Data Article

Data on RNA-seq analysis of Drosophila melanogaster during ageing

Morteza Bajgiran\textsuperscript{a}, Azali Azlan\textsuperscript{a,b}, Shaharum Shamsuddin\textsuperscript{b,c}, Ghows Azzam\textsuperscript{a,b,∗}, Mardani Abdul Halim\textsuperscript{b,d,∗}

\textsuperscript{a} School of Biological Sciences, Universiti Sains Malaysia, Gelugor, Penang 11800, Malaysia  
\textsuperscript{b} USM-RIKEN Interdisciplinary Collaboration for Advanced Sciences, Universiti Sains Malaysia, Gelugor, Penang 11800, Malaysia  
\textsuperscript{c} School of Health Science, Universiti Sains Malaysia Health Campus, Kubang Kerian, Kelantan 16150, Malaysia  
\textsuperscript{d} Biotechnology Research Institute, Universiti Malaysia Sabah, Jalan UMS, Kota Kinabalu, Sabah 88400, Malaysia

Abstract

Ageing is defined as gradual decline of physiological, cellular and molecular state of an organism with time. The age-associated cell dysfunctions usually cause chronic diseases such as diabetes, cancers and other age-related diseases. Many of the genes and pathways involved in ageing are conserved in different species. These genes and pathways have been categorised into nine cellular and molecular hallmarks, namely, genomic instability, telomere attrition, loss of proteostasis, mitochondrial dysfunction, epigenetic alterations, deregulated nutrient sensing, stem cell exhaustion, cellular senescence and altered intercellular communication. Despite countless studies on ageing, the molecular mechanism of ageing is poorly understood. Here, we performed genome wide transcriptome mapping of ageing process in \textit{D. melanogaster}. In which, transcriptomic analysis conducted on the 1 day and 60 days flies. Illumina Hiseq platform were used to generate raw data. Afterwards, further analysis including differential expression analysis, GO classification and KEGG pathway enrichment analysis were performed. The raw data were uploaded to SRA database and the BioProject ID

https://doi.org/10.1016/j.dib.2021.107413  
2352-3409 © 2021 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
is PRJNA718442. These data provide the basis for future research in order to discover the genes and pathways involved in ageing.

© 2021 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

**Specifications Table**

| Subject                  | Ageing                                                                 |
|--------------------------|------------------------------------------------------------------------|
| Specific subject area    | Transcriptomic changes during ageing, comparing expression changes that occur in ageing, investigating the involvement of molecular pathways in ageing |
| Type of data             | RNA-seq data, figures, tables                                          |
| How data were acquired   | RNA sequencing by Illumina Hiseq platform                              |
| Data format              | Raw (FASTQ), excel spreadsheet, image, table                           |
| Parameters for data collection | Total RNA extraction and sequencing of samples in two different conditions, namely, day 1 Drosophila melanogaster (young) and day 60 Drosophila melanogaster (old) performed. |
| Description of data collection | Total RNA was isolated using Trizol reagent and RNeasy MinElute Cleanup Kit. RNA quality was evaluated by electrophoresis, Nanodrop2000 and Agilent2100 Bioanalyzer. rRNA was removed and then samples prepared and sequenced. |
| Data source location     | School of biological sciences, Universiti Sains Malaysia (USM), Malaysia (5.355° N, 100.3012° E) |
| Data accessibility       | Data can be accessed from NCBI SRA (BioProject ID: PRJNA718442)         |

[https://www.ncbi.nlm.nih.gov/bioproject/PRJNA718442](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA718442)

**Value of the Data**

- These data provide a comprehensive picture with a greater resolution of gene expression changes and the pathways involved in the process of ageing in *D. melanogaster*.
- The dataset and analysis provided here can be useful for researchers focusing on ageing and age-related diseases such as Alzheimer, cancer, and cardiovascular diseases in *D. melanogaster*.
- Applying different workflows, the RNA-seq raw data provided here can be used for further analysis to investigate the role of coding and non-coding genes in ageing. Besides, the analysis provided here would shed light on potential genes and pathways involved in ageing process for further molecular research in order to find novel anti-ageing strategies and treatments for age-related diseases.

1. **Data Description**

To investigate changes in molecular landscape in ageing process, day 1 and day 60 flies of *D. melanogaster* were chosen as model system and RNA sequencing was done using Illumina Hiseq platform. Table 1 provides accession numbers and links for raw data generated by RNA sequencing. There are in total three paired end libraries for day 1, and three paired end libraries for day 60 flies. Raw reads generated was mapped by HISAT2 and differential expression analysis was performed using edgeR. Table 2 shows the summary of libraries statistics and mapping including number of raw reads, number of cleaned reads and mapping rates. Differentially expressed genes and their respective fold change and expression levels as count per million (CPM) are listed in supplementary 1. Differentially expressed genes, further, were chosen for GO classification and KEGG pathway analysis. The enriched GO terms featuring biological process, cellular component,
and molecular functions and the number of differentially expressed genes related to those GO terms are presented in Tables 3–5, respectively. Table 6 shows the result of KEGG pathway enrichment analysis in day 60 compared to day 1 flies. Number of differentially expressed genes related to each KEGG pathway is provided in Table 6.

2. Experimental Design, Materials and Methods

2.1. Fly husbandry

Wild-type Oregon-R (OreR) (genotype: Oregon-R-P2; stock no.: 107294) from Kyoto Stock Center was used. The flies were maintained at 25 °C, 12 h light/dark cycle in a corn-based meal consists of 4% (w/v) corn starch, 5% (w/v) polenta, 10% (w/v) brown sugar, 0.7% (w/v) agar, 5% (w/v) yeast, 3% (w/v) nipagin and 0.7% (v/v) propionic acid.

2.2. Total RNA extraction, library construction, and RNA-seq

Equal number of male and female flies was used to extract the total RNA. A combination of Trizol reagent (Invitrogen, USA) and RNeasy MinElute Cleanup Kit (Qiagen, Germany) was used to extract the RNA. The flies were homogenized in 500 μL of Trizol reagent, then, a volume of 100 μL of chloroform was added into the mixture. The sample was thoroughly mixed and centrifuged at 10,000 xg for five minutes. A volume of 1000 μL of isopropanol was added into aqueous layer and thoroughly mixed. The sample was cleanup using MinElute Cleanup Kit according to manufacturer protocol. gDNA was removed using Turbo™ DNase Kit (Thermo Fisher Scientific, USA). The quality of extracted RNA was assessed by agarose gel electrophoresis, Nanodrop2000 (Thermo Fisher Scientific, USA), and Agilent2100 Bioanalyzer (Agilent, USA). High quality RNA (≥ 5 μg; ≥ 200 ng/μL; OD260/280 = 1.8–2.2) will be used for library construction.

Table 1

| Sample             | Accession number | Accession link                   |
|--------------------|------------------|----------------------------------|
| Day1 replicate 1   | SAMN18533764     | https://www.ncbi.nlm.nih.gov/biosample/18533764 |
| Day1 replicate 2   | SAMN18533765     | https://www.ncbi.nlm.nih.gov/biosample/18533765 |
| Day1 replicate 3   | SAMN18533766     | https://www.ncbi.nlm.nih.gov/biosample/18533766 |
| Day60 replicate 1  | SAMN18533767     | https://www.ncbi.nlm.nih.gov/biosample/18533767 |
| Day60 replicate 2  | SAMN18533768     | https://www.ncbi.nlm.nih.gov/biosample/18533768 |
| Day60 replicate 3  | SAMN18533769     | https://www.ncbi.nlm.nih.gov/biosample/18533769 |

Table 2

| Library            | %GC | Number of raw reads | Number of cleaned reads | Mapping rate |
|--------------------|-----|---------------------|-------------------------|--------------|
| Day1 replicate 1   | 45  | 49210530            | 48883744                | 96.7         |
| Day1 replicate 2   | 45  | 51138168            | 50773952                | 96.4         |
| Day1 replicate 3   | 44  | 49970634            | 49629740                | 92.4         |
| Day60 replicate 1  | 51  | 56678370            | 56334634                | 95.9         |
| Day60 replicate 2  | 50  | 50948514            | 50564212                | 95.7         |
| Day60 replicate 3  | 50  | 53317220            | 52898742                | 95.3         |
For library construction, standard Illumina protocol was employed. The first step involving the enrichment of mRNA using poly-T oligo attached magnetic beads. Then, the mRNA was fragmented using divalent cations. First strand cDNA synthesis was performed using SuperScript II followed by second strand. End repair was performed to remove any overhangs prior to adenylation of 3’ends. Then, adapter was ligated, and size selection (150–200 bp) was performed. The purified size-selected RNA was sequenced using Illumina HiSeq platform. Raw data generated was trimmed and cleaned by removing low quality reads and removing the adaptor.

2.3. Differential expression analysis

RNA-seq reads were aligned to the reference genome of *D. melanogaster* by using HISAT2 version 2.1.0 [1]. The genome was Drosophila_melanogaster.BDGP6.28.dna_sm.toplevel.fa.gz downloaded from Ensembl. Afterwards, in order to quantify the expression level of transcripts the alignment files generated by HISAT2 were used as inputs for featurecount [2]. These counts were then used as input for differential analysis using edgeR [3]. The statistical program edgeR
### Table 4

Enriched GO terms featuring cellular component. Significantly differentially expressed genes in day 60 versus day 1 are categorised into 25 GO terms featuring cellular component with significant of P-value < 0.05. The number of differentially expressed genes related to the GO terms are presented as count with their respective P-value.

| ID       | GO term                                      | Count | P-Value                      |
|----------|----------------------------------------------|-------|------------------------------|
| GO:0005634 | nucleus                                      | 1447  | 3.27917457291437E-20         |
| GO:0071011 | precatalytic spliceosome                     | 141   | 2.51427093226562E-12         |
| GO:0005737 | cytoplasm                                    | 1203  | 1.23883014505389E-11         |
| GO:0005875 | microtubule associated complex              | 260   | 3.97418659471114E-11         |
| GO:0071013 | catalytic step 2 spliceosome                | 122   | 1.36639696934484E-10         |
| GO:0005730 | nucleolus                                    | 150   | 6.93025838404035E-08         |
| GO:0012505 | endomembrane system                          | 177   | 2.8756207140668E-07          |
| GO:0005622 | intracellular                                | 245   | 4.95714684182731E-05         |
| GO:0005813 | centrosome                                   | 84    | 7.7186956064813E-05          |
| GO:0005739 | mitochondrion                               | 371   | 1.99536454552943E-04         |
| GO:0005681 | spliceosomal complex                         | 45    | 3.85332448204499E-04         |
| GO:0003532 | small nuclear ribonucleoprotein complex      | 34    | 5.51522366393124E-04         |
| GO:0000775 | chromosome, centromeric region               | 37    | 9.26713989307493E-04         |
| GO:0032040 | small-subunit processome                     | 32    | 9.89089142354131E-04         |
| GO:0005819 | spindle                                      | 45    | 0.001109693298146730         |
| GO:0005635 | nuclear envelope                             | 45    | 0.001109693298146730         |
| GO:0022625 | cytosolic large ribosomal subunit            | 52    | 0.001409411399342990         |
| GO:0005840 | ribosome                                     | 83    | 0.00177795170627050          |
| GO:0005654 | nucleoplasm                                  | 123   | 0.002370412221655130         |
| GO:0043234 | protein complex                              | 61    | 0.004578428094727970         |
| GO:0009222 | spindle pole                                 | 31    | 0.004752710283904030         |
| GO:0016020 | membrane                                     | 296   | 0.00699314564473070          |
| GO:0005747 | mitochondrial respiratory chain complex I    | 41    | 0.00733298819206281800       |
| GO:0005643 | nuclear pore                                 | 33    | 0.008068390352879900         |
| GO:0005912 | adherens junction                            | 44    | 0.008423004737610120         |

### Table 5

Enriched GO terms featuring molecular function. Significantly differentially expressed genes in day 60 compare to day 1 are categorised into 42 GO terms featuring molecular function with significant of P-value < 0.05. The number of differentially expressed genes related to the GO terms are presented as count with their respective P-value.

| ID       | GO term                                      | Count | P-Value                      |
|----------|----------------------------------------------|-------|------------------------------|
| GO:0005524 | ATP binding                                  | 612   | 1.05918346247829E-12         |
| GO:0005515 | protein binding                              | 527   | 5.96469126229741E-11         |
| GO:0003676 | nucleic acid binding                         | 359   | 1.02969701996623E-09         |
| GO:0008270 | zinc ion binding                             | 519   | 2.0315421977325E-09          |
| GO:0003723 | RNA binding                                  | 230   | 4.35245434691068E-08         |
| GO:0005509 | calcium ion binding                          | 181   | 1.49307405316839E-07         |
| GO:0046872 | metal ion binding                            | 519   | 2.59437961279962E-07         |
| GO:0003713 | transcription coactivator activity           | 54    | 3.138357146616E-05           |
| GO:0004386 | helicase activity                            | 47    | 5.6129988650304E-05          |
| GO:0003682 | chromatin binding                            | 111   | 6.63049880227043E-05         |
| GO:0000166 | nucleotide binding                           | 175   | 6.76405674789166E-05         |
| GO:0003729 | mRNA binding                                 | 143   | 1.0322440384053E-04          |
| GO:0008017 | microtubule binding                          | 91    | 4.19839962666167E-04         |
| GO:0004722 | protein serine/threonine phosphatase activity | 42    | 8.43623758676245E-04         |
| GO:0040004 | ATP-dependent RNA helicase activity          | 45    | 0.001167979760263700         |
| GO:003954  | NADH dehydrogenase activity                 | 31    | 0.001394636413593100         |
| GO:0044822 | poly(A) RNA binding                          | 63    | 0.0015759251370282100        |
| GO:0016887 | ATPase activity                              | 123   | 0.0016145164103400           |
| GO:0001104 | RNA polymerase II transcription cofactor activity | 32    | 0.00378153889304200          |
| GO:003714  | transcription corepressor activity           | 27    | 0.0044196473951650           |
| GO:0004842 | ubiquitin-protein transferase activity       | 143   | 0.00467428235042100          |
| GO:0003684 | damaged DNA binding                          | 26    | 0.005837637092831700         |
| GO:0019843 | rRNA binding                                 | 26    | 0.005837637092831700         |
| GO:0008134 | transcription factor binding                 | 72    | 0.012680395459747200         |

(continued on next page)
Table 5 (continued)

| ID          | GO term                                           | Count | P-Value       |
|-------------|---------------------------------------------------|-------|---------------|
| GO:0003743  | translation initiation factor activity            | 48    | 0.01586414858335640 |
| GO:0003924  | GTPase activity                                    | 106   | 0.017835528186287700 |
| GO:0003899  | DNA-directed RNA polymerase activity               | 26    | 0.0178667512397270 |
| GO:0051539  | 4 iron, 4 sulfur cluster binding                   | 26    | 0.0178667512397270 |
| GO:0003755  | peptidyl-prolyl cis-trans isomerase activity       | 29    | 0.02168192134654930 |
| GO:0016853  | isomerase activity                                 | 29    | 0.02168192134654930 |
| GO:0003705  | transcription factor activity, RNA polymerase II   | 46    | 0.023058171236249100 |
|            | distal enhancer sequence-specific binding          |       |               |
| GO:0051082  | unfolded protein binding                           | 45    | 0.0276767064792650 |
| GO:0042393  | histone binding                                    | 24    | 0.02925615687200600 |
| GO:0016740  | transferase activity                               | 31    | 0.030883614952758100 |
| GO:0003824  | catalytic activity                                 | 98    | 0.03357953732093580 |
| GO:0005484  | SNAP receptor activity                             | 23    | 0.037228197776595800 |
| GO:0004693  | cyclin-dependent protein serine/threonine kinase   | 15    | 0.03845359676428300 |
|            | activity                                           |       |               |
| GO:0030515  | snoRNA binding                                     | 15    | 0.03845359676428300 |
| GO:0001075  | transcription factor activity, RNA polymerase II   | 19    | 0.03922287285579800 |
|            | core promoter sequence-specific binding involved in |       |               |
|            | preinitiation complex assembly                     |       |               |
| GO:0042803  | protein homodimerization activity                  | 110   | 0.04323401342660000 |
| GO:0042623  | ATPase activity, coupled                           | 39    | 0.04556019342003400 |
| GO:0000977  | RNA polymerase II regulatory region                | 48    | 0.04871352258031190 |
|            | sequence-specific DNA binding                      |       |               |

Table 6
KEGG pathway enrichment analysis. 10 KEGG pathways are significantly enriched by differentially expressed genes in day 60 compared to day 1 with significant of P-value < 0.05. The number of differentially expressed genes related to the pathway are presented as count with their respective P-value.

| Term                                    | Count | P-Value       |
|-----------------------------------------|-------|---------------|
| Spliceosome                             | 110   | 5.73666837052742E-04 |
| DNA replication                         | 34    | 0.00347665731881500 |
| Nucleotide excision repair              | 37    | 0.00607462745663300 |
| Basal transcription factors             | 36    | 0.007531383638678190 |
| Protein processing in endoplasmic reticulum | 107   | 0.009490293668130400 |
| mRNA surveillance pathway               | 61    | 0.022952383725918000 |
| Mismatch repair                         | 20    | 0.02518504502307330 |
| Purine metabolism                       | 110   | 0.02557908482786070 |
| Fanconi anemia pathway                  | 25    | 0.02684084015079530 |
| Ubiquitin mediated proteolysis          | 82    | 0.0493859761510847 |

was analyzed in R/Bioconductor environment. FDR < 0.05 were set as the threshold for significantly differential expression genes [4].

2.4. GO classification and enrichment analysis

DAVID online tool was used to identify significantly enriched GO terms featuring biological process, cellular component, molecular function and KEGG pathways with corrected P-value less than 0.05 [5,6].

Ethics Statements

All animal handlings complied with guidelines set forth by the National Institutes of Health for the care and use of laboratory animals, and the protocol of this study followed the National
Institutes of Health guide for the care and use of laboratory animals (NIH Publications No. 8023, revised 1978) and Guide for the Care and Use of Laboratory Animals: Table 4 8th Edition.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

**CRediT Author Statement**

**Morteza Bajgiran:** Methodology, Resources, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing; **Azali Azlan:** Software, Data curation, Formal analysis; **Shaharum Shamsuddin:** Supervision, Funding acquisition; **Ghows Azzam:** Supervision, Funding acquisition; **Mardani Abdul Halim:** Conceptualization, Methodology, Resources, Investigation, Data curation, Writing – original draft, Supervision.

**Acknowledgments**

The authors would like to acknowledge financial support from Universiti Sains Malaysia and Malaysian Ministry of Higher Education grant (FRGS: 203.PPSK.6171226) and URICAS RU-Top Down Research Grant (1001/ PBIIOLOGI/870040), Universiti Sains Malaysia

**Supplementary materials**

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107413.

**References**

[1] D. Kim, et al., Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells, Nat. Methods 12 (3) (2015) 237–243, doi:10.1038/nmeth.3284.
[2] Y. Liao, G.K. Smyth, W. Shi, FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics 30 (7) (Apr. 2014) 923–930 doi:, doi: 10.1093/bioinformatics/btt656.
[3] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (Nov. 2009) 139–140 doi:, doi:10.1093/bioinformatics/btp616.
[4] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1) (Jan. 1995) 289–300 doi:, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[5] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (1) (2009) 44–57 doi:, doi: 10.1038/nprot.2008.211.
[6] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res. 37 (1) (2009) 1–13 doi:, doi:10.1093/nar/gkn923.