RNA-Seq Based Gene Expression Profiling of Pasilla gene, the Drosophila ortholog of mammalian NOVA1 and NOVA2

NGS Data Analysis project

Tools: Galaxy, DAVID, Excel

INTRODUCTION

RNA sequencing is widely used for gene expression profiling to identify genes or molecular pathways that are differentially expressed between two different biological condition.

In this project, I used a computational workflow to investigate the impact of depleting the *Pasilla* (ps) gene in *Drosophila melanogaster* cells. RNA-Seq data from both Pasilla-depleted (treated) and control (untreated) samples were analyzed to determine how the loss of ps affects gene expression and associated biological pathways

The *Pasilla* gene in *Drosophila melanogaster* (fruit fly) encodes an RNA-binding protein that is the functional equivalent (ortholog) of the mammalian NOVA1 and NOVA2 proteins. It plays a crucial role in the regulation of alternative splicing, developmental and regulatory mechanisms. The *Pasilla* gene and NOVA1/2 (in insects and mammals respectively) are orthologues and their regulatory map/mechanism are highly conserved, however their target genes are distinctly different. This points to a conserved "RNA regulatory map" for splicing factors across evolution.

AIMS & OBJECTIVES

AIM: To identify genes and molecular pathways that are differentially expressed following depletion of the *Pasilla* gene in *Drosophila melanogaster*

OBJECTIVES

- Perform raw RNA-Seq data processing and quality control to ensure reliable results.
- Develop and implement a reproducible computational workflow using the Galaxy platform.
- Conduct downstream analysis to identify differentially expressed genes and enriched pathways resulting from *Pasilla* depletion.

WORKFLOW

Phase 1 (Data Processing & QC)

\FastQ file upload for dataset collection (treated &untreated paired end files)

\FastQC

\Trimmomatic. Ouput Fastq

\MultiQC to aggregate fastq results on both treated and untreated paired end files

Phase 2 (Read Alignment/Mapping)

\HISAT2 ref genome *Drosophila melanogaster* dm6. Ouput:BAM

Phase 3 (Downstream Analysis)

-Gene Expression Quantification & Analysis

\FeatureCounts

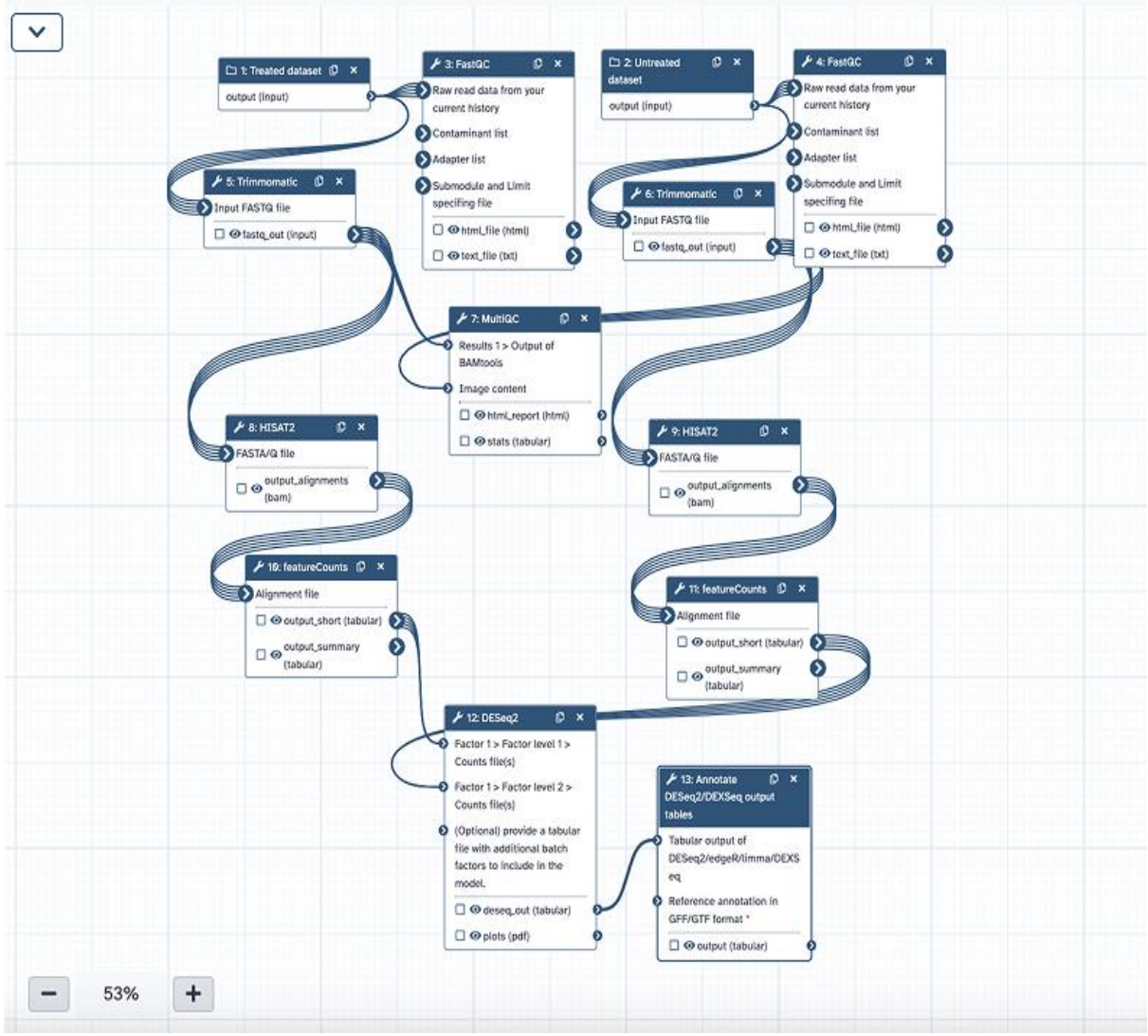
\DESeq2 - plots, tabular result file & normalized counts

\AnnotateDESeq2 - tabular file

-Differential analysis and Functional enrichment, Gene ontology

\Excel to sort DEs (adj p-value <0.05, log2 FC ≤ -1 for downregulated, ≥ 1 for upregulated genes)

\DAVID for GO annotation and KEGG pathway



METHODOLOGY

The analysis followed three main stages: data processing and quality control, read alignment, and downstream analysis for differential expression and pathway enrichment.

I began by uploading the paired-end raw FASTQ files for both treated and untreated samples from the Galaxy data library. Quality control was performed with FastQC, and Trimmomatic was used to trim low-quality bases, ensuring the data was clean and ready for analysis. The processed reads were then mapped to the *Drosophila melanogaster* reference genome (dm6) using HISAT2, which produced BAM files for each sample.

For quantifying gene expression, I used FeatureCounts to assign the aligned reads to genes, generating raw count data. Differential expression analysis was carried out with DESeq2, which provided statistical results, normalized counts, and visualizations. Since the initial DESeq2 output didn't provide gene annotations, I used AnnotateDESeq2 to add gene information, making the results easier to interpret.

To sort and filter the differentially expressed genes, I worked with the annotated results in Excel, selecting genes with an adjusted p-value below 0.05 and a log2 fold change of ≤ -1 for downregulated or ≥ 1 for upregulated genes. Finally, I used DAVID to perform Gene Ontology (GO) and KEGG pathway enrichment analysis, which helped identify the biological processes and pathways most affected by the loss of *Pasilla*.

RESULT INTERPRETATION

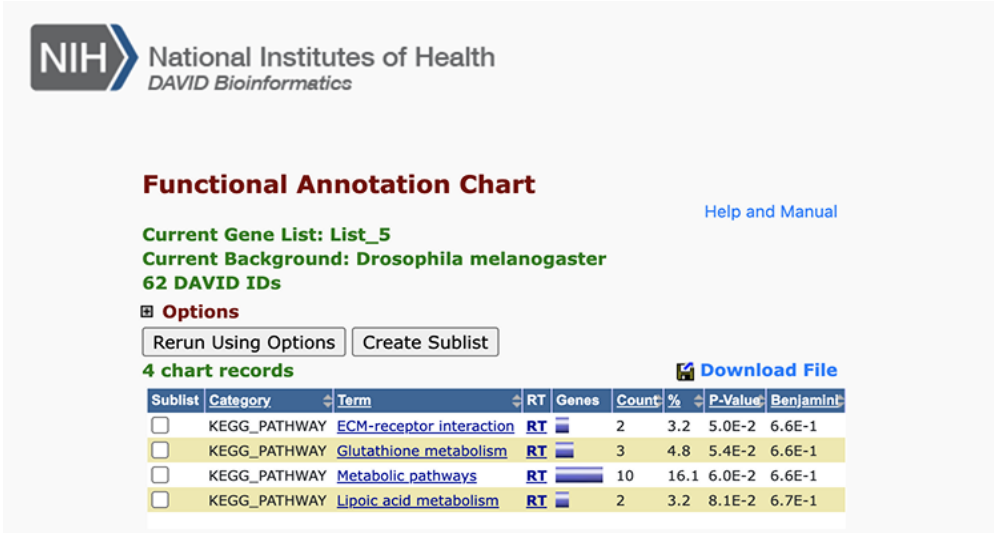
Downregulated Genes - (p-value < 0.05 and a log2 fold change of ≤ -1)

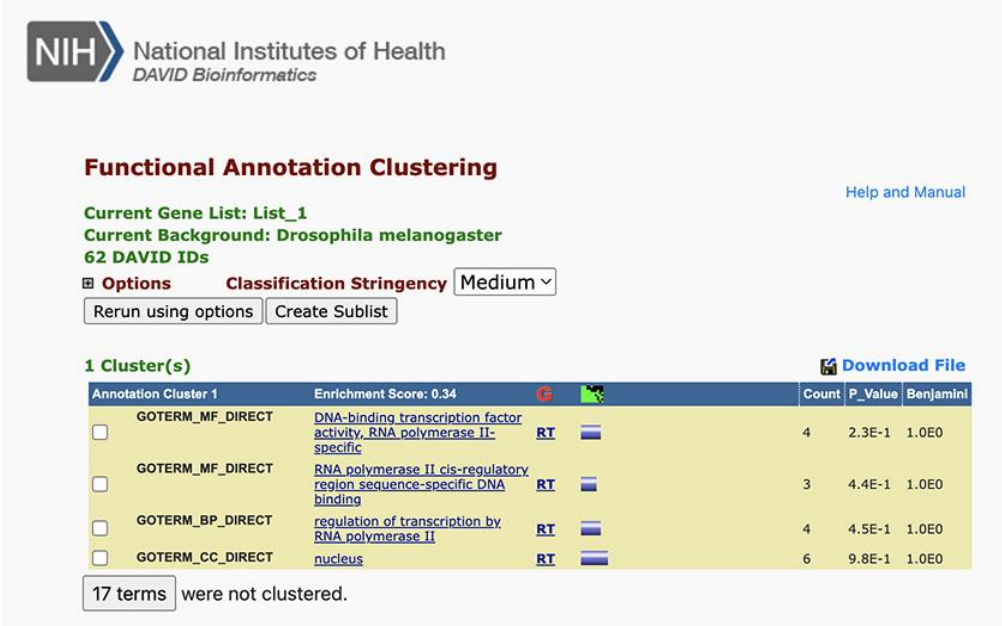
The KEGG Pathways identified and genes involved

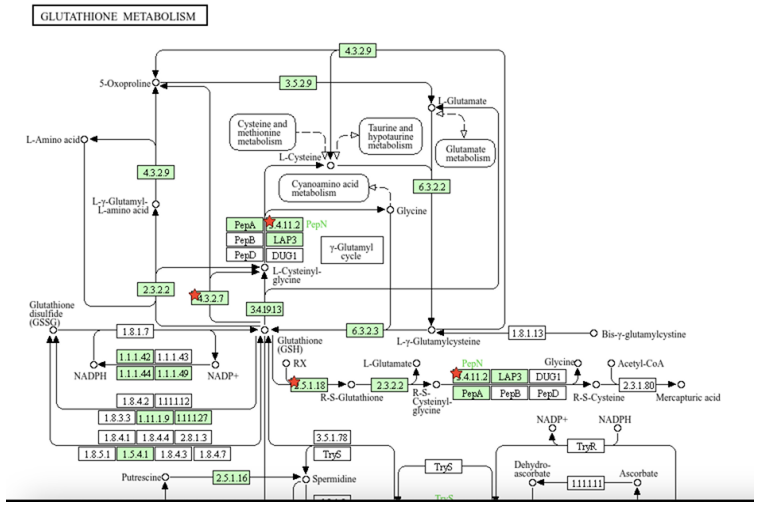
1. Metabolic pathways- CAHbeta
2. Lipoic acid metabolism- muc gene and CG1544 gene
3. Glutathione metabolism- GstE2 gene
4. ECM-receptor interaction – Hml gene nad CG3168 gene

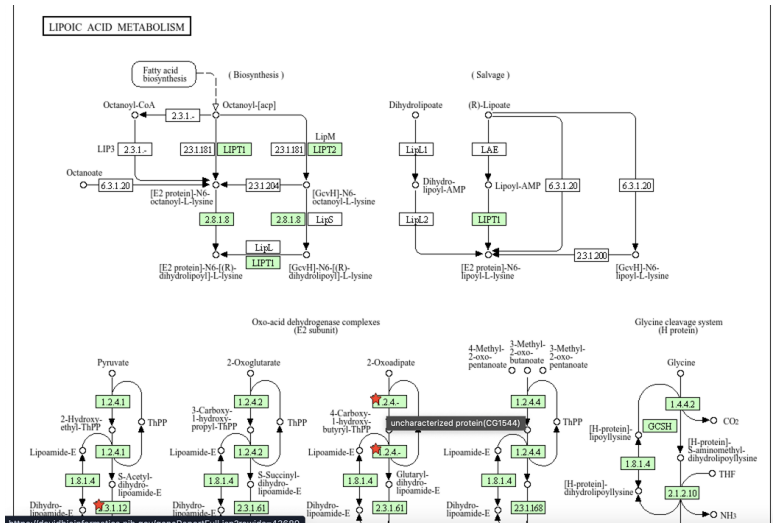
Gene Ontology (GO)

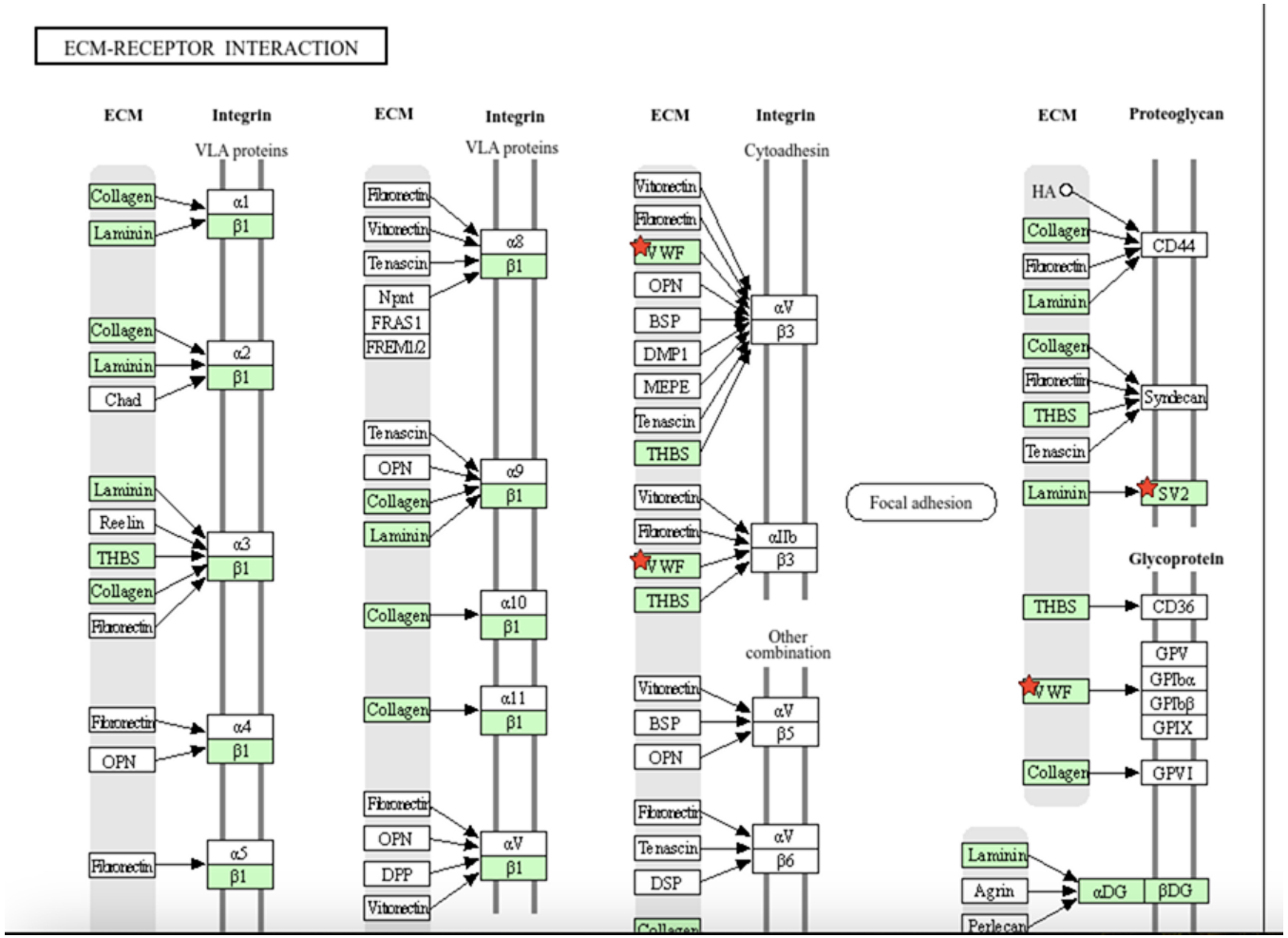
GO terms identified : DNA-binding transcription factor activity, RNA polymerase II-specific, Nucleus, RNA polymerase II Cis- regulatory region sequence specific DNA binding






GLUTATHIONE METABOLISM
5-Oxoproline
L-Amino acid
L-Glutamate
Cysteine and methionine metabolism
Taurine and hypotaurine metabolism
Glutamate metabolism
L-Cysteine
Cyanoamino acid metabolism
Glycine
L-γ-Glutamyl-L-amino acid
PepA
PepB
PepD
LAP3
DUG1
PepN
γ-Glutamyl cycle
L-Cysteinyl-glycine
Glutathione disulfide (GSSG)
Glutathione (GSH)
L-γ-Glutamylcysteine
Bis-γ-glutamylcystine
NADPH
NADP+
RX
R-S-Glutathione
L-Glutamate
R-S-Cysteinyl-glycine
Glycine
Acetyl-CoA
R-S-Cysteine
Mercapturic acid
Putrescine
Spermidine
TryS
TryR
Dehydro-ascorbate
Ascorbate


LIPOIC ACID METABOLISM
Fatty acid biosynthesis
(Biosynthesis)
(Salvage)
Octanoyl-CoA
Octanoyl-[acp]
Dihydrolipoate
(R)-Lipoate
LIP3
LIPT1
LipM
LIPT2
Octanoate
[E2 protein]-N6-octanoyl-L-lysine
[GcvH]-N6-octanoyl-L-lysine
LipL
LipS
LipL1
LAE
Dihydro-lipoyl-AMP
Lipoyl-AMP
LipL2
[E2 protein]-N6-dihydrolipoyl]-L-lysine
[GcvH]-N6-[(R)-dihydrolipoyl]-L-lysine
[E2 protein]-N6-lipoyl-L-lysine
[GcvH]-N6-lipoyl-L-lysine
Oxo-acid dehydrogenase complexes (E2 subunit)
Glycine cleavage system (H protein)
Pyruvate
2-Oxoglutarate
2-Oxoadipate
Glycine
ThPP
Lipoamide-E
S-Acetyl-dihydro-lipoamide-E
S-Succinyl-dihydro-lipoamide-E
Glutaryl-dihydro-lipoamide-E
uncharacterized protein(CG1544)
GCSH
THF
CO2
NH3


ECM-RECEPTOR INTERACTION
ECM
Integrin
VLA proteins
Collagen
Laminin
Chad
Reelin
THBS
Fibronectin
OPN
α1
α2
α3
α4
α5
β1
Fibronectin
Vitronectin
Tenascin
Npnt
FRAS1
FREM1/2
α8
α9
α10
α11
αV
DPP
Cytoadhesin
VWF
BSP
DMP1
MEPE
αV
β3
αIIb
β3
αV
β5
αV
β6
DSP
Other combination
Focal adhesion
Proteoglycan
HA
CD44
Syndecan
SV2
Glycoprotein
CD36
GPV
GPIbα
GPIbβ
GPIX
GPVI
Agrin
αDG
βDG

RESULT INTERPRETATION

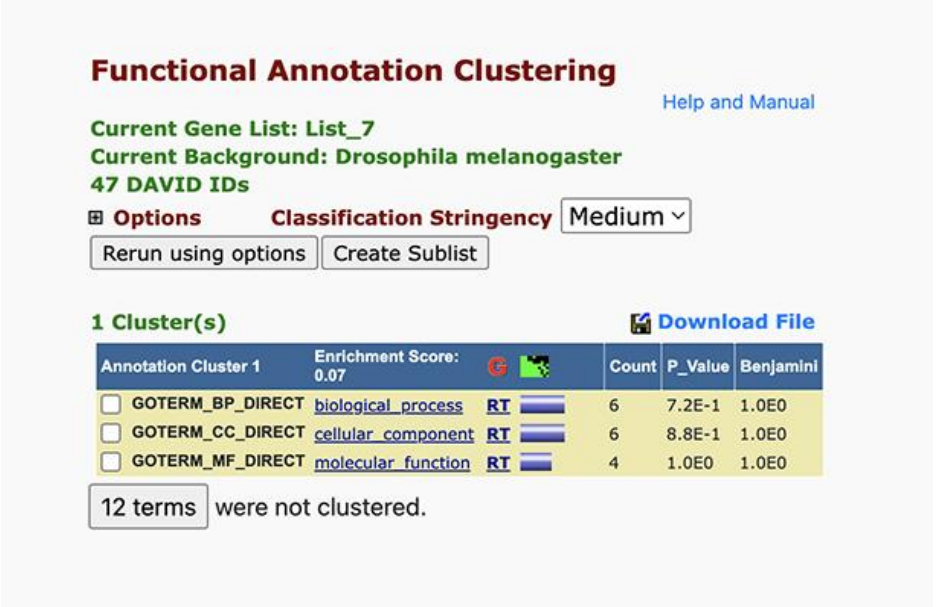
Upregulated Genes - (p-value < 0.05 and a log2 fold change of ≥ 1)

The KEGG Pathways identified and genes involved

1. Starch & Sucrose metabolism- Treh gene
2. Metabolic pathways – CAHbeta gene
3. Ribosome biogenesis in eukaryotes- NXF3 gene
4. Arginine and proline metabolism- ArgK1 gene
5. MAPK Signalling Pathways-Fly- Src64b gene
6. Retinol metabolism & Biosynthesis cofactor – Firl gene
7. Mucin Type O-Glycan. Biosynthesis – Pgant 9 gene
8. Tyrosine – PP01 gene

Gene Ontology (GO)

GO terms identified : Stress response and unfolded protein response (UPR), Apoptotic process and cell death regulation, Immune response and defense mechanisms, RNA metabolic process and regulation of gene expression, negative regulation of developmental processes



Functional Annotation Clustering

Help and Manual

Current Gene List: List_7
Current Background: Drosophila melanogaster
47 DAVID IDs

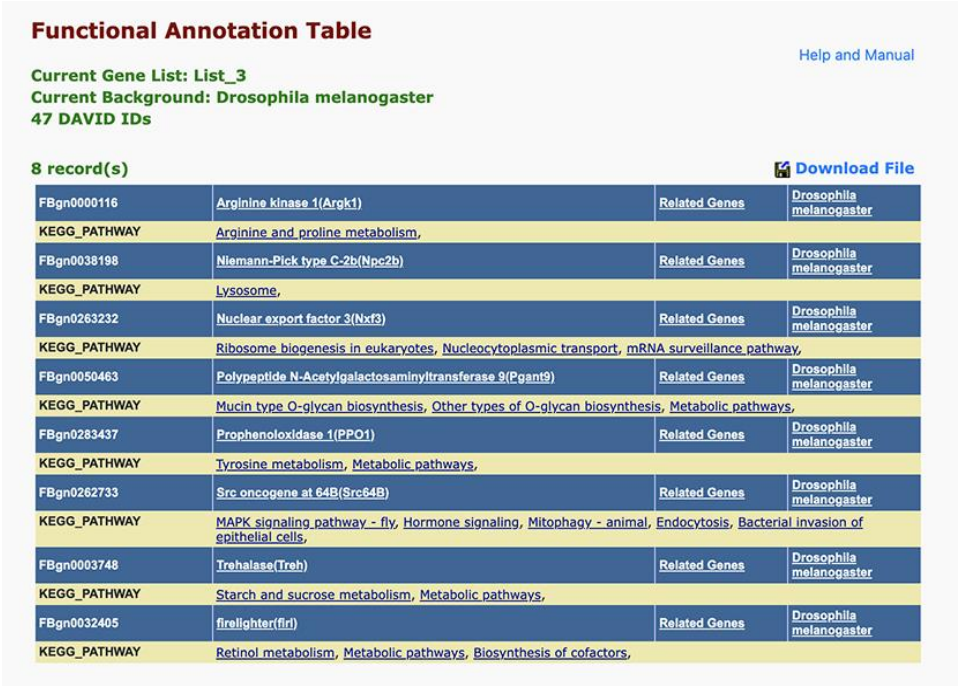
Options Classification Stringency: Medium

Rerun using options | Create Sublist

1 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 0.07	Count	P_Value	Benjamini
GOTERM_BP_DIRECT	biological_process RT	6	7.2E-1	1.0E0
GOTERM_CC_DIRECT	cellular_component RT	6	8.8E-1	1.0E0
GOTERM_MF_DIRECT	molecular_function RT	4	1.0E0	1.0E0

12 terms were not clustered.



Functional Annotation Table

Help and Manual

Current Gene List: List_3
Current Background: Drosophila melanogaster
47 DAVID IDs

8 record(s) Download File

FBgn0000116	Arginine kinase 1(Argk1)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Arginine and proline metabolism,		
FBgn0038198	Niemann-Pick type C-2b(Npc2b)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Lysosome,		
FBgn0263232	Nuclear export factor 3(Nxf3)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Ribosome biogenesis in eukaryotes, Nucleocytoplasmic transport, mRNA surveillance pathway,		
FBgn0050463	Polypeptide N-Acetylgalactosaminyltransferase 9(Pgant9)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Mucin type O-glycan biosynthesis, Other types of O-glycan biosynthesis, Metabolic pathways,		
FBgn0283437	Prophenoloxidase 1(PPO1)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Tyrosine metabolism, Metabolic pathways,		
FBgn0262733	Src oncogene at 64B(Src64B)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	MAPK signaling pathway - fly, Hormone signaling, Mitophagy - animal, Endocytosis, Bacterial invasion of epithelial cells,		
FBgn0003748	Trehalase(Treh)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Starch and sucrose metabolism, Metabolic pathways,		
FBgn0032405	firelighter(firl)	Related Genes	Drosophila melanogaster
KEGG_PATHWAY	Retinol metabolism, Metabolic pathways, Biosynthesis of cofactors,		

CONCLUSION

Depleting the Pasilla gene in *Drosophila melanogaster* leads to significant changes in gene expression affecting multiple biological pathways. The downregulation of metabolic and ECM-related genes points to impaired cellular metabolism and tissue maintenance. Meanwhile, upregulated genes reflect activation of stress and immune responses, as well as adjustments in protein synthesis and signalling pathways.

The presence of both metabolic and signalling pathway alterations highlights the broad impact of *Pasilla* on cellular physiology, influencing not only gene expression but also metabolic and developmental processes.

For more comprehensive studies, results files and datasets are available here
https://github.com/CreativeLuchi/RNA-seq-NGS-analysis-of-Pasilla-Gene-depletion-

REFERENCES

1. Seshaiah, P., Miller, B., Myat, M. M., & Andrew, D. J. (2001). *pasilla, the Drosophila homologue of the human Nova-1 and Nova-2 proteins, is required for normal secretion in the salivary gland*. Developmental Biology, 239(2), 309–322. https://doi.org/10.1006/dbio.2001.0429
2. Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., & Graveley, B. R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, 21(2), 193–202. https://doi.org/10.1101/gr.108662.110
3. Bérénice Batut, Mallory Freeberg, Mo Heydarian, Anika Erxleben, Pavankumar Videm, Clemens Blank, Maria Doyle, Nicola Soranzo, Peter van Heusden, Lucille Delisle, Reference-based RNA-Seq data analysis (Galaxy Training Materials). https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html Online; accessed Wed Jul 02 2025
4. Hiltemann, Saskia, Rasche, Helena et al., 2023 Galaxy Training: A Powerful Framework for Teaching! PLOS Computational Biology 10.1371/journal.pcbi.1010752
5. Batut et al., 2018 Community-Driven Data Analysis Training for Biology Cell Systems 10.1016/j.cels.2018.05.012