Uncovering the Role of RNA Helicase Vasa in Alternative Splicing in Mouse Oocytes Using Computational Methods

Yoonah Lee

Thesis Submitted in Partial Fulfillment of the Requirements of Bachelor of Science

with Honors in Biology

First Reader: Mamiko Yajima, PhD

Second Reader: Ashok Ragavendran, PhD

Brown University

April 21, 2023

Table of Contents

Abstract

DEAD-box RNA proteins are essential in cellular RNA metabolism, ranging from transcription to translation regulation. Among the DDX helicases, DDX4 is particularly known to be critical in cell-cycle progression and germline development. Alternative splicing is critical for the regulation of generating mRNAs and proteins from one gene, and there are limited studies investigating the relationship between DDX helicases and alternative splicing, let alone DDX4. This study aims to identify candidate genes that might be involved in splicing and to uncover the role and pathways of mouse Vasa homolog (MVH) as an alternative splicing regulator in mouse oocytes using various computational methods. Indeed, 130 genes were identified by MISO to have differential isoform expression in MVH-knockout samples compared to the wildtype, while 461 exon-skipping genes were identified using DEXSeq. The spliceosome pathway was enriched for differentially expressed genes and exon-skipping genes, while RNA and mRNA splicing were enriched for differentially expressed genes and differentially spliced genes. Most notably, GEMIN7, the gene encoding a protein required for pre-mRNA splicing, was significantly downregulated, had differentially expressed isoforms, and was exon-skipping, indicating that MVH regulates GEMIN7 expression and closely interacts with GEMIN7. These results strongly suggest that MVH regulates splicing both directly and indirectly by promoting differential isoform expression of genes involved in RNA splicing and by differentially expressing genes that are involved in the spliceosome.

Introduction

Alternative splicing is a fundamental process that greatly expands the coding capacity of the genome and contributes to the functional diversity of proteins by forming different isoforms.

It is a key regulatory mechanism in eukaryotic gene expression and plays a crucial role in various biological processes, such as development, differentiation, and disease, most notably cancer cell proliferation (Nilsen & Graveley, 2010). Dysregulation of alternative splicing is a hallmark of cancer and is often associated with tumor development and progression (David & Manley, 2010). Mutations that inactivate splice sites are especially known to be good indicators for hereditary cancer genes (Bonnal, López-Oreja, & Valcárcel, 2020). Different cancer types and subtypes have different alternative splicing that can be used as a cancer hallmark. Analyses of 8705 tumor patients across 32 cancer types revealed 30% more alternative splicing events in tumors than in normal samples, meaning splicing might take a role in generating cancer markers and neoantigens (Kahles et al., 2018) that could be used in the development of mRNA vaccines (Bonnal et al., 2020). In oocytes, a disruption in alternative splicing can compromise oogenesis and fertility. Deletion of breast carcinoma amplified sequence 2 (*Bcas2*) in mouse oocytes caused 991 alternative splicing events to be significantly different from the wildtype, which affected 706 genes, including those involved in oogenesis and spindle assembly, and compromised mouse oocyte development (J. Zhang et al., 2022). Alterations in alternative splicing events can also directly impact fertility: female mice with the short isoform of the enhancer of zeste homologue 2 (*Ezh2*), a gene involved in polycomb repressive complex 2 (PRC2), had significantly decreased fertility and mitochondrial function (Guo et al., 2022). Interestingly, proteomic expression profiles of aging oocytes revealed that the differentially expressed proteins were enriched in RNA splicing and that changes in protein expression in aging oocytes from abnormal alternative splicing regulation significantly contributed to impaired oocyte developmental potential (M. Li et al., 2021). Thus, alternative splicing directly and indirectly regulates oocyte development, and dysfunctional alternative splicing events can compromise oogenesis and even be one of the causes of declined developmental potential in aging oocytes.

DEAD-box RNA (DDX) proteins consist of highly conserved core domains and the N- and C- terminals. The core domains are responsible for the regulation of substrate binding and ATPase and helicase unwinding activities, while the function of the terminal regions remains unclear. DDX helicases are essential in cellular RNA metabolism, ranging from transcriptional regulation to translation.

Among the DDX helicases, recent studies have discovered DDX41's role in mRNA splicing. DDX41 regulated alternative splicing of genes involved in cancer pathways, such as EGFR and FGFR pathways, in HeLa cells (Qin et al., 2021), while cells deficient of DDX41 in mice exhibited intron excision in the processing of snoRNAs (Chlon et al., 2021). In the nematode *Caenorhabditis elegans*, the depletion of DDX41 ortholog *sacy-1* resulted in significant alternative splicing, more in the germline-depleted samples than soma-depleted samples, while *sacy-1/ddx41* missense mutations exhibited antimorphic activity, indicating that these mutations would be destructive to spliceosomes (Tsukamoto et al., 2020). Recently, DDX10 was also shown to regulate the development of colorectal cancer in mice by alternatively splicing RPL35 mRNA: the expression levels of mRNA isoforms of RPL35 changed in DDX10-knockdown samples, while DDX10-overexpressed samples showed reduced levels of RPL35 mRNA (Zhou et al., 2022).

DDX4 (the human ortholog of *Vasa*) is an ATP-dependent helicase and is implicated in translational regulation. It has nine DDX-conserved sequence motifs: six of these motifs are in Domain I (DEADc), while the rest are in Domain II (HELICc) (Hickford, Frankenberg, Pask, Shaw, & Renfree, 2011). DDX4 has been found in the germlines of many animal taxa, such as mammals, sponges, and flatworms (Hickford et al., 2011) as well as outside of the germline in the embryo and adult of some animals (Schudrowitz, Takagi, Wessel, & Yajima, 2017). However, DDX4 plays different roles in germline development, gametogenesis, and fertility across a wide range of species. *Vasa* gene is required for abdominal segments and germ cell specification in *Drosophila* (Schüpbach & Wieschaus, 1986). Knockdown of DDX4 causes female *Drosophila* to lack germ cells, while the knockdown does not affect males (Schüpbach & Wieschaus, 1986). In *Caenorhabditis elegans*, germ cells were compromised in the offspring of knockdown of *glh* (DDX4 homologue in *C. elegans*) (Kuznicki et al., 2000). On the other hand, in *Macrostomum lignano* and *Danio rerio*, the loss of DDX4 did not affect germ cells (Braat, van de Water, Korving, & Zivkovic, 2001; Pfister et al., 2008).

Among mammals, DDX4 is mostly studied in the mouse. The loss of DDX4 has a sex-specific effect in mice like *Drosophila* but in the opposite direction. The mouse Vasa homolog (MVH) is crucial for male germ cell development, with its deficiency leading to defects in spermatogenesis and infertility (Tanaka et al., 2000). Further, DDX4 has long been known to interact with other proteins, notably the PIWI family member MILI, in the regulation of germ cell development and function (Kuramochi-Miyagawa et al., 2004). MVH is essential for MILI slicing-triggered piRNA biogenesis (Wenda et al., 2017), while MIWI deficiency affects sub-cellular localization of MVH (Thomson & Lin, 2009). This interaction underscores the complexity of the molecular and biological mechanisms that involve DDX4 and points to the importance of investigating VASA/DDX4 in the context of a broader regulatory network.

Recently, the host lab has found that spliceosomes that regulate mRNA splicing may be downregulated in MVH-knockdown datasets of multiple organisms (Yajima lab, unpublished). Although DEAD-box proteins, including DDX4, have been implicated in diverse aspects of RNA metabolism, such as RNA folding, stability, and translation (Linder & Jankowsky, 2011), the potential role of DDX4 in alternative splicing remains largely unexplored. Using the mouse *Vasa* (MVH) knockout dataset, one of the datasets available in the lab, in this project, I address three objectives: 1) identify candidate genes that undergo and/or participate in differential alternative splicing in MVH-knockout samples, 2) identify downstream pathways in which VASA regulates, and 3) conduct a comparative analysis with the glioblastoma stem cells in humans based on genes and pathways. If VASA/DDX4 is shown to play a role in alternative splicing, it will facilitate a deeper understanding of the greater network of VASA/DDX4 in mice and potentially humans.

Materials and Methods

RNA-seq pre-processing and mapping

RNA-seq data of two wild type (WT) and two MVH-knockout (KO) replicates were provided by the collaborator Dr. Azusa Inoue at RIKEN in Japan as fastq files. All fastq files were quality-checked using FastQC (Andrews, 2010) and pre-processed and trimmed using Trimmomatic (Bolger, 2014), primarily to remove the adapter content. The trimmed alignments were mapped to the reference genome *mm10* (mouse reference genome NCBI build 38, GRCm38) with STAR (Dobin et al., 2013) and sorted by coordinate with samtools (Danecek et al., 2021). Aligned and non-aligned reads were quantified using Qualimap (García-Alcalde et al., 2012). The count matrix of the four output bam files was obtained using featureCounts (Liao, Smyth, & Shi, 2013).

Differential expression gene analysis

With the obtained count matrix from featureCounts, differential expression gene analysis was performed using DESeq2 (Love, Huber, & Anders, 2014). Volcano plots were produced using

EnhancedVolcano (Blighe, 2018). 6 different thresholds were considered to filter significant genes: *p*-value < 0.001 and |log2 fold change (logFC)| > 1.5, *p*-value < 0.05 and |logFC| > 1.5, *p*-value < 0.001, *p*-value < 0.05, false discovery rate (FDR) < 0.1, and FDR < 0.05. Pathway analysis was conducted on the gene list filtered using *p*-value < 0.001 (a total of 498 genes) because the pathway analyses of other thresholds showed no enriched gene ontologies and pathways. Although we could have used genes with *p*-value < 0.05, many of these genes had a low logFC, which means that these genes might not actually be differentially expressed but were only considered statistically significant because of the low read counts. When comparing differentially expressed genes with genes with differential isoform expression (generated by MISO), the gene list filtered using *p*-value < 0.001 and |logFC| > 1.5 (a total of 89 genes) was used. We used this threshold because some genes had a significant *p*-value but a low logFC, in which case the low *p*-value could be attributable to the low counts of those genes. To ensure that the genes we analyze are classified as differentially expressed because of the loss of MVH, not because of the low count, logFC was used to further filter differentially expressed genes. We did not use the FDR for the differential expression gene analysis, because the gene list produced with FDR neither exhibited significant differential expression nor had any enriched pathways when the pathway analysis was conducted.

Differential isoform expression analysis

The trimmed data were processed with mixture-of-isoforms (MISO) model for all genes. MISO is a probabilistic framework that detects differentially regulated isoforms using Bayesian inference (Katz, Wang, Airoldi, & Burge, 2010). A list of genes with corresponding bayes factor (BF), isoforms, and genomic location was generated. BF measures the confidence interval for differential expression, and a higher BF corresponds to a higher likelihood of a differentially expressed isoform between WT and KO replicates. Pathway analysis was conducted with the genes of BF greater than 10. Some genes with a high BF were visualized using Integrative Genomics Viewer (IGV) (Robinson et al., 2011). The genes that were identified to be differentially spliced in humans (produced by Dr. Yusuke Suita, a collaborator in Tapinos Lab, using LeafCutter (Y. I. Li et al., 2018)) were compared with the genes from differential expression analysis and differential isoform expression analysis in our data and visualized using IGV.

Differential exon usage analysis

Using DEXSeq (Anders, Reyes, & Huber, 2012), number of reads for all exons in all genes were obtained. With the number of reads, exon-skipping events were identified with a generalized linear regression model. A pathway analysis was done on the list of exon-skipping genes that were statistically significantly differentially expressed (FDR < 0.05). This gene list was also compared with the differentially expressed gene list as well as the MISO gene list. Transcripts of differentially expressed and differential exon expressed genes were visualized using DEXSeq and IGV.

Pathway analysis

ClusterProfiler (Wu et al., 2021), Gene Ontology (GO) terms (Balakrishnan, Harris, Huntley, Van Auken, & Cherry, 2013), Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), and Kyoto Encyclopedia of Genes and Genomes **(**KEGG) pathway analysis (Du et al., 2016) were performed.

Results

Possibilities for ~35% of mapped reads that were empty

Principal component analysis (PCA) of our RNA-seq data revealed that KO and WT replicates were clustered together, as in the magnitude of PC2 was very small (Figure 1A), confirming that there were no unexpected variations between samples that may skew our results. There were a little less than 35% empty reads (the reads that do not fall in the exonic region) out of all the mapped reads for all four replicates—there were 34,910,682 and 29,454,332 read alignments for WT samples, while there were 16,421,880 and 19,230,885 reads for KO samples (Supplementary Figure 1). The 35% empty reads could be polyadenylated RNAs, unspliced introns, the presence of rRNAs, and alignment reads falling into the intron area or transcript areas that are not annotated yet. There were around 25% intronic and intergenic reads out of all the mapped reads for all samples (Supplementary Figure 1), so 35% was reasonable since our protocol had a polyadenylated RNA enriching step and there were other types of RNAs in our RNA-Seq data. If introns appear to be evenly filled with reads, it could indicate intron retention. On the other hand, if there are small islands of reads within the introns, it could mean that there are exons missing in the annotation. Although these two possibilities were not explored deeper, there were potentially missing annotations as seen in the small islands of reads in an unannotated region for all four replicates (Figure 1C). Intron retention was not found but still remains a possibility.

There were similar numbers of genes being downregulated and upregulated (Figure 1B). In line with the PCA plot, there was little variability found within the experimental groups in the MA plots, since there were no genes with large logFC values that were deemed statistically insignificant, and most genes with big logFC values and expression values were classified as differentially expressed (Figure 1B). Thus, using the *p*-value and logFC value as thresholds to identify differentially expressed genes seemed appropriate.

Cdc5l and the spliceosome pathway are upregulated in MVH-KO samples

There were 49 downregulated genes (1.5-fold downregulation with *p*-value < 0.05), including *Ddx4*, *Gemin7*, *Nubp2*, *Atp5j2*, and *Fis1* (Figures 2A, B). Many of these genes were involved in adenosine triphosphate (ATP) synthesis and mitochondrion morphogenesis. ATP is essential in oocytes as it provides energy for development, fertilization, and embryo formation (D. Zhang, Keilty, Zhang, & Chian, 2017), which may explain the GSEA result in which the suppressed genes were enriched in animal organ development and developmental process in addition to mitochondrion activity (Figure 2D). On the other hand, there were 23 upregulated genes (1.5-fold upregulation with *p*-value < 0.05), including *Cdc5l*, *Zfp937*, *Zfp952*, and *Zdhhc15* (Figures 2A, B). Many of these upregulated genes were associated with zinc finger, many of which functions include DNA recognition, regulation of apoptosis, and RNA packaging (Laity, Lee, & Wright, 2001), which is a potential reason that the activated genes are enriched in DNA-binding activities (Figure 2D). One possible explanation is that the lack of MVH acts as a stressor, promoting biological activities that alter the transcriptome and regulate gene expressions (Deplancke, Alpern, & Gardeux, 2016).

GO enrichment analysis of the differentially expressed genes with *p*-value < 0.001 revealed that RNA splicing was one of the most enriched with a gene ratio of 0.0401, followed by RNA splicing via transesterification reactions and mRNA splicing via the spliceosome (Figure 2C). The genes enriched in RNA splicing were *Snrpd3, Rtraf, Cdc5l, Clns1a, Larp7, Plrg1, Srrm1, Taf12, Frg1, Dnajc17, Aqr, Gemin7, Txnl4a, Snrpg, Phf5a, Fam172a, Ubl5,* and *Scnm1*.

KEGG analysis further revealed that the upregulated genes were associated with the spliceosome pathway with an enrichment score of 0.4444 (Figure 2E), most notably *Cdc5l*

(Supplementary Figure 2), the most differentially expressed gene. This uncovers a new possibility that *Ddx4* might regulate the expression of genes that are involved in the spliceosome, such as *Cdc5l*, thus regulating the spliceosome activity and alternative splicing events via the spliceosome. Indeed, although *Cdc5l* was the most upregulated gene in MVH-KO mice with a p-value of 6.16^{-127}, it was classified as neither differentially spliced by MISO nor to be exon-skipping by DEXSeq. The loss of *Cdc5l* is known to decrease pre-mRNA splicing efficiency of *Sox9* and *Col2a1* and to enhance the splicing of *Wee1* by binding directly to their transcripts in mouse embryo bone cartilage samples (Jokoji et al., 2021). The loss of *Cdc5l* also upregulates *Wee1* gene expression but downregulates *Sox9* and *Col2a1* (Jokoji et al., 2021). However, in oocytes, *Sox9* gene expression is restricted to anamniotes clade (Penrad-Mobayed et al., 2018). In our data, the expression levels of *Sox9*, *Col2A1*, and *Wee1* were too low (close to zero). This is in line with previous studies for *Sox9* and *Wee1*, and the gene expression of *Col2a1* in oocytes could also be limited to fish and amphibians like *Sox9*. Overall, the loss of MVH upregulated genes involved in the spliceosome, including *Cdc51*, meaning that MVH likely plays an indirect role in alternative splicing.

All genes in the mRNA splicing pathway via the spliceosome, including Gemin7, are downregulated and differentially spliced in MVH-KO samples

MISO detected 428 genes with a BF > 10, meaning that these genes were at least 10 times more likely to be differentially spliced (the isoform is at least 10 times more likely to be differentially expressed between treatments) in WT and KO samples. GSEA analysis of the 428 genes identified by MISO revealed that RNA splicing via transesterification reactions and mRNA splicing via spliceosome were one of the most over-represented for suppressed genes (Figure 3A, B). This means that the genes that were differentially spliced and are involved in splicing were more suppressed in KO than WT replicates. This raises a hypothesis that *Ddx4* directly regulates alternative splicing efficiency of the genes identified by MISO that in turn regulates the expression levels of those genes.

To explore this possibility, we identified differentially expressed genes that also had isoforms that were significantly differentially expressed by using MISO. First, we investigated the genes (*Snrpd3, Rtraf, Cdc5l, Clns1a, Larp7, Plrg1, Srrm1, Taf12, Frg1, Dnajc17, Aqr, Gemin7, Txnl4a, Snrpg, Phf5a, Fam172a, Ubl5,* and *Scnm1*) identified as RNA splicing in the GO enrichment analysis. Among these 18 genes, *Gemin7* (variant 2's BF = 1235292364.89, variant 6's BF = 1000000000000.00, and logFC = -2.14649), *Dnajc17* (variant 1's BF = 1707.22, variant 3's BF = 85770.97, and logFC = -1.29355), and *Scnm1* (variant 1's BF = 249.80, and logFC = -1.30347) had differentially expressed isoforms. It should be noted that the genes with differentially expressed isoforms were all downregulated, which raises the hypothesis that the loss of MVH decreases the alternative splicing efficiency in these genes and thereby decreases their expression levels, suggesting that MVH may play a direct role in splicing.

Next, we used a more stringent threshold (*p*-value < 0.001 and |logFC| < 1.5) to identify genes that were both differentially expressed and spliced. Out of the 428 genes, only *Gemin7* and *Gm14444* were differentially expressed with this threshold (Supplementary Table 1). *Gemin7* had a BF of 1235292364.89 for the second variant and 1000000000000.00 for the sixth variant (Figure 3E), which was identified as exon-skipping by DEXSeq (Figure 4D). *Gemin7* is a component of the survival of motor neurons (SMN) complex, which structures spliceosomal snRNPs as well as splices and transcribes pre-mRNA, and likely acts upstream or within mRNA processing (Baccon, Pellizzoni, Rappsilber, Mann, & Dreyfuss, 2002). Genes involved in the mRNA splicing GO with

Gemin7 were *Nono*, *Npm1*, *Nsrp1*, and *Scnm1* (Figure 3C), which all had an isoform that was more than 10 times likely to be differentially expressed in MVH-KO and WT replicates, most notably *Nsrp1* having a BF of 295121820.44 for the first variant. Although not as statistically significant as *Gemin7*, all of these genes were downregulated in MVH-KO samples (p-value < 0.05). This strengthens the hypothesis that the loss of *Ddx4* promotes alterations in the alternative splicing events on genes and thereby regulates their gene expression levels. Furthermore, by regulating the expression levels of the genes involved in mRNA splicing (*Gemin7*, *Nono*, *Npm1*, *Nsrp1*, and *Scnm1*), the loss of *Ddx4* might also indirectly affect mRNA splicing of other genes.

Gm14444 had a BF of 1000000000000 and a p-value of $2.74^{-0.6}$ for being differentially expressed. Although *Gm14444* is still widely unknown, it is predicted to participate in DNA binding activity and regulating transcription via RNA polymerase II ("Gene: Gm14444 (predicted gene 14444) Mus musculus,"). *Gm14444* was not found to be in a particularly enriched gene set, making it difficult to hypothesize *Gm14444*'s function and interaction with MVH and their pathways. It could be that *Gm14444* also interacts with mRNA splicing genes or that MVH regulates the splicing efficiency of *Gm14444*.

Upregulated exon-skipping genes were enriched in the spliceosome pathway in MVH-KO

To identify how many of these genes identified by MISO were exon-skipped, DEXSeq was performed. We confirmed that the fit curve to gene-wise dispersion estimates used in DEXSeq was reasonable, since the dispersions decreased with the increasing mean of normalized counts and the genes clustered around the fitted line appropriately (Figure 4A). In total, DEXSeq identified 292 genes that potentially switch isoforms that involve 461 transcripts. Out of 64132 genes, 461 genes were identified to have differentially expressed exonic regions with FDR < 0.05. Interestingly, although no gene sets or pathways were shown to be enriched, KEGG analysis showed that the suppressed genes that were exon-skipped were involved in cancer pathways (Figure 4B). Like the pathway analysis of differentially expressed genes, the spliceosome pathway was also enriched for activated genes with an enrichment score of 0.5315043 (Figure 4C). This could mean that *Ddx4* upregulates exon-skipping genes that are involved in the spliceosome pathways. One hypothesis is that the irregular transcriptome change from the exon-skipping promotes the spliceosome activity to balance those alterations in the transcriptome.

Out of the 461 genes identified by MISO, 6 genes were identified to be differentially expressed (*p*-value < 0.001 and |logFC| > 1.5), most notably *Gemin7* (Supplementary Table 1). 74 genes were detected by both DEXSeq and MISO (Supplementary Table 2), meaning that the differential isoform expression in MVH-KO samples was due to an exon-skipping event. Regardless, it is clear that the loss of MVH upregulates genes involved in the spliceosome, some of which are also exon-skipped like *Gemin7* (Figure 4D).

Genes undergoing alternative splicing in human GSC are not differentially expressed and spliced in mouse oocytes

Dr. Yusuke Suita identified three candidate genes that undergo alternative splicing in glioblastoma stem cells (GSC) using LeafCutter: *Postn*, *Ppia*, and *hnRNPK* (Suita et al., in preparation). These three genes were not differentially expressed in mouse oocytes (Supplementary Figure 3A), although *Ppia* in KO samples were slightly downregulated compared to the WT samples (*p*-value of 0.058), potentially because LeafCutter identifies differentially spliced transcripts, not differentially expressed genes. Thus, these three genes may have different isoform expressions instead. To confirm this, we used IGV to visualize the quantity of junction reads, which revealed slightly fewer function reads between the last two exons for *Ppia*, but not *Postn* and *Hnrnpk* (Supplementary Figures 3B, C, D), indicating possible exon skipping. Since the read depths were not big, however, we checked if this was a significant difference using MISO and DEXSeq and found no statistical significance. For *Postn*, among the 13 isoforms, the second isoform had a BF of 3.92, showing a slight differential isoform expression, whereas *Ppia* and *hnRNPK* showed no differential splicing at all (BF < 1). POSTN is essential for tissue development and regeneration, and the lack of POSTN in mice exhibits various defects (Walker, McLeod, Kim, Conway, & Hamilton, 2016), but the consequences of alternative splicing of *Postn* are unknown. All three genes were expressed in our samples, lowering the likelihood of the genes simply not being expressed in oocyte samples. Thus, it is likely that *Postn*, *Ppia*, and *hnRNPK* do not undergo alternative splicing or are differentially expressed in MVH-KO samples, unlike human GSC samples. It is also possible that the genes or pathways that are responsible for the splicing of these three genes are either not expressed in mouse oocytes or are not conserved in mice.

Discussion

This paper demonstrates computational methods to understand differential splicing and the role and mechanism of a gene in alternative splicing. Although *Ddx4* is known to be involved in various mechanisms, primarily in mRNA translation, germline development, gametogenesis, and fertility, *Ddx4* has not been discussed as an alternative splicing regulator. Computational tools help illuminate new insights from preexisting data that are difficult to quantify and accurately analyze with conventional biological approaches. This study uses a combination of computational methods even with a small sample size to study a new biological role of a gene in a top-down approach by looking at the differential expression of genes, exons, and isoforms and then looking at the pathways involved. A comparative analysis with human GSC data was also done in a candidate approach, in which the genes that were differentially spliced in GSC were investigated in mouse oocytes. To illuminate the mechanism behind *Ddx4*'s role in splicing, it was critical to computationally determine pathways involved in the interaction between *Ddx4* and other genes.

We detected two genes (*Gemin7* and *Gm14444*) that were both downregulated and experienced differential splicing. *Gemin7* is involved in mRNA splicing via spliceosome and was also identified to be exon-skipping. This could mean that *Ddx4* regulates alternative splicing of genes like *Gemin7* and consequently their gene expression levels, which may also affect other genes via mRNA splicing. Interestingly, the spliceosome pathway was enriched in upregulated genes, such as *Cdc5l*. The upregulation of *Cdc5l* and the spliceosome activity in MVH-KO could cause alterations in alternative splicing events and in gene expression levels. It would be interesting to study how the upregulation of spliceosome affects the downregulated mRNA splicing pathway via the spliceosome in MVH-KO samples. A more in-depth analysis that targets the pathways is required to understand the mechanism behind *Ddx4* in splicing. To do so, we should first quantify different alternative splicing events, such as altered 3' splice site, altered 5' splice site, intron retention, and exon skipping, which can be determined using rMATS (Shen et al., 2014). A limitation of this study is that we only looked at exon analysis (DEXSeq). Future studies could focus on intron retention using iREAD (H.-D. Li, Funk, & Price, 2020), since DEXSeq considers any reads that fall into the intron region as empty, thus eliminating those reads from our exon analysis. After finding statistically significant splicing events, future studies should look at the phenotypes of MVH-KO to detect any antagonistic activities (Tsukamoto et al., 2020). Detection of antagonistic activities would mean that the loss of *Ddx4* is detrimental to the subject of interest, whether it is the spliceosome or mRNA splicing mechanism, and that *Ddx4* can affect the conserved components of those mechanisms in humans, as well.

Although the genes that were alternatively spliced in human GSC (*Postn*, *Ppia,* and *hnRNPK*) were not differentially expressed in mouse oocytes, POSTN showed some differentially spliced variants with a BF of 3.92. Since the change in *Postn* isoform expression were observed in both human GSC and mouse oocytes, we cannot reject the hypothesis that there is a conserved role or pathway of *Ddx4* that regulates the splicing of *Postn*.

This paper primarily uses MISO to detect isoform expressions. Using other computational splicing-detecting software could illuminate new insights. We attempted to install SCISSOR and LeafCutter. However, SCISSOR required at least 7 samples while we had 4 samples. LeafCutter could not be installed properly, as the compilation of LeafCutter failed, likely due to issues with rstan files, although the exact issue could not be pinpointed and resolved. Since SCISSOR focuses on the structural changes in the RNA-seq coverage profile (Choi et al., 2021) while LeafCutter quantifies RNA splicing variation using short-read RNA-seq data (Y. I. Li et al., 2018), comparing the results from SCISSOR and LeafCutter to those from MISO will help us connect RNA splicing variation with shape changes and thus isoform changes.

It is important to test and validate the quality of our data and analyses with other larger datasets. For experimental validation, conventional biological approaches could be used to confirm our computationally produced results. It is also crucial to examine the effect of these splicing alterations and RNA expression levels on *Ddx4*-deficient phenotypes, as discussed above. This paper shows that VASA/DDX4 plays a role in alternative splicing, although its role, whether it is direct or indirect, and the underlying mechanism remains unclear. Alternative splicing is essential for transcriptome maintenance and control of gene expressions (Tian, Li, & Wu, 2020) as well as tissue and organ development (Baralle & Giudice, 2017). Since VASA/DDX4 is critical in germline and oocyte development and function, uncovering the role of VASA in alternative splicing can provide us with insights into the mechanism behind many pathologies, including developmental defects, infertility, and even cancer.

Acknowledgment

Firstly, I am extremely grateful to my PI and thesis advisor, Dr. Mamiko Yajima, as well as my second reader, Dr. Ashok Ragavendran, for guiding me through and entrusting me with this project. I also extend my deepest gratitude to Dr. Eric Salomaki for computational guidance and for helping me install MISO. I thank all of the above for constructive feedback throughout the project and for critical reading of the manuscript. Lastly, I thank Dr. Yusuke Suita for providing me with the human GSC analyses and insightful feedback throughout the early stages of the project.

References

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res*, *22*(10), 2008-2017. doi:10.1101/gr.133744.111

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Baccon, J., Pellizzoni, L., Rappsilber, J., Mann, M., & Dreyfuss, G. (2002). Identification and characterization of Gemin7, a novel component of the survival of motor neuron complex. *J Biol Chem, 277*(35), 31957-31962. doi:10.1074/jbc.M203478200

Balakrishnan, R., Harris, M. A., Huntley, R., Van Auken, K., & Cherry, J. M. (2013). A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford), 2013*, bat054. doi:10.1093/database/bat054

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol, 18*(7), 437-451. doi:10.1038/nrm.2017.27

Blighe, K., S Rana, & M Lewis. (2018). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. Retrieved from https://github.com/kevinblighe/EnhancedVolcano.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*.

Bonnal, S. C., López-Oreja, I., & Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer — implications for care. *Nature Reviews Clinical Oncology, 17*(8), 457-474. doi:10.1038/s41571-020-0350-x

Braat, A. K., van de Water, S., Korving, J., & Zivkovic, D. (2001). A zebrafish vasa morphant abolishes vasa protein but does not affect the establishment of the germline. *Genesis, 30*(3), 183-185. doi:10.1002/gene.1060

Chlon, T. M., Stepanchick, E., Hershberger, C. E., Daniels, N. J., Hueneman, K. M., Kuenzi Davis, A., . . . Starczynowski, D. T. (2021). Germline DDX41 mutations cause ineffective hematopoiesis and myelodysplasia. *Cell Stem Cell, 28*(11), 1966-1981.e1966. doi:10.1016/j.stem.2021.08.004

Choi, H. Y., Jo, H., Zhao, X., Hoadley, K. A., Newman, S., Holt, J., . . . Hayes, D. N. (2021). SCISSOR: a framework for identifying structural changes in RNA transcripts. *Nature Communications, 12*(1), 286. doi:10.1038/s41467-020-20593-3

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2). doi:10.1093/gigascience/giab008

David, C. J., & Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev, 24*(21), 2343-2364. doi:10.1101/gad.1973010

Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell, 166*(3), 538-554. doi:10.1016/j.cell.2016.07.012

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Du, J., Li, M., Yuan, Z., Guo, M., Song, J., Xie, X., & Chen, Y. (2016). A decision analysis model for KEGG pathway analysis. *BMC Bioinformatics, 17*(1), 407. doi:10.1186/s12859-016-1285-1

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., . . . Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics, 28*(20), 2678-2679. doi:10.1093/bioinformatics/bts503

Gene: Gm14444 (predicted gene 14444) Mus musculus. Retrieved from https://rgd.mcw.edu/rgdweb/report/gene/main.html?id=10000418

Guo, S.-m., Liu, X.-p., Tian, Q., Fei, C.-f., Zhang, Y.-r., Li, Z.-m., . . . Zhou, L.-q. (2022). Regulatory roles of alternative splicing at Ezh2 gene in mouse oocytes. *Reproductive Biology and Endocrinology, 20*(1), 99. doi:10.1186/s12958-022-00962-x

Hickford, D. E., Frankenberg, S., Pask, A. J., Shaw, G., & Renfree, M. B. (2011). DDX4 (VASA) Is Conserved in Germ Cell Development in Marsupials and Monotremes1. *Biology of Reproduction, 85*(4), 733-743. doi:10.1095/biolreprod.111.091629

Jokoji, G., Maeda, S., Oishi, K., Ijuin, T., Nakajima, M., Tawaratsumida, H., . . . Taniguchi, N. (2021). CDC5L promotes early chondrocyte differentiation and proliferation by modulating pre-mRNA splicing of SOX9, COL2A1, and WEE1. *J Biol Chem, 297*(2), 100994. doi:10.1016/j.jbc.2021.100994

Kahles, A., Lehmann, K. V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., . . . Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell, 34*(2), 211-224.e216. doi:10.1016/j.ccell.2018.07.001

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods, 7*(12), 1009-1015. doi:10.1038/nmeth.1528

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T. W., Isobe, T., Asada, N., Fujita, Y., . . . Nakano, T. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development, 131*(4), 839-849. doi:10.1242/dev.00973

Kuznicki, K. A., Smith, P. A., Leung-Chiu, W. M., Estevez, A. O., Scott, H. C., & Bennett, K. L. (2000). Combinatorial RNA interference indicates GLH-4 can compensate for GLH-1; these two P granule components are critical for fertility in C. elegans. *Development, 127*(13), 2907-2916. doi:10.1242/dev.127.13.2907

Laity, J. H., Lee, B. M., & Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol, 11*(1), 39-46. doi:10.1016/s0959-440x(00)00167-6

Li, H.-D., Funk, C. C., & Price, N. D. (2020). iREAD: a tool for intron retention detection from RNA-seq data. *BMC Genomics, 21*(1), 128. doi:10.1186/s12864-020-6541-0

Li, M., Ren, C., Zhou, S., He, Y., Guo, Y., Zhang, H., . . . Huo, R. (2021). Integrative proteome analysis implicates aberrant RNA splicing in impaired developmental potential of aged mouse oocytes. *Aging Cell, 20*(10), e13482. doi:10.1111/acel.13482

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics, 50*(1), 151-158. doi:10.1038/s41588-017-0004-9

Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923-930. doi:10.1093/bioinformatics/btt656

Linder, P., & Jankowsky, E. (2011). From unwinding to clamping — the DEAD box RNA helicase family. *Nature Reviews Molecular Cell Biology, 12*(8), 505-516. doi:10.1038/nrm3154

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Luo, W., Brouwer, Cory. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics, 29*(14), 1830-1831. doi:10.1093/bioinformatics/btt285

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature, 463*(7280), 457-463. doi:10.1038/nature08909

Penrad-Mobayed, M., Perrin, C., L'Hôte, D., Contremoulins, V., Lepesant, J. A., Boizet-Bonhoure, B., . . . Veitia, R. A. (2018). A role for SOX9 in post-transcriptional processes: insights from the amphibian oocyte. *Scientific Reports, 8*(1), 7191. doi:10.1038/s41598-018-25356-1

Pfister, D., De Mulder, K., Hartenstein, V., Kuales, G., Borgonie, G., Marx, F., . . . Ladurner, P. (2008). Flatworm stem cells and the germ line: Developmental and evolutionary implications of macvasa expression in Macrostomum lignano. *Developmental Biology, 319*(1), 146-159. doi:https://doi.org/10.1016/j.ydbio.2008.02.045

Qin, K., Jian, D., Xue, Y., Cheng, Y., Zhang, P., Wei, Y., . . . Yuan, X. (2021). DDX41 regulates the expression and alternative splicing of genes involved in tumorigenesis and immune response. *Oncol Rep, 45*(3), 1213-1225. doi:10.3892/or.2021.7951

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Schudrowitz, N., Takagi, S., Wessel, G. M., & Yajima, M. (2017). Germline factor DDX4 functions in blood-derived cancer cell phenotypes. *Cancer Sci, 108*(8), 1612-1619. doi:10.1111/cas.13299

Schüpbach, T., & Wieschaus, E. (1986). Maternal-effect mutations altering the anterior-posterior pattern of the Drosophila embryo. *Roux's archives of developmental biology, 195*(5), 302-317. doi:10.1007/BF00376063

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., . . . Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences, 111*(51), E5593-E5601. doi:doi:10.1073/pnas.1419161111

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545-15550. doi:doi:10.1073/pnas.0506580102

Tanaka, S. S., Toyooka, Y., Akasu, R., Katoh-Fukui, Y., Nakahara, Y., Suzuki, R., . . . Noce, T. (2000). The mouse homolog of Drosophila Vasa is required for the development of male germ cells. *Genes Dev, 14*(7), 841-853.

Thomson, T., & Lin, H. (2009). The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu Rev Cell Dev Biol, 25*, 355-376. doi:10.1146/annurev.cellbio.24.110707.175327

Tian, G. G., Li, J., & Wu, J. (2020). Alternative splicing signatures in preimplantation embryo development. *Cell & Bioscience, 10*(1), 33. doi:10.1186/s13578-020-00399-y

Tsukamoto, T., Gearhart, M. D., Kim, S., Mekonnen, G., Spike, C. A., & Greenstein, D. (2020). Insights into the Involvement of Spliceosomal Mutations in Myelodysplastic Disorders from Analysis of SACY-1/DDX41 in Caenorhabditis elegans. *Genetics, 214*(4), 869-893. doi:10.1534/genetics.119.302973

Walker, J. T., McLeod, K., Kim, S., Conway, S. J., & Hamilton, D. W. (2016). Periostin as a multifunctional modulator of the wound healing response. *Cell Tissue Res, 365*(3), 453-465. doi:10.1007/s00441-016-2426-6

Wenda, J. M., Homolka, D., Yang, Z., Spinelli, P., Sachidanandam, R., Pandey, R. R., & Pillai, R. S. (2017). Distinct Roles of RNA Helicases MVH and TDRD9 in PIWI Slicing-

Triggered Mammalian piRNA Biogenesis and Function. *Dev Cell, 41*(6), 623-637.e629. doi:10.1016/j.devcel.2017.05.021

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., . . . Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation, 2*(3), 100141. doi:https://doi.org/10.1016/j.xinn.2021.100141

Zhang, D., Keilty, D., Zhang, Z. F., & Chian, R. C. (2017). Mitochondria in oocyte aging: current understanding. *Facts Views Vis Obgyn, 9*(1), 29-38.

Zhang, J., Liu, W., Li, G., Xu, C., Nie, X., Qin, D., . . . Li, L. (2022). BCAS2 is involved in alternative splicing and mouse oocyte development. *Faseb j, 36*(2), e22128. doi:10.1096/fj.202101279R

Zhou, X., Liu, Z., He, T., Zhang, C., Jiang, M., Jin, Y., . . . Yang, X. (2022). DDX10 promotes the proliferation and metastasis of colorectal cancer cells via splicing RPL35. *Cancer Cell Int, 22*(1), 58. doi:10.1186/s12935-022-02478-1

Figures

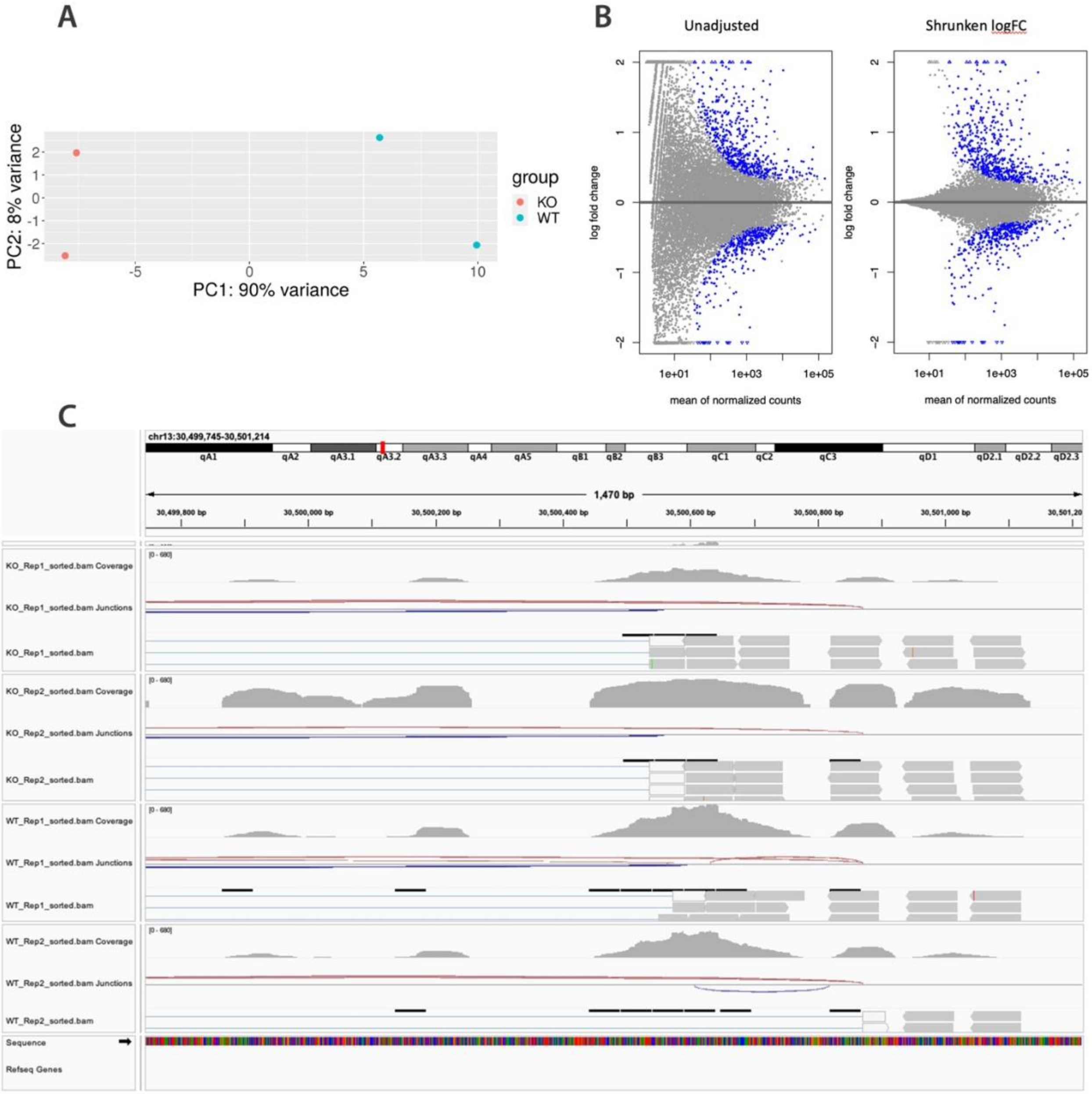


Figure 1: Overview of the mapped reads for each sample. (A) PCA plot of RNA-seq data of wildtype and MVH-knockout mouse oocyte samples. Two replicates per group were analyzed, as indicated. **(B)** MA plot of logFC over the mean of normalized counts for all samples in the DESeqDataSet. Blue points represent datapoints with adjusted *p*-values less than 0.1 (differentially expressed genes). Triangles represent the points that fall out of the window. Left MA plot is the

unadjusted plot, while the right MA plot is the shrunken logFC without low count genes. **(C)** Visualization of transcripts using IGV of potentially missing annotation. Arcs represent reads spanning exon junctions, and peaks represent exon coverage.

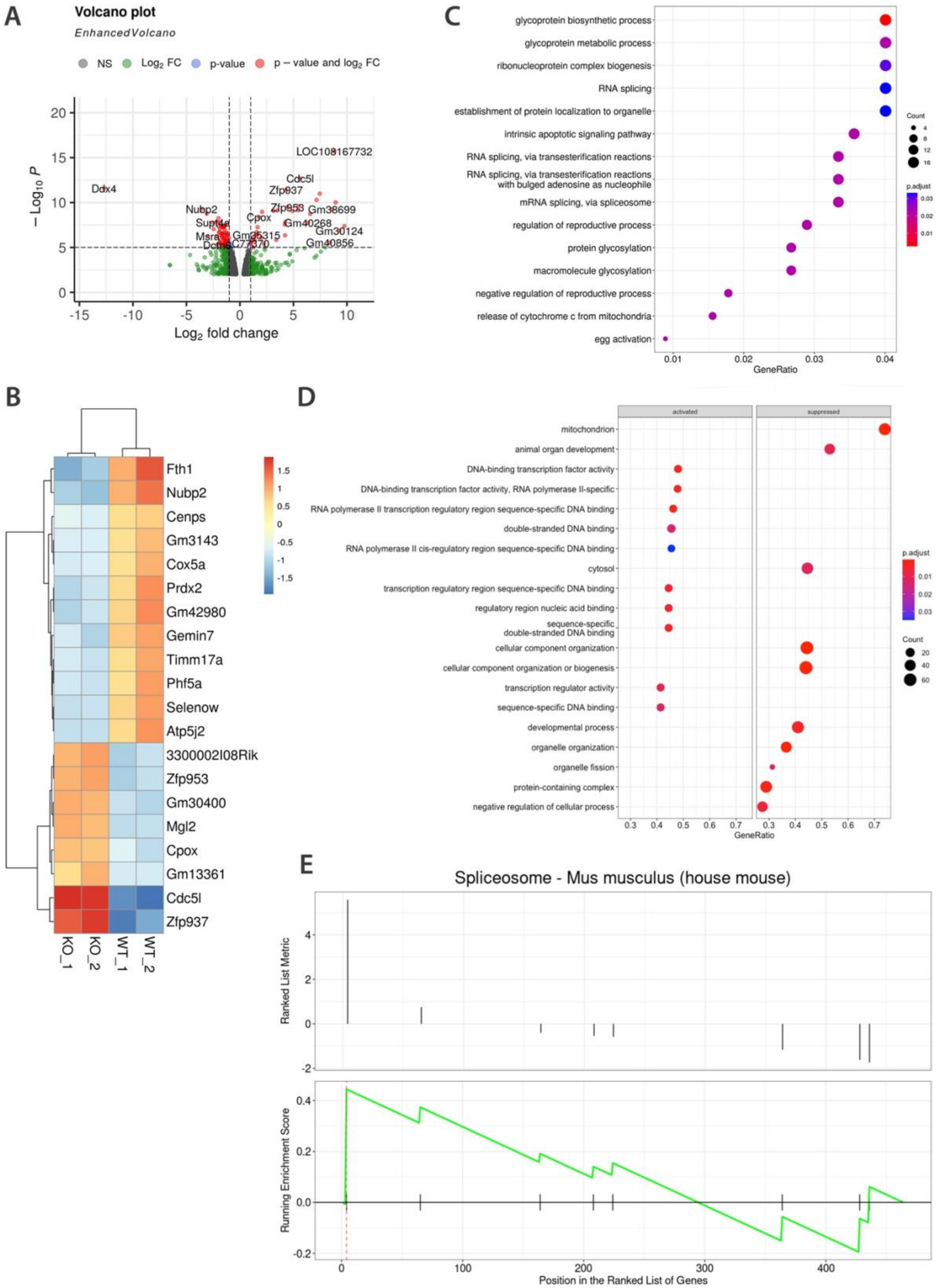

A
Volcano plot
EnhancedVolcano
NS
Log2 FC
p-value
p – value and log2 FC
– Log10 P
Log2 fold change
LOC108167732
Cdc5l
Ddx4
Zfp937
Zfp953
Gm36699
Nubp2
Cpox
Supt4a
Gm40268
Gm30124
Msra
Gm35315
Dctn6
C77370
Gm40856
20
15
10
5
0
-15
-10
-5
0
5
10
B
Fth1
Nubp2
Cenps
Gm3143
Cox5a
Prdx2
Gm42980
Gemin7
Timm17a
Phf5a
Selenow
Atp5j2
3300002I08Rik
Zfp953
Gm30400
Mgl2
Cpox
Gm13361
Cdc5l
Zfp937
KO_1
KO_2
WT_1
WT_2
1.5
1
0.5
0
-0.5
-1
-1.5
C
glycoprotein biosynthetic process
glycoprotein metabolic process
ribonucleoprotein complex biogenesis
RNA splicing
establishment of protein localization to organelle
intrinsic apoptotic signaling pathway
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via spliceosome
regulation of reproductive process
protein glycosylation
macromolecule glycosylation
negative regulation of reproductive process
release of cytochrome c from mitochondria
egg activation
Count
4
8
12
16
p.adjust
0.03
0.02
0.01
0.01
0.02
0.03
0.04
GeneRatio
D
activated
suppressed
mitochondrion
animal organ development
DNA-binding transcription factor activity
DNA-binding transcription factor activity, RNA polymerase II-specific
RNA polymerase II transcription regulatory region sequence-specific DNA binding
double-stranded DNA binding
RNA polymerase II cis-regulatory region sequence-specific DNA binding
cytosol
transcription regulatory region sequence-specific DNA binding
regulatory region nucleic acid binding
sequence-specific double-stranded DNA binding
cellular component organization
cellular component organization or biogenesis
transcription regulator activity
sequence-specific DNA binding
developmental process
organelle organization
organelle fission
protein-containing complex
negative regulation of cellular process
p.adjust
0.01
0.02
0.03
Count
20
40
60
0.3
0.4
0.5
0.6
0.7
GeneRatio
E
Spliceosome - Mus musculus (house mouse)
Ranked List Metric
4
2
0
-2
Running Enrichment Score
0.4
0.2
0.0
-0.2
0
100
200
300
400
Position in the Ranked List of Genes

Figure 2: Pathway analysis of differentially expressed genes. (A) Volcano plot with logFC and the −log10 of the adjusted *p*-value of genes between MVH-KO and WT samples. **(B)** GO enrichment analysis performed using the enrichgo function in ClusterProfiler. The top 5 gene ontologies, including RNA splicing, all had a gene ratio of 0.0401. **(C)** Heatmap of 30 genes that were the most differentially expressed. Each column represents a sample with each gene's expression level. The color represents the relative standing of the reads count data. **(D)** GSEA on differentially expressed genes. **(E)** The positive enrichment score of the spliceosome pathway (0.4444) visualized with DEXSeq.

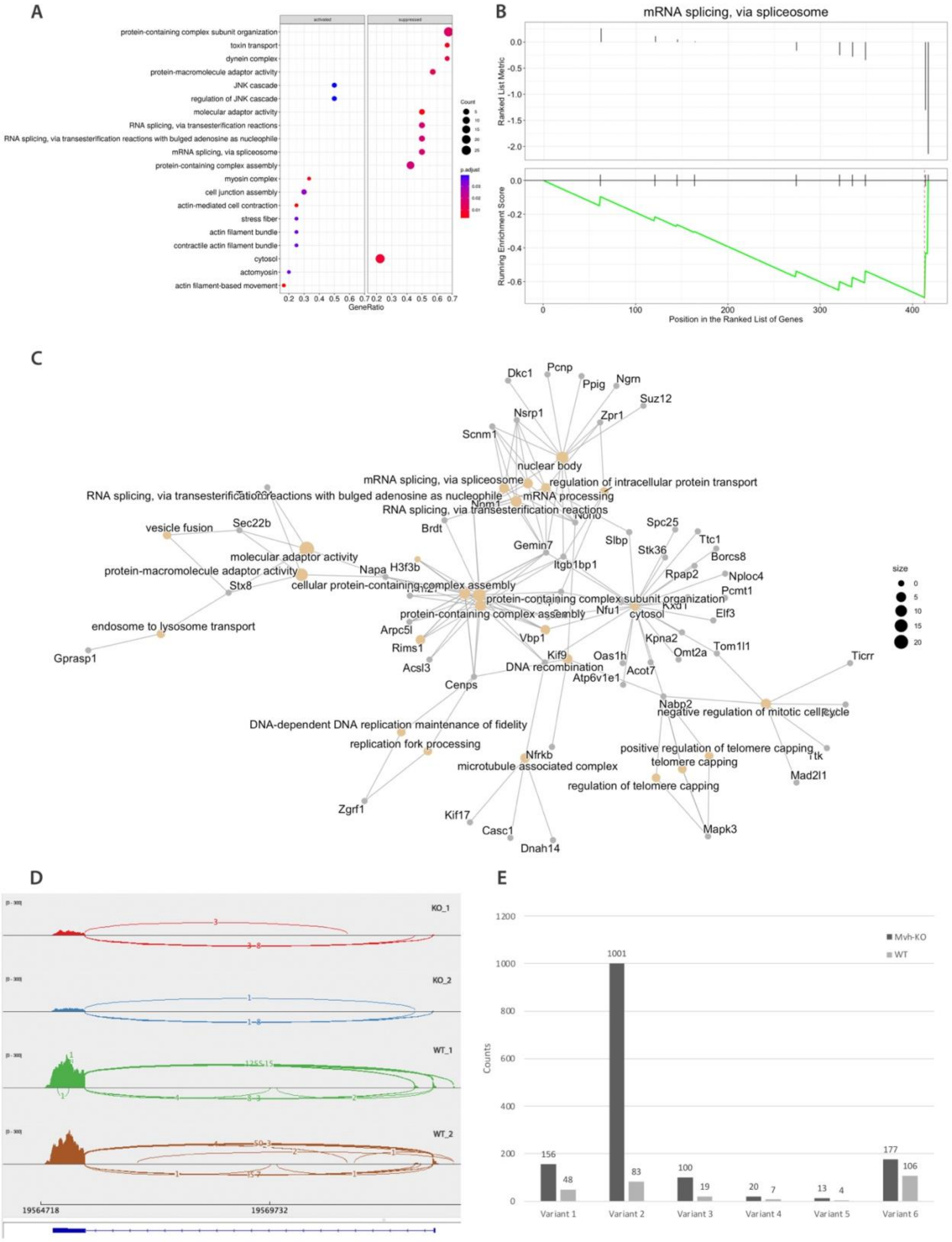

A
activated
suppressed
protein-containing complex subunit organization
toxin transport
dynein complex
protein-macromolecule adaptor activity
JNK cascade
regulation of JNK cascade
molecular adaptor activity
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via spliceosome
protein-containing complex assembly
myosin complex
cell junction assembly
actin-mediated cell contraction
stress fiber
actin filament bundle
contractile actin filament bundle
cytosol
actomyosin
actin filament-based movement
0.2 0.3 0.4 0.5 0.6 0.7 0.2 0.3 0.4 0.5 0.6 0.7
GeneRatio
Count
5
10
15
20
25
p.adjust
0.03
0.02
0.01
B
mRNA splicing, via spliceosome
Ranked List Metric
0.0
-0.5
-1.0
-1.5
-2.0
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
0
100
200
300
400
Position in the Ranked List of Genes
C
Dkc1
Pcnp
Ppig
Ngrn
Suz12
Nsrp1
Zpr1
Scnm1
nuclear body
mRNA splicing, via spliceosome
regulation of intracellular protein transport
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA processing
Npm1
RNA splicing, via transesterification reactions
Nono
vesicle fusion
Sec22b
Brdt
Spc25
Slbp
Ttc1
Gemin7
Stk36
Borcs8
molecular adaptor activity
Itgb1bp1
Rpap2
Nploc4
Napa
H3f3b
protein-macromolecule adaptor activity
Stx8
cellular protein-containing complex assembly
Pcmt1
protein-containing complex subunit organization
Xxd1
protein-containing complex assembly
Nfu1
cytosol
Elf3
endosome to lysosome transport
Arpc5l
Rims1
Vbp1
Kpna2
Tom1l1
Gprasp1
Kif9
Oas1h
Omt2a
Ticrr
Acsl3
DNA recombination
Acot7
Atp6v1e1
Cenps
Nabp2
negative regulation of mitotic cell cycle
DNA-dependent DNA replication maintenance of fidelity
replication fork processing
positive regulation of telomere capping
Ttk
Nfrkb
telomere capping
microtubule associated complex
Mad2l1
regulation of telomere capping
Zgrf1
Kif17
Casc1
Mapk3
Dnah14
size
0
5
10
15
20
D
KO_1
3
3 8
KO_2
1
1 8
WT_1
1
1255 15
1
4
8 3
2
WT_2
4
50 3
2
1
1
15 7
1
19564718
19569732
E
1200
1000
800
600
400
200
0
Counts
Mvh-KO
WT
156
48
1001
83
100
19
20
7
13
4
177
106
Variant 1
Variant 2
Variant 3
Variant 4
Variant 5
Variant 6

Figure 3: Pathway analysis on genes with differential isoform expression identified by MISO and analysis of GEMIN7. (A) GSEA on genes identified by MISO. RNA splicing and mRNA splicing were the most enriched among the suppressed genes with an enrichment score of -0.6931041. **(B)** Negative enrichment score of mRNA splicing via spliceosome visualized with DEXSeq. **(C)** Cnet plot for the GSEA result on genes identified by MISO produced with ClusterProfiler. **(D)** IGV visualization of splice junctions of GEMIN7. Arcs represent splice junctions that connect the exons. Numbers on the arcs represent the junction depth (number of reads for each junction). Blue track on the bottom represents the gene annotation track. **(E)** MISO isoform counts of GEMIN7 for KO and WT samples.

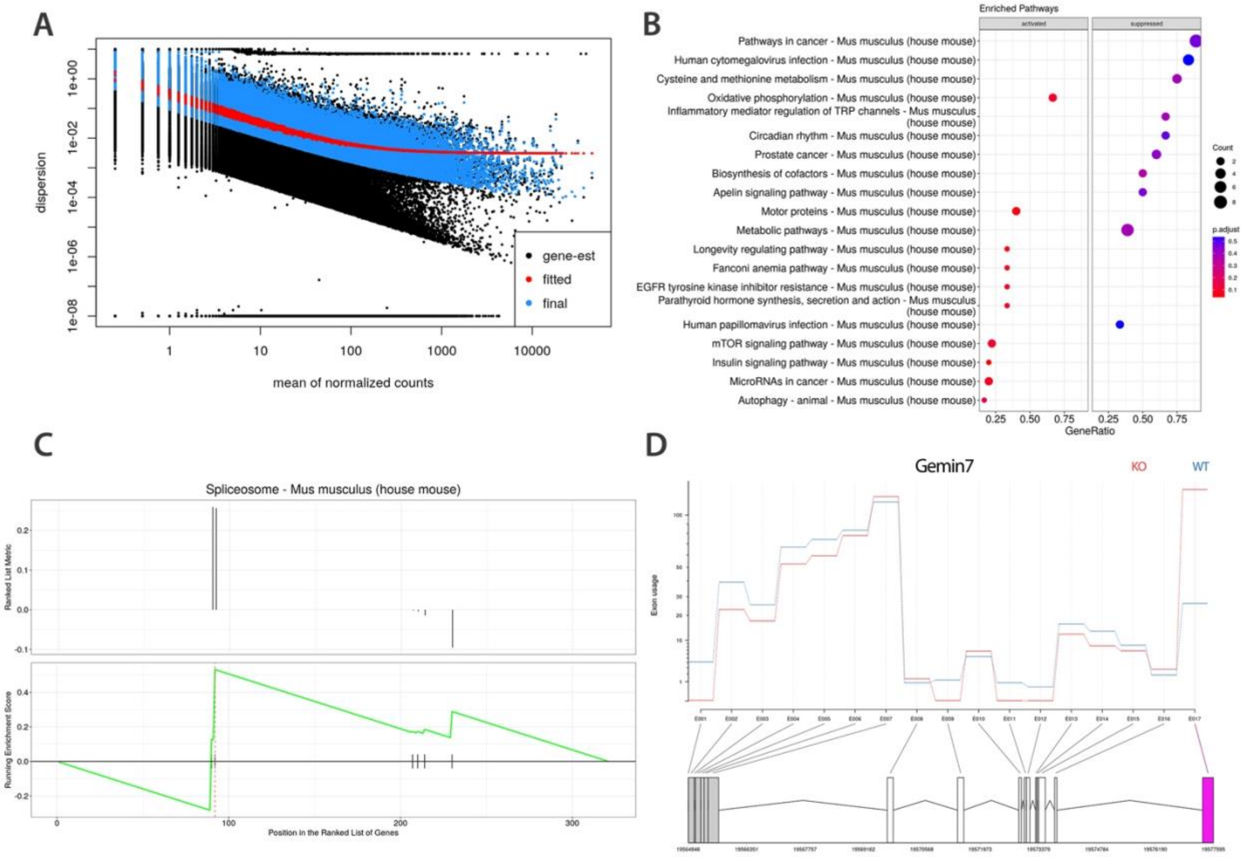
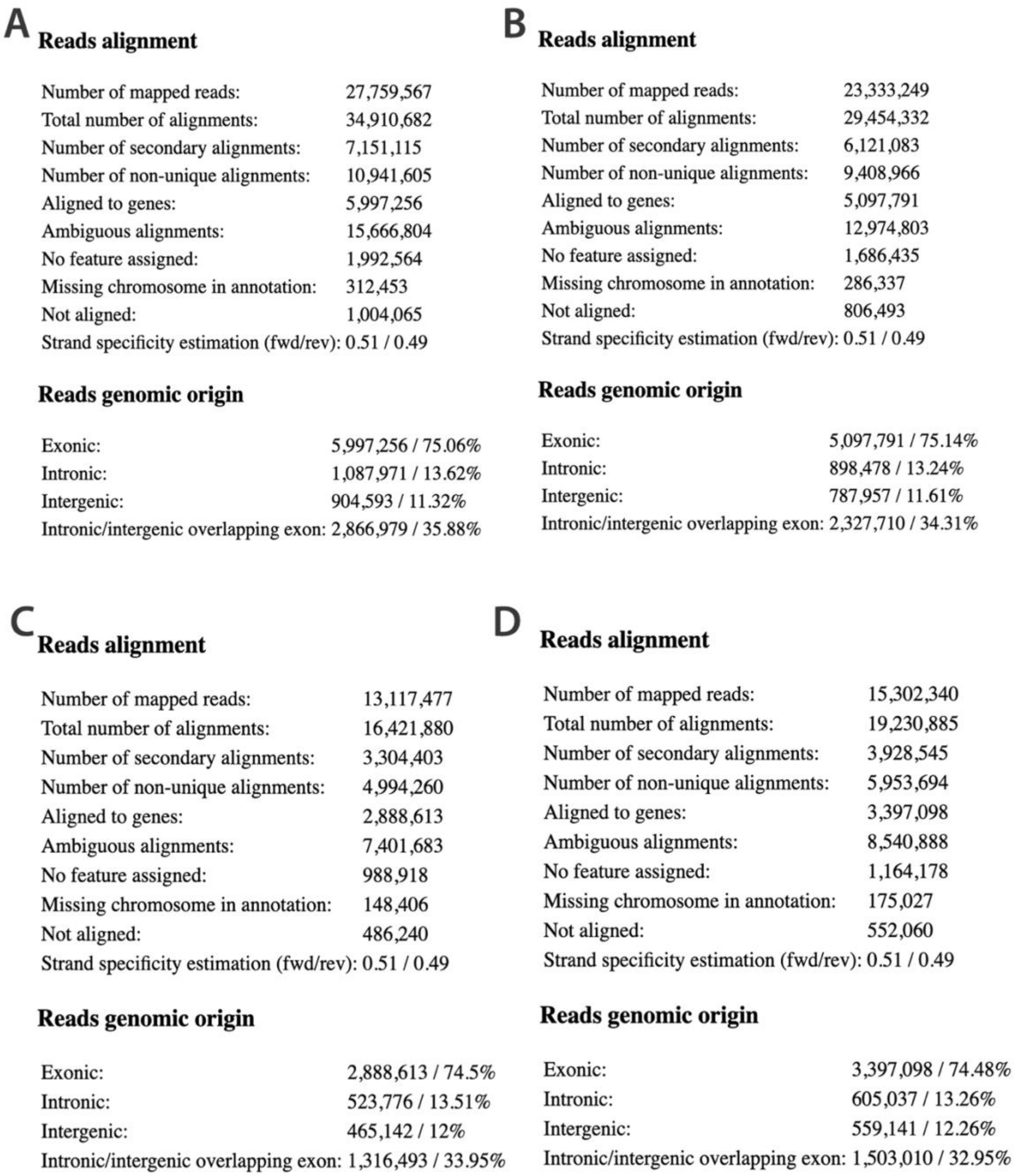


Figure 4: Using DEXSeq to detect exon-skipping genes and visualize alternative splicing of *Gemin7*. (A) Per-exon dispersion plot. Black points represent the initial per-exon dispersion estimates. Red line represents the fitted mean-dispersion values function. Blue points represent the shrunken values. **(B)** KEGG analysis of the exon-skipping genes. Pathway in cancer was the most enriched for suppressed genes with an enrichment score of -0.4180282. **(C)** Positive enrichment score of 0.5315043 for spliceosome pathway in exon-skipping genes visualized with DEXSeq. **(D)** Differential exon usage analysis of *Gemin7* after the counts were normalized to better visualize alternative splicing. Bars below are exons, lines between exons are introns, the numbers on the y-axis represent significant differential exon usage (FDR < 0.05), and pink exon is the differentially expressed exon.

Supplementary Figures



A

Reads alignment

Number of mapped reads:	27,759,567
Total number of alignments:	34,910,682
Number of secondary alignments:	7,151,115
Number of non-unique alignments:	10,941,605
Aligned to genes:	5,997,256
Ambiguous alignments:	15,666,804
No feature assigned:	1,992,564
Missing chromosome in annotation:	312,453
Not aligned:	1,004,065
Strand specificity estimation (fwd/rev):	0.51 / 0.49

Reads genomic origin

Exonic:	5,997,256 / 75.06%
Intronic:	1,087,971 / 13.62%
Intergenic:	904,593 / 11.32%
Intronic/intergenic overlapping exon:	2,866,979 / 35.88%

B

Reads alignment

Number of mapped reads:	23,333,249
Total number of alignments:	29,454,332
Number of secondary alignments:	6,121,083
Number of non-unique alignments:	9,408,966
Aligned to genes:	5,097,791
Ambiguous alignments:	12,974,803
No feature assigned:	1,686,435
Missing chromosome in annotation:	286,337
Not aligned:	806,493
Strand specificity estimation (fwd/rev):	0.51 / 0.49

Reads genomic origin

Exonic:	5,097,791 / 75.14%
Intronic:	898,478 / 13.24%
Intergenic:	787,957 / 11.61%
Intronic/intergenic overlapping exon:	2,327,710 / 34.31%

C

Reads alignment

Number of mapped reads:	13,117,477
Total number of alignments:	16,421,880
Number of secondary alignments:	3,304,403
Number of non-unique alignments:	4,994,260
Aligned to genes:	2,888,613
Ambiguous alignments:	7,401,683
No feature assigned:	988,918
Missing chromosome in annotation:	148,406
Not aligned:	486,240
Strand specificity estimation (fwd/rev):	0.51 / 0.49

Reads genomic origin

Exonic:	2,888,613 / 74.5%
Intronic:	523,776 / 13.51%
Intergenic:	465,142 / 12%
Intronic/intergenic overlapping exon:	1,316,493 / 33.95%

D

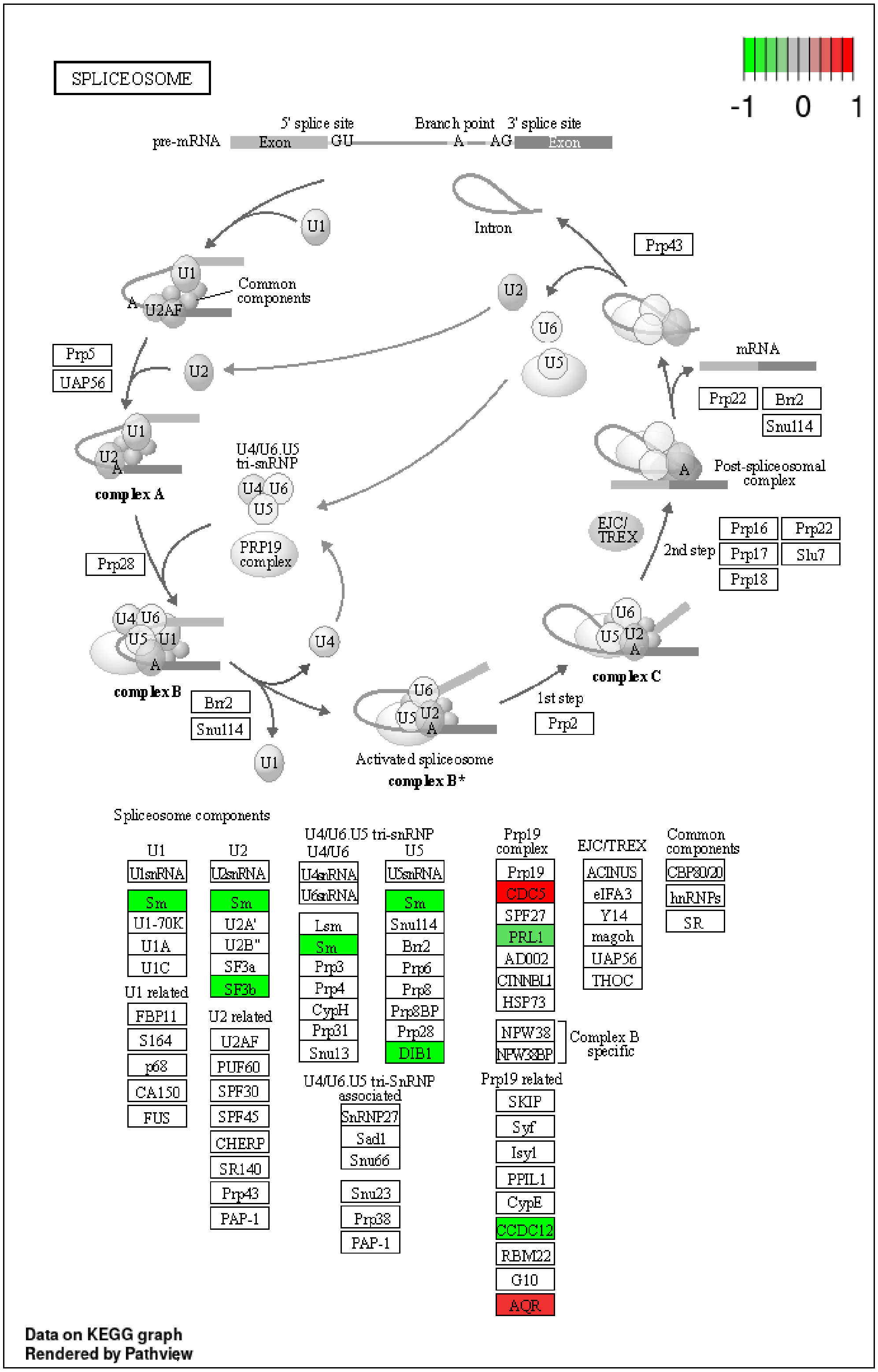
Reads alignment

Number of mapped reads:	15,302,340
Total number of alignments:	19,230,885
Number of secondary alignments:	3,928,545
Number of non-unique alignments:	5,953,694
Aligned to genes:	3,397,098
Ambiguous alignments:	8,540,888
No feature assigned:	1,164,178
Missing chromosome in annotation:	175,027
Not aligned:	552,060
Strand specificity estimation (fwd/rev):	0.51 / 0.49

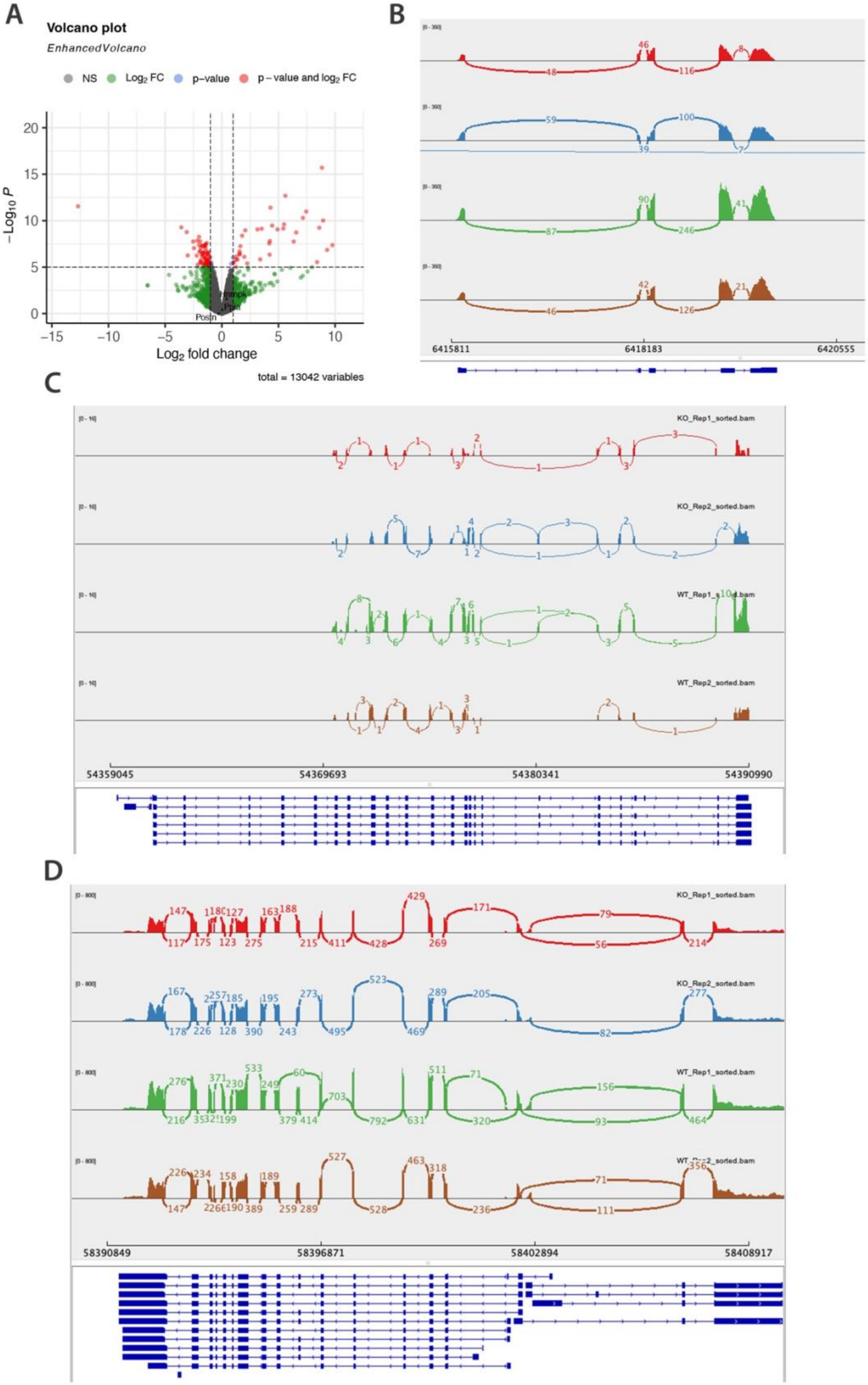
Reads genomic origin

Exonic:	3,397,098 / 74.48%
Intronic:	605,037 / 13.26%
Intergenic:	559,141 / 12.26%
Intronic/intergenic overlapping exon:	1,503,010 / 32.95%

Supplementary Figure 1: Qualimap summary of all samples. (A, B, C, D) Number of aligned reads and read genomic origin of **(A)** the first replicate of wildtype, **(B)** second replicate of wildtype, **(C)** first replicate of MVH-knockout, and **(D)** second replicate of MVH-knockout.



Supplementary Figure 2: KEGG spliceosome pathway and fold enrichment visualization using pathview (Luo, 2013). Red genes are upregulated. Green genes are downregulated.


A
Volcano plot
EnhancedVolcano
NS
Log2 FC
p−value
p − value and log2 FC
−Log10 P
Log2 fold change
Hnrnpk
Ppia
Posln
total = 13042 variables
B
6415811
6418183
6420555
C
KO_Rep1_sorted.bam
KO_Rep2_sorted.bam
WT_Rep1_sorted.bam
WT_Rep2_sorted.bam
54359045
54369693
54380341
54390990
D
KO_Rep1_sorted.bam
KO_Rep2_sorted.bam
WT_Rep1_sorted.bam
WT_Rep2_sorted.bam
58390849
58396871
58402894
58408917

Supplementary Figure 3: Comparative Analysis of PPIA, POSTN, and HNRNPK in GSC and mouse oocytes. (A) Volcano plot of *Postn, Ppia,* and *Hnrnpk*. **(B, C, D)** IGV visualization of the transcripts of **(B)** *Ppia*, **(C)** *Postn,* and **(D)** *Hnrnpk*.

Supplementary Tables

Gene Symbol	Log2FC	P-value	Bayes Factor
Gemin7	-2.1464944	3.66E-27	0.08, 1235292364.89, 0.14, 0.06, 0.03, 1000000000000, 0.05
Gm14444	3.6687339	2.74E-06	1E+12

Supplementary Table 1: 2 genes that are differentially expressed (DESeq2) and have differentially expressed isoforms (MISO).

Gene Symbol	Log2FC	P-value	Bayes Factor
Eif2b1	-0.22583326	0.06694722	47.38, 1.9, 1.36, 0.09, 0, 0
Oas1h	-0.49297973	0.00038692	1E+12
Rpap2	-0.37266838	0.00798616	0.04,0.08,0.07,0.18,10.46,0.08,0.09,0.45
Rnf216	0.1358222	0.39789388	39997474.90,0.10,0.00,0.00,0.00,0.00,0.08,0.34,0.54
Mphosph9	-0.08542244	0.58938086	11.59,0.54,0.38,0.03,0.08,0.24,0.05,0.15,0.09,0.00,0.06
Ywhag	0.2481941	0.09723803	1708352036.20,0.16,0.00,0.33,276.80
Srp72	-0.5305243	0.00042518	64466145.87,37.23,2.75,0.00,105.61,0.16,0.00,0.00,0.03
Rac1	-0.19331957	0.10812794	1000000000000.00,1.25,23398.84,0.00
Orc5	0.2911553	0.03099593	0.05,0.16,0.00,37.47,0.06,0.00
Brdt	-0.25917895	0.04681604	1000000000000.00,1000000000000.00,1000000000000.00,0.00
St13	-0.23649783	0.07830029	0.05,9.86,0.12,2946817.82,1.16
Sub1	-0.04800824	0.71351711	0.00,431632.37,1000000000000.00,0.00
Dcaf8	0.0110982	0.93182407	0.21,0.00,0.15,0.00,0.07,1108236.23,0.00,0.00,0.00,0.00,0.00
Stk36	-0.45374049	0.00624003	1000000000000.00,0.00,0.17,0.25,0.00,0.58,0.00,0.00,0.00
Tgfb2	-0.19795174	0.09857469	0.00,11076.77,0.07,0.04,0.00,0.00
Acsl3	-0.26374481	0.03526016	22836.46,1000000000000.00,267610.31,0.00,0.00,0.00,0.00
Dnah14	-0.62116425	0.00039658	0.00,1000000000000.00,195172.20,0.00,0.00,0.00,0.00,0.00,1205.25
Aspm	-0.04713134	0.68874034	1000000000000.00,0.00,0.00,0.00,1000000000000.00
Zfp57	-0.2611439	0.02941947	0.00,1.11,1000000000000.00,516407.48,0.00,0.00,0.06,0.00,0.03,0.00,0.00,0.00,0.00
Anapc16	-0.22476486	0.14871724	0.14,12.65,4.17,0.22,0.15,0.22,0.00,0.08
Ipmk	-0.1963838	0.13062177	0.76,2.76,0.06,0.03,0.00,0.03,91.62
Cand1	-0.40874403	0.00732546	9034416.48,0.12,0.00,1014.95
Rdh11	-0.60467561	0.00094232	24.71,0.89,0.27
Itgb1bp1	-0.92934039	7.20E-09	0.00,10800701865.68,0.04,2.02,0.69,0.04
Rbm25	0.0141143	0.90577721	0.05,31.36,4817.05,0.08,0.00,2.37,0.00,1.09,0.00,0.11,0.00,0.00,0.06,0.00,0.00,0.00,0.00,0.00

Wdr20	0.0813218	0.53610286	710.24,0.00,0.00,6.45,0.00,0.00,0.03
Npm1	-0.2514939	0.04421346	0.06,162.26,0.00,0.00,14.71,0.00
Skp1	-0.04686412	0.69288434	1478.84,130273.33,0.00,0.03,0.00,0.00
Mat2b	0.1678484	0.18961555	190.05,0.11,22.64,0.00,0.00,0.00,0.00
Slc22a23	0.3365686	0.04665682	6.62,0.08,0.03,0.05,24.82
Trim27	-0.24982965	0.12265224	1.11,0.31,0.02,73.26
Larp4b	0.1637056	0.21925328	0.18,0.00,0.00,1000000000000.00,0.54,0.04,0.00
Ccnb1	0.1392461	0.40718254	1000000000000.00,0.15,0.00,0.06,1000000000000.00,0.00,0.00
Cul1	-0.01175499	0.92449482	38.75,7.22,2.13,0.38,0.00,0.03,0.00,0.00,0.00
Mkrn2	0.2489767	0.09401989	1566.42,1845.21,0.00
Rad18	0.1547502	0.23071369	0.10,0.13,0.13,0.13,0.04,275.36,0.46,0.00
Atp6v1e1	-0.48863787	0.00026568	0.00,1000000000000.00,0.00,1000000000000.00,0.09,0.13
Oosp3	0.1672209	0.26922643	167.06
Ermp1	-0.00837935	0.94432066	0.00,0.00,1000000000000.00,0.00,0.00,1000000000000.00,0.00,0.00
Oosp1	0.3077957	0.01438299	92899879.07,0.00,92294.89
Arl6ip1	-0.24029021	0.03839489	1000000000000.00,0.00,0.05,0.08,0.00,3.12
Gemin7	-2.1464944	3.66E-27	0.08,1235292364.89,0.14,0.06,0.03,1000000000000.00,0.05
Nlrp9c	-0.1114054	0.47163293	0.00,1000000000000.00,1000000000000.00,0.00,0.00
C86187	-0.36059588	0.00379145	0.03,4372166.88,0.68,0.00
Mapk3	-0.07427819	0.66438677	0.00,0.09,0.00,0.00,0.00,0.51,0.00,11.34
AI987944	0.0752524	0.63971065	0.12,57170929.38,1000000000000.00,0.00
Obox2	-0.37102677	0.00346552	1000000000000.00,0.00,805670485743.19,1000000000000.00
Orc6	-0.17097343	0.18730952	0.00,0.00,0.00,14.48,0.00,14.11,0.00
Dkc1	-0.274683	0.03221666	12.76,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.34,0.00,0.06,0.00
Yipf6	0.3260191	0.04005083	77.46,0.41,0.04,0.00,2540.59,0.00
Zbtb33	0.1099899	0.55585571	107944245.21,0.36,0.94
Ttk	-0.43627648	0.00231335	0.91,0.25,182.56,0.00,0.34,0.11
Fbxw20	-0.56912535	0.00011551	28727.79,0.00,1708.08
Izumo1r	-0.39062274	0.01136888	21.36,0.00,1000000000000.00,1.61,0.12
Omt2a	-0.58049665	2.29E-06	0.00,1000000000000.00,1000000000000.00,0.00,0.00
Cacnb2	-0.16089263	0.52438852	0.07,0.05,0.10,0.04,6.00,0.06,0.03,0.06,0.04,0.03,0.04,7065525.81,0.00,0.14,0.04,0.00
Oser1	-1.00240036	4.45E-08	1000000000000.00,0.00,1000000000000.00,0.04,0.00,0.00,0.00
Serf2	-0.18081427	0.45533128	0.00,0.00,0.00,1000000000000.00,0.00,1000000000000.00
Gorasp2	-0.14451448	0.30903165	0.00,1000000000000.00,1000000000000.00,0.25,0.04,0.00
Dync1i2	-0.01958476	0.87416134	0.07,0.10,0.00,0.25,17391716.39,0.15,0.00,1.52,0.00,0.00,0.00,0.00
Mphosph8	-0.14386678	0.2144487	68560134084.19,14298432075.90,0.00,0.00,0.00
Naa30	-0.04096956	0.78581875	7373287.28,0.27,0.30,1000000000000.00,0.00

Fndc3a	-0.06492512	0.60236439	1000000000000.00,5.94,0.00,0.04,0.05,0.00,0.00
Ap3m1	-0.02025548	0.89897971	2127.61,0.00,0.00,2.69,0.00,0.00,0.00,0.08,0.03
Sec22b	-0.35165866	0.02386847	2759920231.71,1.00,0.09,0.00,0.05
Fxr1	-0.16584199	0.16503117	0.06,0.00,0.00,0.11,0.31,1000000000000.00,0.00,0.56,0.00
Bcar3	-0.54520854	0.00049186	0.18,0.29,1000000000000.00,0.22,0.00,0.00,0.00,0.00,0.00,0.00
Kif17	-0.65718976	1.63E-06	89.19,3.04,0.12,0.00
Cenps	1.35276164	1.73E-24	0.70,1000000000000.00,90.57,0.05,2.45,0.27,0.32
Snapc3	0.7464975	0.00088253	8.08,0.04,0.00,0.03,0.00,0.00,0.10,0.00,47.50
Msantd3	0.2099729	0.19174926	0.19,0.13,0.36,0.85,12.02
Sec22a	0.4823436	0.00068156	0.12,251699866.24,1000000000000.00,0.00,0.07
Nectin3	-0.15079422	0.22419127	0.00,0.66,0.09,0.04,0.00,0.00,0.00,1299.86,0.00,0.00
Rsl1d1	-0.21931635	0.0842793	1000000000000.00,0.00,1000000000000.00

Supplementary Table 2: 74 exon-skipping genes (DEXSeq) that have differentially expressed isoforms (MISO).

Gene Symbol	Log2FC	P-value	Bayes Factor
Prdx2	-1.88414256	5.34E-29	0.04,2.86,0.00,6.04,0.00,0.00,2.21,5.71
Cpox	1.7850287	1.41E-28	N/A
Gemin7	-2.1464944	3.66E-27	0.08, 1235292364.89, 0.14, 0.06, 0.03, 1000000000000, 0.05
Gm30400	5.490878	8.28E-24	3.4
Stmn1	-1.85389133	2.35E-05	0.07,0.00,0.00,0.06,0.00
A930017M01Rik	1.9480068	0.00805925	N/A

Supplementary Table 3: 6 exon-skipping genes (DEXSeq) that are differentially expressed (DESeq2).