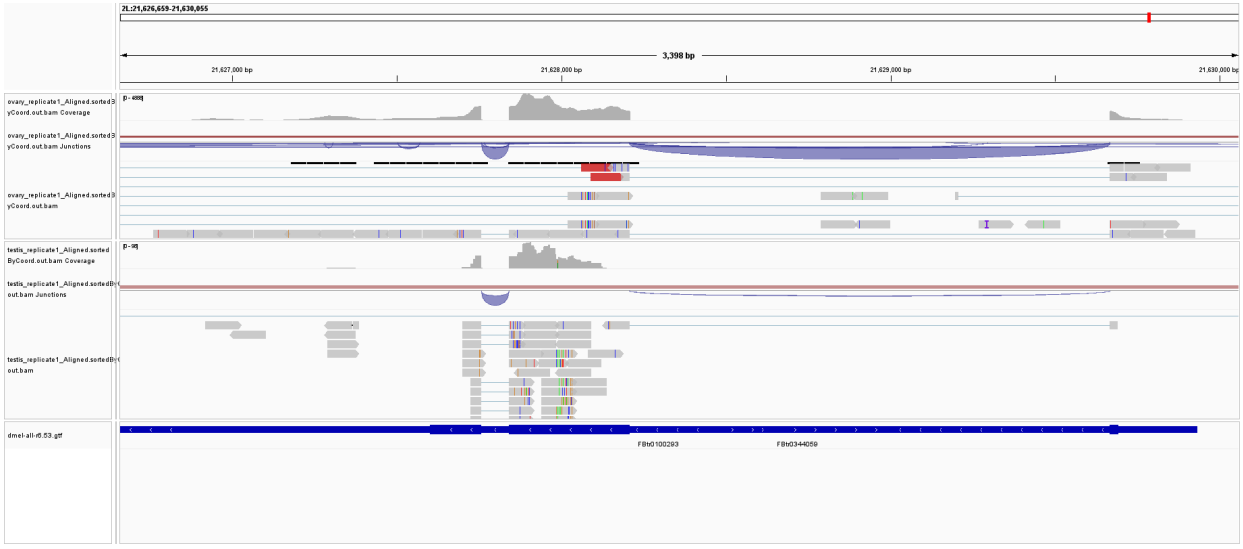


FBgn0267726-FBtr0347276

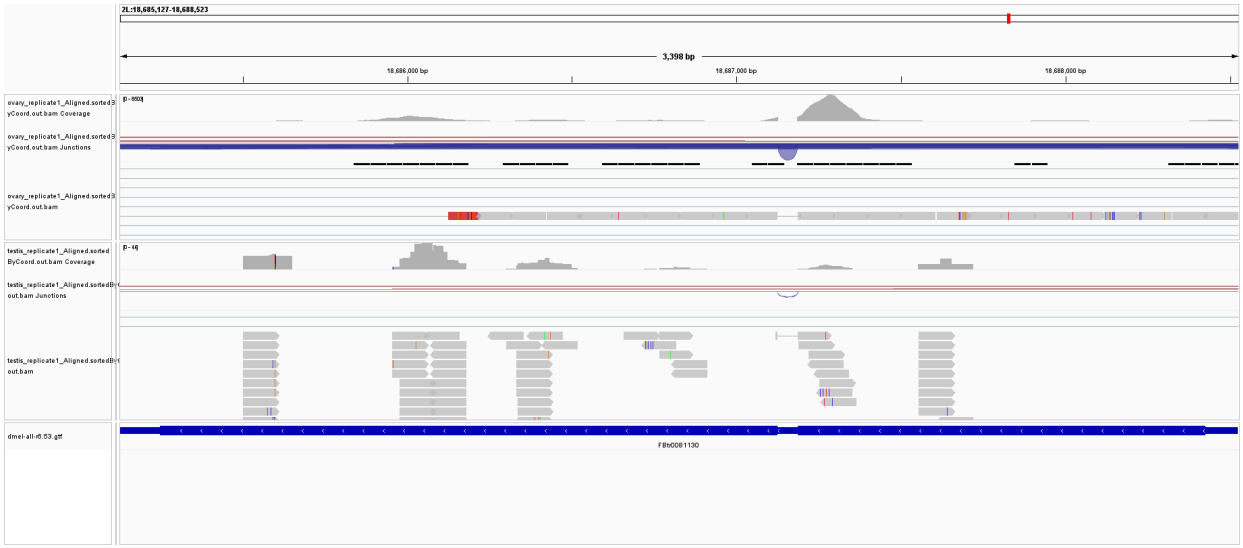
Fbgn0037888-FBtr0082362



FBgn0022893-FBtr0100293



FBgn0005617-FBtr0081130



Out of the top 5 highest fold-changes in splice junction reads, the top 3 highest genes were associated with male reproduction(FBgn0036091, FBgn0267726, FBgn0037888), while the other two were highest in females and associated with histone modification (FBgn0022893) or associated with transcription (FBgn0005617).

When FBgn0267726 was searched for in FlyBase, the annotation said ‘The **biological processes** in which it is involved are not known’ which made interpretation difficult. In order to determine what biological processes this transcript is involved in, the sequence was entered into NCBI BLAST (Basic Local Alignment Search Tool) which will determine if the sequence matches any other known sequences in the database. While most of the hits matched D. melanogaster as unknown matches again, there was a match between our sequence and another fly species, *Drosophila mauritiana* male-specific sperm protein Mst84Dd. Images below show the top hits from BLAST and the specific alignment of our query sequence and Mst8Dd. The match shows 92.82% identity match, an e value of 6e-166 and 100% coverage.

Top Blast Hits for **FBgn0267726**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Drosophila melanogaster uncharacterized protein, transcript variant B (CG46059), mRNA	Drosophila mel...	756	756	100%	0.0	100.00%	409	NM_001316400.1
Drosophila melanogaster uncharacterized protein, transcript variant C (CG46059), mRNA	Drosophila mel...	695	695	100%	0.0	96.92%	596	NM_001316401.1
Drosophila melanogaster uncharacterized protein, transcript variant A (CG46059), mRNA	Drosophila mel...	689	689	96%	0.0	97.78%	405	NM_001316399.1
Drosophila melanogaster GH05530 full insert cDNA	Drosophila mel...	682	682	95%	0.0	97.76%	424	AY118741.1
PREDICTED: Drosophila mauritiana male-specific sperm protein Mst84Dd (LOC117148076), mRNA	Drosophila mau...	597	597	100%	6e-166	92.82%	435	XM_033315287.1
PREDICTED: Drosophila sechellia male-specific sperm protein Mst84Dd (LOC6614659), mRNA	Drosophila sec...	590	590	96%	1e-163	93.33%	451	XM_002039050.2
PREDICTED: Drosophila teissieri male-specific sperm protein Mst84Dd (LOC122612897), transcript var...	Drosophila teis...	468	468	99%	5e-127	87.50%	412	XM_043786768.1
Drosophila melanogaster isolate dmeA_05_F0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121943.1
Drosophila melanogaster isolate dmeA_01_M0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121931.1
Drosophila melanogaster isolate dmeA_05_F0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121937.1
Drosophila melanogaster isolate dmeA_05_M0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121949.1
Drosophila melanogaster isolate dmeA_15_F0 chromosome 2L	Drosophila mel...	455	753	100%	4e-123	100.00%	23513712	CP121960.1
Drosophila melanogaster isolate dmeA_15_M0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121972.1
Drosophila melanogaster isolate dmeA_18_F0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121995.1
Drosophila melanogaster isolate dmeA_18_M0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP122001.1
Drosophila melanogaster isolate dmeA_18_F0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP121984.1
Drosophila melanogaster isolate dmeE_27_M0 chromosome 2L	Drosophila mel...	455	758	100%	4e-123	100.00%	23513712	CP122073.1

Alignment between “unknown” **FBgn0267726** and *Drosophila mauritiana* male-specific sperm protein Mst84Dd

PREDICTED: Drosophila mauritiana male-specific sperm protein Mst84Dd (LOC117148076), mRNA

Sequence ID: XM_033315287.1 **Length:** 435 **Number of Matches:** 1

Range 1: 9 to 425 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
597 bits(323)	6e-166	388/418(93%)	10/418(2%)	Plus/Plus

```
Query  1    CAAAATTCGATGTGTTAACACCTTTGAAAATAGTTTGTGTCTTACGTTTTCCGGTACCTA  60
            ||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||
Sbjct  9    CAAAATTCGATGTGTTAACACCTTTGGAAATAGTTTGTGTCTTACGTTTTCCGGTACCTA  68

Query  61   TTTTGTTGTTTCTAAGATCCTATTTCTGTAAATAGGCC-AA----AT----CTGCCCTTT  111
            ||||||||||||||| ||||||||||||||||||||| ||    ||    |||||||||
Sbjct  69   TTTTGTTGTTTCTAATATCCTATTTCTGTAAATAGGCCGAATTCCATTTAGCTGCCCTTT  128

Query  112  TAGAAGGAAATTAGGATACCCTGAACGCTTTTCCCCAAGACAAAATAAAATCATCATGTG  171
            ||||||||||||||| |||||||||||| ||||| |||||| |||| ||||||||||||||
Sbjct  129  TAGAAGGAAATTAGCATACCCTGAACACTTTT-CCCAAGGCAAACTAAAATCATCATGTG  187

Query  172  CTGCGGACCCTGTGGACCTCGCTGCTGCGATCCGTGCGGCGGATGCTACAACTGCTGCGT  231
             |||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
Sbjct  188  TTGCGGACCCTGTGGACCTCGCTGCTGCGATCCCTGCGGCGGATGCTACAACTGCTGCGT  247

Query  232  GGAACTCTGCTGTGTACCCTGCACCCCAGCCTACATCCAGTGCTCATTTATGCCCTGCGG  291
            |||||||||||||| ||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  248  GGAACTCTGCTGTGTGCCCTGCACCCCAGCCTACATCCAGTGCTCCTTTATGCCCTGCGG  307

Query  292  ACCAAGAGGCTGTTGCTGAAGTGGGGATGTGCCAGGTGCCGAAACACGTTCAACCATATT  351
             |||||||||||||||||||| ||  | ||||||||||||| || |||| |||||
Sbjct  308  TCCAAGAGGCTGTTGCTGAAGTGGGAATAAGTCAGGTGCCGAAACAAGTCCAACAATATT  367

Query  352  GTACCTGAAACACTCGTAGATACCCAACATGTCCCAATAAACGAATTTTATAAATGTT  409
            |||||||||||||||||||| |||||||||||||||||||||||| |||||||||
Sbjct  368  GTACCTGAAACACTCGTAGATACACAACATGTCCCAATAAACGAATTTAATAAATGTT  425
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

Discussion

Alternative splicing (AS) is a feature of metazoan genomes that allows for the production of a wide variety of transcripts from a single gene. These transcripts vary in their expression across tissues, as well as in a temporal dimension, and are critical to cell differentiation and regulation. Reproductive tissues (testis) and the brain, represent highly specialized organs that have been found to have high levels of AS across many species, including mammals (Naro et al 2021).

As spermatogenesis is a highly complex process and occurs throughout the lifetime of the fly, it requires constant regulation. Some studies have shown how important AS is to the genetic diversity of spermatozoa and subsequent embryonic viability (Song et al 2020). We actually saw one of these viability genes from our study, FBgn0005617, male-specific lethal 1 (msl-1) which is homozygous lethal in males (death at larval stage) but is viable in females.

References

Naro C, Cesari E, Sette C. Splicing regulation in brain and testis: common themes for highly specialized organs. Cell Cycle. 2021 Mar-Mar;20(5-6):480-489. doi: 10.1080/15384101.2021.1889187. Epub 2021 Feb 26. PMID: 33632061; PMCID: PMC8018374.

Song H, Wang L, Chen D, Li F. The Function of Pre-mRNA Alternative Splicing in Mammal Spermatogenesis. Int J Biol Sci. 2020 Jan 1;16(1):38-48. doi: 10.7150/ijbs.34422. PMID: 31892844; PMCID: PMC6930371.