



Norwegian University
of Life Sciences

Master's Thesis 2021 60 ECTS
Faculty of Biosciences

Developing a workflow for the multi-omics analysis of Daphnia

Wanxin Lai
Genome Science

Acknowledgement

I would like to express my heartfelt gratitude to my supervisor Prof Torgeir R. Hvidsten for his dedicated support, precious advice, and most importantly, his academic spirit which has inspired and assisted me throughout this project. Special appreciation is also extended to my co-supervisor Prof Knut-Erik Tollefsen for his valuable guidance and insights. Also, a big thanks to Dr You Song for his overall support. Last but not least, I would also like to thank my family for their love, tolerance and accommodation whenever I was stressed and caught up in my work. This endeavour would never have been the same without any of them.

30th October 2021

Abstract

In the era of multi-omics, making reasonable statistical inferences through data integration is challenged by data heterogeneity, dimensionality constraints, and data harmonization. The biological system is presumed to function as a network where the physical relationships between genes (nodes) are represented by links (edges) connecting genes that interact. This thesis aims to develop a new and efficient workflow to analyse non-model organism multi-omics data for researchers who are entangled in the biology questions by using readily available software tools. The proposed approach was applied to the transcriptome and metabolome data of *Daphnia magna* under various dose rates of gamma radiation.

The first part of this workflow compares and contrasts the transcriptional regulation of short-and long-term gamma radiation exposure. A group of genes which share a similar expression across different samples under the same conditions are known as modules, because they are likely to be functionally relevant. Modules were identified using WGCNA but biologically meaningful modules (significant modules) were selected through a novel approach that associates genes with significantly altered expression levels as a result of radiation (i.e. differentially expressed genes) with these candidate modules. Dynamic transcriptional regulation was modelled using transcription factor (TF) DNA binding patterns to associate TFs with expression responses captured by the modules. The biological functions of significant modules and their TF regulators were verified with functional annotations and mapped into the proposed Adverse Outcome Pathways (AOP) of *D. magna*, which describes the key events which contribute to fecundity reduction. The findings demonstrate that short term radiation impacts are entirely different from long term and cannot be used for long term prediction.

The second part investigates the coordination of gene expression and metabolites with differential abundances induced by different gamma dose rates and the underlying mechanisms contributing to the varying extent of the reduction in fecundity. Significant modules which belong to the same design model of dose rates were combined and annotated with new functionality. The abundance of metabolites was also modelled with the same design model. Integrated pathway enrichment analysis was performed to discover and create pathway diagrams for visualising the multi-omics output. Finally, the performance of this workflow on explaining the reduction of fecundity of *D. magna*, which has not been described in previous studies, has been evaluated. Combining the information from the metabolome and transcriptome data, new insights suggest that the alteration to the cell cycle is the underlying mechanism contributing to the varying reduction of fecundity under the effect of different dose rates of radiation.

Contents

Figures	7
Tables	10
Chapter 1 Introduction	12
1.1 Background.....	14
1.1.1 Omics and the central dogma of molecular biology	14
1.1.2 Integration of transcriptomics and metabolomics.....	15
1.1.3 Transcription Factor Binding Sites (TFBS).....	15
1.1.4 Machine learning	15
1.1.5 Differential testing with counts data transformations.....	20
1.1.6 Functional enrichment analysis.....	21
1.2 Aims of the thesis	22
Chapter 2 Materials and methods	25
2.1 Culture conditions and exposure studies with <i>Daphnia magna</i>	26
2.1.1 Maintenance of Daphnia culture.....	26
2.1.2 Laboratory settings for gamma radiation.....	26
2.1.3 RNA extraction.....	27
2.1.4 Next Generation Sequencing and Library Preparation	27
2.1.5 Metabolite extraction.....	27
2.1.6 GC-MS method and raw data preparation.....	28
2.2 Data analysis and pipelines	29
2.2.1 Data pre-processing.....	30
2.2.2 Identification of modules and DEGs	30
2.2.3 Identification and detection of biologically meaningful modules.....	31
2.2.4 Conversion of identifiers	31
2.2.5 Detection of biologically meaningful modules and construction of transcriptional regulatory network.....	32
2.2.6 Integration of metabolomics and transcriptomics data.....	33
Chapter 3 Results	35
3.1 Identification of dose rate responsive co-expression modules and transcriptional regulatory networks.....	35

3.1.1 Initial assessment on gene expression profile with PCA- and scatterplots	35
3.1.2 Creation of a co-expression network	37
3.1.3 Differential expression analysis.....	40
3.1.4 Modules with non-random association with DEGS.....	40
3.1.5 Whole proteome comparison and ortholog identification shared by Drosophila and Daphnia.....	42
3.1.6 Mapping of transcripts to gene identifiers.....	44
3.1.7 Pathway and GO Enrichment Analysis (Gene modules).....	45
3.1.8 Transcriptional regulation by active transcription factors.....	46
3.1.9 Transcriptional regulatory network between 4 days and 8 days exposure to gamma radiation	48
3.1.10 Module similarity assessment to predict the outcome of 8 days of exposure to gamma radiation based on 4 days.....	50
3.1.11 AOP Integration	51
3.2 Multiomics integration of differentially expressed and co-expressed genes (DEACGS) with metabolite profiles.....	53
3.2.1 Differential expression analysis of metabolomics.....	53
3.2.2 Generate DEACGs corresponding to the design models.....	55
3.2.3 Integrative pathway analysis (Paintomics3) (DEACGs and metabolites)	56
3.2.4 Pathway and GO enrichment analysis (DEACGs)	59
Chapter 4 Discussion.....	63
4.1 Analysis and integration of data (technical discussion).....	63
4.1.1 Non-conventional modules selection.....	63
4.1.2 Software selections	64
4.1.3 Design model	64
4.1.4 The conversion of identifiers and loss of information	65
4.1.5 Significant modules in the network of transcriptional regulation	66
4.1.6 Choices on motif database in AME.....	67
4.1.7 Confounding variables in the metabolomics data	68
4.1.8 Conventional linear regression on the metabolomics data	68
4.1.9 Creating DEACGs based on design model	68
4.2 Network based inter-modular transcriptional regulation	69

4.2.1 Networks of TF regulation after 4 days of radiation exposure	69
4.2.2 Investigating a cyclical relationship in the 4 days regulatory network	71
4.2.3 Module specific findings from a longer exposure to gamma radiation	74
4.3 Limitation of GO-based module integration with AOPs and the new prospects	76
4.4 Summary toxicity pathways: Integrating the response of differential metabolites with differential transcriptomics pathways	79
4.5 Contributions and future prospect: Integrating multi-omics revealed an altered nature, prioritizing survivorship over reproduction	87
4.5.1 The energy consumption of DNA repair enzymes and regaining homeostasis	87
4.5.2 Perturbances from cell cycle arrest and the maintenance of genomic stability.....	88
4.5.3 Accelerated cellular metabolism	88
Chapter 5 Summary	90
Chapter 6 Supplementary Data	92
S1 Functional Enrichment – modules from 4 days data.....	92
S2 TF gene expression profiles – 4days	95
S3 Functional Enrichment - modules from 8 days data	97
S4 TF expression profiles – 8days	100
References	103

Figures

Figure 1: The network of Adverse Outcome Pathways (AOPs) demonstrates the impact of excessive ROS recently proposed (Song et al., 2020).	13
Figure 2: Overall workflow for the multi-omics data analysis. The colour partition separates the workflow into Part 1 and Part 2.....	25
Figure 3: Overview of software in each step of data analysis. R commands are shown in italic. 29	
Figure 4: Scatterplot showing the log2 fold changes of the same genes exposed to 4 and 8 days of gamma radiation. Exclusive DEGs for 8 days are marked in yellow, exclusive DEGs for 4 days are marked in blue and DEGs present in both are marked in red.	37
Figure 5: WGCNA. Diagnostic plots showing various beta fits to reach a scale-free topology network. Analysis of scale-free network topology using different soft-thresholding power on 4 days transcriptomics data. On the upper left, A shows the scale free fit index (y-axis) and the upper right (B) shows the mean connectivity (degree, y-axis). C shows the numbers of genes in every module, which was given an arbitrary colour.	40
Figure 6: The bar plots show the association between WGCNA and DESeq2 after Fisher's Exact Test. A & B: Association of gene module with the linear model. Modules in red passed the p-value cut-off < 0.05. C, D, E & F: Association of gene module with the linear combination (contrast) model. Modules marked with asterisk (*) passed the significant p-value cut-off < 0.05. The 4 days data yielded 11 significant modules while the 8 days data yielded 12.	42
Figure 7: Comparison of orthologues genes between different clones of <i>Daphnia magna</i> , <i>Daphnia pulex</i> and <i>Drosophila melanogaster</i> . A: Venn diagram showing the numbers of shared orthologous groups between <i>D. pulex</i> , <i>D. magna xinb3</i> , <i>D. magna KIT</i> and <i>D. melanogaster</i> . B: The bar graph above shows the numbers of protein clusters found in each species, while the bar plot below displays the number of orthologous clusters shared by 1, 2, 3 and 4 species. C: Pairwise heatmap with number of overlapping clusters between different pairs of species. The overlapping cluster numbers were indicated in the cells and the colour intensity followed the shared number of orthologous groups: the darker the colour, the more orthologs shared between species.	43
Figure 8: Venn diagram comparing GO terms and Reactome pathway between 4 days and 8 days data. The upper diagrams (8A,8B,8C) correspond to the GO domain: biological processes, cellular components, and molecular functions. Diagram 8D shows the difference of radiation affected pathways between 4 and 8 days of gamma radiation exposure arranged in a descending order according to the number of genes.	46
Figure 9: Network showing the changes of transcripts abundance and the eigengene expression across different dose rates. The size of the nodes corresponds to the number of enriched motifs. The pointing direction of the arrows indicate the regulation of TFs on the targeted node. The transcriptional regulatory direction is highly focussed on the central module, blue, and shows that it is the key module as it contains the most motif binding sites.	49

Figure 10: Network showing the changes of transcripts abundance and the eigengene expression across different dose rates. Size of the nodes are corresponding to the numbers of enriched motifs. The pointing direction of the arrows indicates the regulation of TF on the targeted node. The interactions between most modules were bidirectional.	50
Figure 11: The eigengene expression of the 4 and 8days module networks show an entirely different regulatory relationship. 4d and 8d in the end of every row name represents the radiation exposure period of each module.	51
Figure 12: Plots show the initial analysis on metabolites abundance. A: PCA plot showed the sample of metabolites in 2D plane spanned by the first two principal components which explained the most variance. No clustering pattern observed indicates an extremely small difference between samples. B: Venn diagram comparing the numbers of DEMs shared and uniquely existing between low dose-responsive, high-dose responsive and linear model groups. C: Heatmap showing the gene expression of all metabolites and all samples; red to blue colour scale represents high to low gene expression and the colour of dose rates was represented by the legend on the right.	54
Figure 13: Heatmap showing gene expression profiles of all significant modules and all samples, red to blue colour scale represent high to low expression and the colour of dose rates was represented by the legend on the right.	56
Figure 14: Parts of the output from Paintomics3. A: Pie chart demonstrating the pathway categories resulting from the overall transcriptomics and metabolites data. B: Pathway network from low-dose rate responsive group, the node represents the names of the pathway, and the edges show the shared features (DEACGs) between pathways. C, D & E: Pathway enrichment results from low dose-, high dose rate responsive group and linear model. Tables of enriched pathways are ordered in ascending order of P-value, the red colour intensity changes according to the level of enrichment/significance and the grey scale means no corresponding omics data was found in the pathway.	57
Figure 15: Network of top over-represented pathways and features genes associated with (A) low level of dose rates exposure, (B) high level of dose rates exposure and (C) linear response to gamma radiation. Nodes coloured in red, and green indicate whether the log fold change is positive or negative; the name of the node shown where the gene symbol corresponds to Entrez ID; the central node coloured in cream shows the name of the pathway with the size representing the number of genes involved in the pathway.	60
Figure 16: GO enrichment analysis made with the Cytoscape plugins BiNGO, EnrichmentMap and AutoAnnotate. Nodes represent the enriched GO term; node size corresponds to the number of genes and the thickness of edge depicts the number of overlapping genes	62
Figure 17: Significant modules that are excluded from the regulatory networks. The module with black border is from 4 days data while the rest are from 8 days data (top left).	67
Figure 18: Top over-represented pathways from Reactome pathway analysis in blue-, pink-, turquoise- and red modules.	71
Figure 19: The gene expression of NK7.1 from the blue modules (left) and the output from GO analysis (right).....	72

Figure 20: The gene expression profile of ken and barbie from the turquoise and lightcyan modules.....	73
Figure 21: The expression profile of TF encoding genes from the turquoise-, red- and pink modules.....	73
Figure 22: The expression profile of TF encoding genes from the green and black module.	74
Figure 23: Top enriched pathways from Paintomics3: oxidative phosphorylation(left), TCA cycle(right). Map of the most significantly enriched pathway, blue shades represent down-regulated genes, red shades represent up-regulated genes, and the DEACGs are enclosed in a thicker border.	80
Figure 24: Enriched transcriptomics and metabolomics pathways from Paintomics3 output. A: Biosynthesis of amino acid. B: Valine, leucine and isoleucine degradation. C: glutathione metabolism. D: fatty acid biosynthesis.....	82
Figure 25: The relationship between Glutathione (GSH) synthesis, Tricarboxylic acid cycle (TCA) and methionine metabolism as the critical alteration in carbon metabolism.	83
Figure 26: Schematic of purine and pyrimidine metabolism regulated by PI3K/Akt signaling. The level of phosphoribosylpyrophosphate (PRPP) is the key substrate regulated by PI3K/Akt.	85

Tables

Table 1: The number of differentially expressed genes generated by DESeq2. Except for Ctrl vs 1 which is a low dose rate- responsive group, the data from 8 days has more DEGs than from 4 days in the high dose-rate responsive group (1 vs 100) and the linear model.	40
Table 2: Summary of qualified (significant) modules having an overlap between WGCNA and DESeq2. Double asterisks (**) indicate modules that are exclusively found at the intersection of the specific design model ('linear model', 'Ctrl vs 1' or '1 vs 100') and WGCNA modules. Pounds sign (£) refers to those that are at the intersection of linear combinations (exists in both 'Ctrl vs 1' and '1 vs 100') and WGCNA. Modules with pound sign only found in 4 days data but not in the 8 days data.....	40
Table 3: Mapping of identifiers from expressed transcripts of <i>D. magna</i> to Entrez IDs of <i>D. melanogaster</i> and to KO ID of <i>D. pulex</i> . Entrez ID was chosen for the mapping of ID from <i>D. magna</i> to <i>D. melanogaster</i> because it retained more transcripts than KO IDs from <i>D. pulex</i>	44
Table 4: TF orthologs and their corresponding modules from the data of 4 days of gamma radiation exposure. Gene symbol and gene names followed the nomenclatures of <i>D. melanogaster</i> as documented in Flybase. TFs are considered activated if a corresponding enriched motif was found and the ortholog genes which encoded for the TFs were present in that module.	47
Table 5: TF orthologs and their corresponding modules from the data of 8 days of gamma radiation exposure. TFs are considered activated if a corresponding enriched motif was found and the ortholog genes which encoded for the TFs were present in that module.....	47
Table 6: Integration of significant modules with key events derived from AOP for the 4 days- transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a 'tick' to represent the presence of the gene in a specific significant module.....	52
Table 7: Integration of significant modules with key events derived from AOP for 8 days- transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a 'tick' to represent the presence of the gene in a specific significant module.....	52
Table 8: The number of differentially expressed metabolites from each group of the design model. The linear expression group has the highest number of metabolites, but it also shares about half of them (62 from Figure 12B) with the high dose rate responsive group.	55
Table 9: Significant modules from the previous chapter and their corresponding groups. Double asterisk (**) indicates modules that are exclusively found in the design model. The low dose rate responsive group has more exclusive modules than the linear model- and high dose rate responsive- groups.....	55
Table 10: Significant modules and the values in the bracket indicate the numbers of transcript paired with Entrez ID. Double asterisk (**) indicates modules that are exclusively found in the design model. The linear model- and high dose rate responsive- group share many modules,	

with the green module containing the highest number of genes. The blue module has the highest number of genes overall and it is exclusive to the linear model group. 57

Table 11: The number of unique Entrez ID from each dose group and their composition to up- and down-regulation. The number of enriched pathways from Reactome PA is shown in the last column. The low-dose responsive group contains mostly up-regulated DEACGs in the enriched pathway, the high dose rate responsive group contains a similar amount of DEACGs in terms of the directionality, lastly the linear model consists of the most down-regulated DEACGs in the top enriched pathways. 60

Table 12: Numbers of genes in each module. 67

Table 13: Alternative keywords used in the search for Key events..... 77

Table 14: Integration of significant modules with potential key events for the 4 days-transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a ‘tick’ to represent the presence of the gene. Hashtag (#) indicates that alternative terms have been used in the search for key events. 78

Table 15: Integration of significant modules with potential key events for the 8 days-transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a ‘tick’ to represent the presence of the gene. Hashtag (#) indicates that alternative terms have been used in the search for key events. 78

Chapter 1 Introduction

Organisms are consistently exposed to oxidative stress, ranging from exposure to ultraviolet rays from the sun and anthropogenic activities to medical treatments destroying cancerous cells with radiotherapy. As a result, cells are equipped with a healing mechanism to maintain the integrity of the genome and prevent the onset of tumorigenesis. The reactive oxygen species (ROS) is a group of oxygen-derived free radicals generated, mainly, through aerobic respiration in mitochondria. Under the influence of environmental stresses, such as ionizing radiation, the cell's interactions with water through radiolysis can lead to an uncontrollable amount of ROS generation and ineffective elimination. Excessive accumulation of ROS lets the highly reactive, unpaired electrons from the radicals damage all macromolecules including DNA, which can famously lead to DNA lesions. Regardless of cell type, a single diploid cell exposed to 1Gy of gamma radiation was reported to cause approximately 1000 single-strand breaks, 40 double strand breaks and alterations of bases (Olive, 1998). This however is almost negligible compared to the roughly 200,000 single strand breaks which occur daily within mammalian cells (Billen, 1990). However, DNA lesions induced by radiation are more complicated than endogenous damage within a cell because it can lead to cluster lesions, double stranded breaks with heterogeneous ends, and clusters of non-double-strand break lesions which are more likely to trigger apoptosis (Olive, 1998). The self-repair response varies with species and cell type with factors including chromatin structure, spectrum of lesion, availability of repair proteins, gene induction on cell cycle control and the cellular environment. A study on the effects of ionizing radiation on spermatozoa reported that the average mutation rate in mice is higher than in *Drosophila* by approximately 10 to 15 times, and the self-repairing process was postulated to be dose rate dependent (Alpen, 1998).

There are growing concerns around the early life stages of living organisms exposed to ionizing radiation because cells are actively dividing, proliferating, and differentiating at these times. Stages such as gametogenesis, embryogenesis, and organogenesis have been primarily targeted for the study of low-dosage radiation induced by-stander effects, adaptive response and genome instability (Streffer, 2004). Studies of the first two effects show that small amounts of radiation exposure stimulate cellular response and develops better radioactive resistance but that the latter causes chromosomal mutation and uncontrollable gene expression, which could lead to permanent toxic impact in life and even be inherited by future generations (Nations, 2000, Streffer, 2004).

The discharge of industrial and municipal waste into the water environment occurs in a continuous manner all around the world. Aquatic life in the food web ranging from producers (phytoplankton, algae) to consumers (fish, shrimps) are constantly exposed to the radioactive waste chronically before they are ingested by the higher consumers (bears, birds, humans) in the food web. Based on acute radiation exposure to adult organisms, benchmarks lower than 0.42 mGy/h and even 0.01 mGy/h were proposed as the maximum ecotoxicological assessment value

confidently resulting in no adverse effects (Nations, 2000, Garnier-Laplace et al., 2010). Study on chronic sublethal effects in the early development life stage of aquatic invertebrates has slowly gained the interest of academia in the last five years, especially for organisms from the family of Daphniidae. Due to practical benefits such as ecological habitat in a wide variety of water bodies, short life spans and colourless bodies, it has regularly been used in toxicity and hazard assessments (Oecd, 2008, Cui et al., 2017).

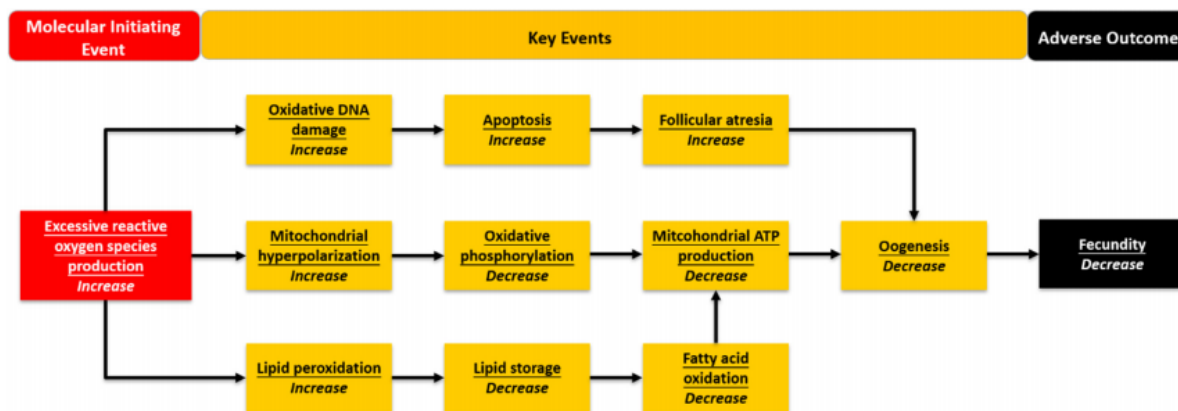


Figure 1: The network of Adverse Outcome Pathways (AOPs) demonstrates the impact of excessive ROS recently proposed (Song et al., 2020).

The network of adverse outcome pathways (Figure 1) describes the mode of actions which contributes to the decrease in fecundity for *Daphnia magna*, exposed to moderately to high gamma radiation (1-100 mGy/h), has recently been proposed (Song et al., 2020). The production of excessive ROS as a molecular initiating event causes a series of key events that eventually lead to the reduction of fecundity. These adverse outcome pathways (AOP) also cause DNA damage, ATP depletion and lipid peroxidation-associated reproduction dysfunction. This project focuses on developing a systematic workflow to evaluate the adverse outcome pathways using two transcriptomic datasets of juvenile *D. magna* measuring gene expression at different dose rates and after 4 days and 8 days of exposure to gamma radiation (Figure 1) as described in detail by Song et al. (2020), and finally presented as AOPs (<https://aopwiki.org/>, AOP #216, #238 and #311)

The preliminary findings (Song et al., 2020) also suggests that different mechanisms were activated upon different dose rates which contributed to varying degrees of reduced fecundity. At the low dose rate (1 mGy/h) reproduction was delayed, whereas at the high dose rate (100 mGy/h) the number of progenies and the brood size were reduced. With the transcriptomic and metabolomic data of *D. magna* exposed to 8 days of gamma radiation provided by Norwegian Institute for Water Research (NIVA), this projects also intends to perform an integrated analysis to refine the understanding of toxicity pathways and the early impact on the reduction in fecundity. This may enhance the biological knowledge of gamma radiation in the biota of other species.

1.1 Background

1.1.1 Omics and the central dogma of molecular biology

It has long been known that the central dogma of molecular biology describes the cellular flow of genetic information from DNA, transcribed into RNA, and translated into protein, with metabolites as the products of cellular metabolism. In the era of multi-omics, the “central dogma” is proposed to be a single layer of macroscopic integration of information, from an omics wide perspective (dos Santos et al., 2021). Metabolomics should be placed at the centre controlling the changes to other fields, such as epigenomics, genomics, proteomics, transcriptomics, because the metabolite states and constituents can reveal the mode of action of these tiny molecules intracellularly and extracellularly. Nonetheless, macro and micro molecules constantly work together in every biological process. The biological pathways linking these processes are highly chaotic where the smallest changes in any omics field at the molecular levels can cause radically different outcomes (Kellert and Sklar, 1997). The emergence of multi-omics technologies has created chances to monitor and study the biological system from different omics perspectives with each type offering a unique view. Conducting multi-omics research can therefore provide a global view which will benefit research in biology.

The term “omics-” describes a field of molecular study in biological science. In this study, transcriptomics and metabolomics data are used in the analysis. **Transcriptomics** is the study of transcriptomes, which are a set of all RNA molecules decoded from DNA (a genome) and generally refer to messenger RNAs (mRNA) (dos Santos et al., 2021). The level of gene expression is equivalent to the level of mRNA, while the term differentially expressed genes (DEGs) refers to genes with a significant difference of expression when exposed to different conditions. The expression of genes in the coding region is known as mRNA transcription, and this harbours all the necessary information responsible for protein synthesis acting as a blueprint.

The metabolome of a species is the collection of all small, lightweight molecules known as metabolites (Sun and Hu, 2016). **Metabolomics** refers to the study of these tiny substrates and products which are involved in metabolism. It is also a promising approach which directly reflects the physiological state of cellular activity, and the underlying biochemistry is strongly related to molecular phenotypes. Metabolomics is a crucial piece of the puzzle in omics organisation as it is more closely related to the phenotypes than the other omics fields. Metabolites and metabolic fluxes represent the end products from upstream regulation in particular conditions and moments. Monitoring the level of metabolomes hence revealed the transition of physiological states, even on the smallest intra- and extra- cellular perturbations.

1.1.2 Integration of transcriptomics and metabolomics

The crosstalk between metabolites and gene expression has also been interpreted as intracellular signalling in transcriptional regulation (Donati et al., 2018). The canonical flow of genetic information started from the regulation of transcription factor to the end products, metabolites, which are then conveyed metabolic feedback to re-interact and change the activity of transcription factors. Macromolecules like sugars and amino acids are metabolites that not only serve as signalling molecules, but also as the building blocks and precursors of other metabolites in response to external stress stimuli. Hence, integrating the transcriptomics and metabolomics data provides a deeper set of information elucidating this bidirectional relationship, and enables better prediction of responses to the mode of action triggered by gamma radiation.

1.1.3 Transcription Factor Binding Sites (TFBS)

Transcription factors (TFs) are proteins involved in the regulation of transcription through DNA binding activity in the upstream of a transcription start site (TSS) or bind to the enhancers which located far away from the TSS (Lawler, 2010). TFs can either refer to a transcriptional activator or repressor which binds specifically to the regions of a gene promoter or a distal region from the TSS. A DNA binding motif is a known specific binding site of TFs. Unlike prokaryotes (microscopic single-celled organisms that have neither a distinct nucleus with a membrane nor other specialized organelles), regulation of gene expression in eukaryotes (those that do have a distinct nucleus and nuclear membrane) usually requires more than one TF working in a combinatorial manner under different conditions (Cole, 2016). The regulatory networks constructed in this thesis propose a complex combinatorial transcriptional control involving more than one TF in a feedback loop within and beyond the clusters of interest, focussing only on TFs from the upstream region on TSS.

1.1.4 Machine learning

The advent of machine learning allows for the promising discovery of hidden structures within sets of highly complex omics data. Machine learning is a broad term that refers to the method of fitting a predictive model to a dataset with a large number of features through the steps of identifying new patterns, recognizing existing patterns and making decisions. The first type of machine learning used is supervised learning, this is creating a predictive model which fits the data given by feeding some ground truths (true labels) into the training so the predictive models generated can predict the answers in a test set as accurately as possible without overfitting. Overfitting is when a system attaches itself too strongly to a particular set of data, for example a model that predicts the number of days in a year and is given the years 2017, 2018, 2019 will give wrong results when trying to predict for 2020 (a leap year). Unsupervised learning identifies patterns without using predetermined labels. If building a model requires accurate, efficient,

cost-effective, and unbiased annotations; then combining limited amounts of labelled data with unlimited unlabelled data can be a powerful solution. This is known as semi-supervised learning.

1.1.4.1 Unsupervised learning - Weighted gene co-expression network analysis (WGCNA)

The Weighted Genes Co-Expression Network Analysis (WGCNA) software package provides a comprehensive collection of R functions for construction of networks, identification of gene clusters (modules), calculations of topological properties (e.g., network centrality), simulation, and visualization of data (Langfelder and Horvath, 2008). WGCNA refers genes, transcripts or protein as nodes. In general, the default setting of Pearson's correlation measures the magnitude of the genes/transcripts co-expression values by evaluating the linear correlation between node pairs, in which a high correlation coefficient suggests that these genes are likely to co-regulate in the same biological process and therefore potentially possess similar functionality. After the construction of the correlation network, WGCNA employs unsupervised clustering for module detection by using topological overlap measures (TOM) as a proximity measure of network interconnectedness. The creation of TOM is to reduce the sensitivity of a network towards random or missing connections resulting from noise. A TOM matrix includes the adjacency of two nodes and the numbers of 1-step neighbours shared between two nodes. This further creates a dissimilarity matrix, $\text{disstOM} = 1 - \text{TOM}$ which serves as an input to average linkage in hierarchical clustering for grouping highly co-expressed nodes into modules.

Hierarchical clustering is a common method that works well in high dimensional data because it provides visualization and does not limit or specify the number of clusters. However, it is hard to control the number of clusters and cluster size generated, especially when capturing prominent clusters from a complicated tree through a typical pruning method such as static cut (Langfelder and Horvath, 2008). Various tree cutting techniques are available but the default in WGCNA is Dynamic Tree Cut as it takes the shape criteria (core scatter, branch gaps and cluster size) into consideration. When examining whether the detected modules are biologically meaningful or simply a technical artifact/contamination, gene ontology (GO) enrichment analysis is very useful. The representative (weighted average) expression of a given module is called a module eigengene, which is also known as the first principal component in a principle component analysis (PCA).

Two common options used to identify condition-associated modules are 1) Correlating module eigengene with traits/disease/treatment of interest to associate the module with biological meaning, 2) measuring the gene significance (GS) (correlating genes with traits/disease of interest) and the module membership (MM) (correlating genes with selected module eigengene) and setting a cut off of at least $\text{MM} > 0.3$ and $\text{GS} > 0.3$ for module selection. However, neither of these options were chosen in this study because the non-linear relationship between modules and traits will not be detected. This study creatively includes the advantages of

supervised learning using software package DESeq2 to select differentially expressed modules. The details are addressed in the following chapter.

1.1.4.2 Unsupervised learning - Automatic network construction and blockwise module detection

The WGCNA workflow that is presented in this study implements the function `blockwiseModules` to build the correlation network and cluster tree, determine, and merge modules with highly correlated eigengenes automatically (and quickly). This function overcomes the limitations of memory size and processor speed when handling large datasets by separating the scalar variables into clusters beforehand. After pre-clustering of nodes into large blocks (a variant of k-means clustering, hierarchical clustering, and automatic module detection), merging is carried out on each block. Module membership is hence recalculated. However, it is recommended to always keep a dataset in one block of memory.

Other parameters used in the function `blockwisemodule` in this workflow were:

1. `networkType = Signed`
2. `TOMType = Signed`

The standard workflow of WGCNA requires an adjacency matrix which is then converted into a topological overlap matrix (TOM). There are two options for building any of these networks, signed or unsigned treatment of pairs of nodes in a weighted correlation network. An unsigned network defines the relations of two genes in a pair of nodes as the absolute value of Pearson correlation. It means that the sign of the correlation does not matter, positive and negative correlations are treated equally. However, in the context of a gene expression study, positive correlation and negative correlation between nodes/genes/transcripts imply different node profiles as in where and when the genes are up- or down- regulated, mixing the signs will simply ignore this piece of information. Moreover, negatively correlated nodes could belong to a different biological category than the positively correlated network, which is usually the case. Signed networks preserve the sign information and scale the correlation interval from $[-1, 1]$ into $[0,1]$ in an adjacency matrix. Previous study of embryonic stem cells show that signed network generated modules outperformed unsigned networks by capturing more specific gene expression patterns, despite negatively correlated nodes being classified as unconnected (Mason et al., 2009). Therefore, the author recommends using signed networks due to its simplicity in biological interpretation and to retain the underlying correlations of node pairs.

3. Setting a threshold for scale-free topology

It is assumed that the biological network resembles a scale free graph and the correlation between genes vary. The topology of a scale free network is dominated by a few centralised, highly connected genes (hub genes) whereas the rest of the genes (most genes) have significantly fewer neighbours compared to the hub genes. Hence, the scale free network also refers to

networks that follow a power law distribution, in which the average degree of node k (number of connections to the nearest neighbours) within a network is not representative, despite having finite numbers of genes. But the variation in the first neighbours $P(k)$ is a proportional change that varies as a power of the average degree (Zhang and Horvath, 2005).

$$P(k) \sim k^{-\gamma}$$

In this workflow, a soft threshold is chosen over a hard threshold because the idea is to focus on strongly correlated genes over weakly correlated genes but without losing them. Hard thresholding turns absolute value (coefficients) which are lower than the threshold into zero. Soft thresholding does the same, however it also shrinks the passing values towards zero. Weak interactions between genes that are equivalent to noise are represented by smaller coefficients whereas the stronger interactions are those with larger coefficients. Choosing a soft threshold thus preserves gene pairs with weak interactions and magnifies gene pairs with strong interactions, the idea is to resemble the continuous nature of gene interactions in a biological system in real life, where all genes are connected in principle, and the strength of connections differs from time to time when different responses or reactions are triggered. The value of soft thresholding (β) is used to raise the similarity between genes by powering the correlation coefficients. The choice on β affects the degree of scale free index (R^2) (Zhang and Horvath, 2005).

Scale-free topology fit index or scale-free index (R^2) describes how good a network fulfils a scale free topology. R^2 comes from the squared correlation of degree distribution $\log P(k)$ and average degree, $\log(k)$. However, there is a trade-off between maximizing the R^2 and maintaining a high k . So, $R^2 > 0.80$ is a good rule of thumb, because an R^2 approaching 1 indicates a very good fit which is close enough to achieve scale free topology (Langfelder and Horvath, 2008). Meanwhile a mean connectivity, not more than a few hundred, but high enough to be informative is important to ensure the detection of meaningful modules.

1.1.4.3 Supervised learning - Regression analysis

Simple linear regression can be considered as the simplest form of supervised learning because it aims to explain the observed response variable with an explanatory variable. The explanatory variable can be represented by features whereas the responding variable is the continuous response. The linear regression algorithm will attempt to fit an arbitrary straight line as close to as many data points (samples) as possible. After multiple iterations, the best-fit line is found and the resulting distance between the line and data points should be minimal. The predicted quantitative response and errors for future data relies on the predicted coefficients, which shows how well the linear model fits the data. In transcriptomics, it is challenging to model sequencing counts and control the prediction performance based on the expression of genes due to treatment effects. The expression of genes due to extraneous sources such as experimental manipulations, known and unknown technical variability as well as unknown biological variations make it challenging to separate the variable of interest from these interferences. Software like DESeq2 can come to the rescue by normalising gene counts associated with the changes between

the conditions by using a more advanced regression algorithm, the negative binomial generalized linear model before fitting of the linear model.

1.1.4.4 Supervised learning - Differential expression analysis on discrete data with DESeq2

Transcriptomics data is discrete data, so it requires a software package which is designed for this data type, such as DESeq2. The statistical model of DESeq2 assumes that most genes are not differentially expressed (DE), so the null hypothesis is true when the log₂ fold changes of a gene is 0 (Love et al., 2014). Firstly, the transcript's raw counts, Y_{ij} of gene i from sample j , were modelled internally following a negative binomial (NB) distribution $Y_{ij} \sim \text{NB}(\mu_{ij}, \sigma^2_{ij})$ where μ_{ij} is the fitted mean scaled from normalization and σ_{ij} is the dispersion.

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = X_j \beta_i$$

s_j represents the size factor for normalization and q_{ij} is proportional to the transcript's abundance of sample j . The application of general linear model (GLM) exponentiates the predictors, whereas coefficient β_i shows the log₂ transformed fold changes for gene i and vector x_j indicating the design matrix elements for sample j . The expected counts follow $\log(E(Y_{ij})) = X_j \beta_i + \log(s_j)$. The simplest comparison of the same gene between two groups in different conditions, g_1 and g_2 , is depicted in the hypothesis testing differential expression shown below:

$$H_0: \beta_{g_1} = \beta_{g_2}$$

$$H_1: \beta_{g_1} \neq \beta_{g_2}$$

Secondly, the coefficient and dispersion parameters (mean and variance) are estimated with a Bayesian shrinkage which shares the information across all genes.

Thirdly, counts from each gene are fitted into the negative binomial GLM followed by performing a Wald test to identify DE genes. Typically, the p-values < 0.05 indicates 5 % chance that the finding results in false positives. As the numbers of genes increase, the false positive rate are correspondingly inflated and therefore in this study, DE genes are filtered only according to the criteria of multiple testing. The correcting method, Benjamini-Hochberg (FDR) is applied in which the genes are ranked according to the p-value, the rank is then divided by the total number of tests and multiply with the false discovery rate of interest.

The normalization implemented by the DESeq2 package accounts for the sequencing depth (the average number of times a portion from the total nucleotides is sequenced), gene length (different genes differ in length but which have the same expression level) and RNA composition (the varied expression of the same gene in different samples; the high counts from highly

expressed gene can mask the counts of other differentially expressed genes, etc.). DESeq2 focusses on the samples comparability by normalizing the counts with a median-of-ratios method. This approach divides the counts by a sample specific scaling factor, which is also the median ratio of each sample relative to the geometric mean per gene. Non-DE genes for each sample should therefore have similar count values after the correction of the estimated size factor.

Using DESeq2

DESeq2 requires two input files: (i) un-normalised count data in the form of matrix and (ii) metadata with sample names as row names and the grouping variable (dose rates) in the next column. The input files are used in the function `DESeqDataSetFromMatrix` associated with tilde (~) followed by the design matrix. The design matrix indicated by coefficient vector X_j in the GLM tells whether a sample j is controlled or treated. The differential gene expression analysis is implemented in one step using the `DESeq` function, which covers size factors, dispersion estimation, negative binomial GLM fitting, and hypothesis testing. The `result` function generates the output in a table format with Base Mean, log₂ Fold Change, p-values and adjusted p-values.

1.1.4.5 Supervised learning - Differential expression analysis on continuous data with limma

Metabolomics data is continuous data and can be handled by limma, but not DESeq2. For differential metabolite analysis, the limma package creates a gene-wise linear model by estimating the gene-specific variance for all samples (Ritchie et al., 2015). Within the matrix of metabolomics data, the coefficients and standard errors of the linear regression for each row is estimated across all the sample comparisons of interest. The flexibility of the design model provided is the same as DESeq2, where the fitted objects that are created can be separated according to the groups or factors of interest and can also be compared. This creates a contrast matrix which can further be used for the calculation of log₂ fold changes and t-statistics. As the variance between genes can differ greatly, information of the estimated variances from all the genes is borrowed by empirical Bayes. A trend line is hence formed in which the gene-wise variances are pushed together to reduce variation among extremely large variances and the effects of outliers, as well as to exert strong push for consistently expressed genes with similar variance. Extremely small and large variances are adjusted to reduce the number of false positives and improve the detecting power for DEGs. The significance of expression (T-stats and p-value) of each gene between the contrast or linear model is also computed by the framework of empirical Bayes (Ritchie et al., 2015).

1.1.5 Differential testing with counts data transformations

In RNA-sequencing (RNA-seq), highly expressed genes tend to have a larger variance in expression terms across all samples than lowly expressed genes. The standard deviation per gene increased and spanned a large variance range, as the rank of the average expression grew. This phenomenon is known as heteroscedasticity as the variance is not evenly distributed across

different means. This affects the presentation of the plots which rely heavily on the genes with the highest counts in differential testing, for example, PCA plot which is a sample clustering plots for data quality control before the differential expression analysis. The actual differences between low and high-count samples made the interpretation of plots difficult.

Theory behind VST

Variance stabilizing transformation (VST) is a normalisation method from the package of DESeq2, ensuring a more equal variance along the range of dynamics when measuring either the within-group or between samples. Genes with low counts are shrunk towards the averages of genes of all samples. VST works better than regular logarithmic methods when dealing with lower counts as it does not require the additional pseudo count of 1 for the case of 0 counts. VST also prevents the inflation of noise from low counts by compressing the differences, particularly when the values are very close to zero. The transformed data makes the visualisation on sample clustering possible.

1.1.6 Functional enrichment analysis

Hypergeometric test is identical to one tailed Fisher's Exact test. It is a popular method calculating the statistical significance on variables of interest, especially to deduce the significance of enrichment. Fisher's exact test is based on the hypergeometric probability distribution in which a 2 x 2 contingency table was set up to calculate the probability of non-random association between two categorical variables. This study focuses on the over-representation test from sets of genes. In the case of radiation exposure on a set of genes, an over-represented hypergeometric probability indicates that the chance of a certain biological term, pathway or functionality represented by this set of genes happened to be more frequent than expectations.

Gene Ontology (GO) analysis

The Gene Ontology (GO) database provides systematic and hierarchical classification for the annotation of gene functions with three formalized GO terms: biological process, molecular function, and cellular component (Gene Ontology, 2015). Each ontology is organised into a directed acyclic graph with each node labelled with a corresponding GO term to facilitate large scale computational analysis. The functional coherence of the detected clusters or gene sets can be verified by conducting a hypergeometric test which further reveals their speciality associated with the given conditions by relating with enriched GO terms. If a subset of genes from an input list was consistently associated with a few GO terms, the functionality of the selected module is thus represented by those enriched GO terms.

Reactome Pathway Enrichment analysis

Reactome pathway analysis makes use of the over-representation test to determine if pathways are statistically enriched by several genes submitted from a gene list or a module. All the enriched pathways are documented in the Reactome database (<https://reactome.org>). If a subset of genes

from a selected module was consistently associated with a few pathways, the functionality of the selected module can be inferred using the knowledge from the over-represented pathway.

Integrating Transcriptomics and Metabolomics data

Paintomics3 (v0.4.5) is a webtool built in 2018 for multi-omics pathway analysis and visualization based on the KEGG pathway database (Hernández-de-Diego et al., 2018). According to the recommendation, the tab-delimited input data for each omics (gene, metabolite, region or regulatory) requires normalised data, with identifiers in the first column and the log fold changes between two conditions as the second column. The second input, feature file, is optional but usually it is a list of genes of interest, e.g., DEGs or DEMs. In default Paintomics 3 performs a Fisher's Exact test when the feature files are included and generates a combined P-value (Fisher combined probability test) if multi-omics data is provided. The Fisher combine probability test calculates the combined probability of separate tests from independent data based on the assumptions that the true effect (null hypothesis) of the combination is zero.

1.2 Aims of the thesis

Omics data, provided by NIVA, were collected from *Daphnia magna* which were exposed to low dose rates (0, 0.4, 1, 4, 10, 40, and 100 mGy/h) of gamma radiation at the neonatal stage. The gene expression profiles reflect the impact of the radiation in the juvenile stage at 4 days and in the transition from juvenile to adulthood stage at 8 days. Intuitively, the early transcriptional pattern in the juvenile stage should be capable of predicting the adverse outcome pathways involved in the late juvenile stage. Nonetheless, such investigations have not been conducted in the past, let alone the interactions between transcriptional factor bindings and the regulation of gene expression. Based on the transcriptomics data generated from the exposure to different dose rates, this study aims to identify the key genes and gene modules (from the co-expression network) associated with different exposure periods. The research problem can be solved by using linear modelling to make predictions, where the genes act as the predictor and the gamma dose rate is the response variable.

For this purpose, a prediction model provides a precise inference by fitting a statistical model to given data to retain potential genes. However, gene regulation is highly dynamic and there is no one-model-fits-all algorithm. Popular software packages like DESeq2 models the change in gene expression between conditions by conducting a univariate test on each gene individually to test for their significance. Based on the assumption that genes of similar expression levels share a similar dispersion, DESeq2 does not account for the concordance and correlation between clusters of genes. In most conditions, genes within a dynamic co-expression network are adequately interconnected to each other, while each gene presents a non-zero effect (though the effect of a single gene by itself is almost negligible). While the large amounts of data yielded by RNA-sequencing makes the search of these genes important, and in certain conditions difficult, there is an insufficient level of aquatic invertebrate molecular understanding documented.

Certainly not enough to decipher the transcriptional profile in response to the adverse outcome pathways. Therefore, without prior knowledge of the gene-expression of toxicity pathways in *D. magna*, an unbiased method which is able to prioritize the most susceptible genes and reveal the hidden structures is necessary.

Unsupervised learning, which reveals gene expression patterns without using prior labelled data, is therefore used as it is unbiased. Under the efficient network detection method, modules detected from the co-expression network infers a group of highly co-expressed genes which correspond to the gamma radiation in a static state. Genes within such clusters are functionally connected and correspond to both the exposure period and the dose rate. Yet, the identified gene modules from unsupervised learning (WGCNA) are not guaranteed to be biologically meaningful. To select the meaningful feature modules for transcriptional response, supervised learning is therefore implemented by using generalized linear regression from DESeq2. The modules, firstly, are analysed using a linear model to see how well the gene expression fits. Secondly a linear combination model (Contrast) is used to compare whether there is a difference in expression between groups of genes of different dose rates, based on the estimated log₂ fold changes.

The cells respond to external stimuli by assigning transcription factors (TFs) to specific motifs of stress-responsive genes. The causal linkage between TFs and observed transcriptional changes show the effects of radiation on the unobserved numbers of activated TFs. Genes within a module can co-regulate each other or even affect other genes in other modules. Such clusters consisting of up-regulated TF-coding genes suggest a potential dose-specific functional role, whereas the downregulation of transcript abundance from TF coding gene clusters may indicate a suppression or no non-functional role. The regulation on target genes only happens when TFs are activated by ligand binding or post-translational modification. Therefore, the transcription level of a TF coding gene does not imply the regulatory/binding activity of that TF. Rather, the expression profiles of such TF-coding genes represent the quantitative effect of a TF on the expression level of transcription. This could further be described as TF-driven hidden regulatory effects.

The first part of the study focuses on finding activated TFs that were primarily regulated by their expression level. Based on the expression data, motif enrichment analysis is a useful method to detect the enrichment of known binding motifs from the upstream of coding regions. The advance of high-throughput technologies has increased the number of TFs with known DNA-binding models drastically. Identification of enriched motifs from modules of interest may help capture any potential dose-rate sensitive TF encoding genes. As there are two different periods of gamma radiation exposure, this study also sought to explore the differences in TF-responsive genes that react exclusively between different conditions.

While there is only one set of metabolomics data from the 8 days radiation exposure available, the second part of this study hence focusses only on 8 days gamma radiation exposure using the integration of transcriptomics and metabolomics to explain the differences in reduction of

fecundity under low dose rate (1mGy/h) and (100mGy/h) as previously reported by Song et al. (2020).

To identify differentially expressed metabolites (DEMs) whose expression is correlated to the incrementing of dose rate, supervised learning can be used by following the design model of DESeq2 for DEGs discovery. DEMs can be found by fitting a linear model with genes as predictors, and the dose rates as response variables. To relate the regulatory mechanisms of the pathways to the levels of the metabolites, while there were many modules identified as significantly relevant to the impact of gamma radiation in the previous chapter, selecting relevant genes from these modules makes full use of the advantages provided by semi-supervised learning. Significant modules corresponding to the design group where the DEGs of DESeq2 belong, were combined accordingly. Software performing the integrated pathway analysis could be implemented to investigate if there is a significant overlap between the enriched transcriptomics and the enriched metabolomics pathways. There are software tools which have been developed for this purpose but only to work with non-model organisms, a systematic workflow is required to address the technical challenges.

Compared to fruit fly, nematodes and mouse, there is very little annotation data available for *Daphnia magna* in the publicly available databases, such as UCSC, Gene Ontology (GO) and KEGG. Mapping *D. magna* to a well-studied species allows us to study the gene functions for development and the response to abiotic stress. Nonetheless, the choice of model organism used for mapping and databases used for investigation will lead to varied results. The advancement of systematic collections of biological data comes with several challenges such as data heterogeneity, annotation, image construction, updating, architecture, and storage systems across different databases. Thus, this study addresses this issue and aims to incorporate this into the workflow to help the users decide on the model organism for mapping, setting up pathway analysis (PA) based on the KEGG and Reactome databases, and gene ontology (GO) enrichment for the study of adverse outcome pathways. Potential candidates for species mapping will be compared in terms of average orthologous amino acid identity and amount of documentation in the databases, to retain as much biological information as possible.

Chapter 2 Materials and methods

In this chapter, a systematic workflow is described for predicting regulation patterns for ionizing radiation using transcriptomics data (Part 1), and for investigating the reduction in fecundity using transcriptomics and metabolomics data (Part 2). The schematic of overall data analysis workflow is present in Figure 2. The software description and computing environment for every step of data analysis was documented in Figure 3.

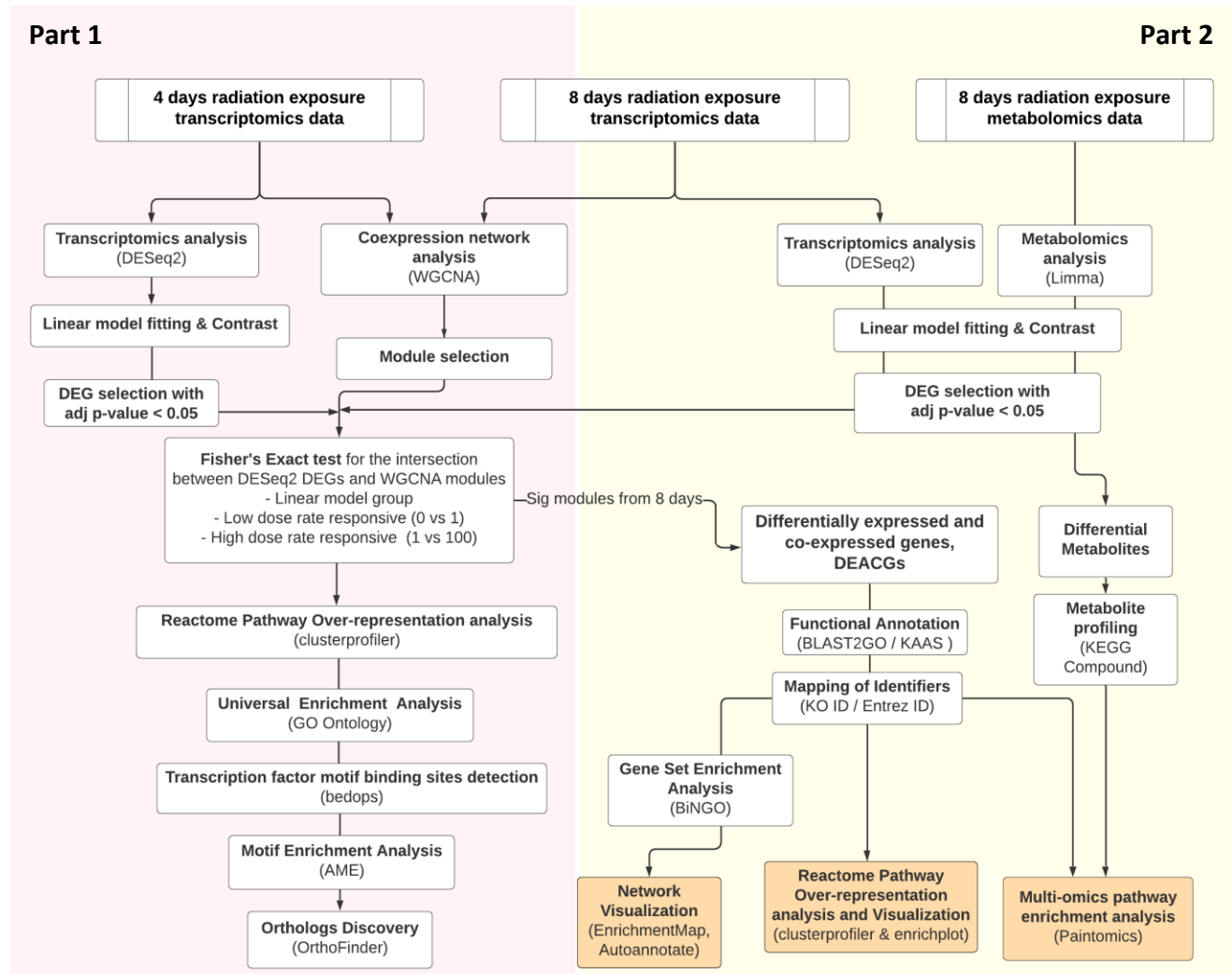


Figure 2: Overall workflow for the multi-omics data analysis. The colour partition separates the workflow into Part 1 and Part 2.

2.1 Culture conditions and exposure studies with *Daphnia magna*

Three sets of data with two types of omics data in this project were obtained from the Norwegian Institute of Water Research (Oslo, Norway). Details regarding the materials performed by NIVA is documented below:

2.1.1 Maintenance of *Daphnia* culture

The culture of *Daphnia magna* DHI strain (DHI Water and Environment, Hørsholm, Denmark), was placed in a climate room (20 ± 1 °C and 16 h light: 8 h dark), maintained using M7 media (pH 7.8 ± 0.2) which was renewed twice every week and fed daily with green algae, *Raphidocelis subcapitata*.

2.1.2 Laboratory settings for gamma radiation

The exposure period of 8 days reflects the transitional stage of daphnids from juvenile to adulthood (visible and unreleased embryo) while the 4 days exposure reflects the temporal change in the juvenile stage. As the juvenile stage was prone to the low dose ionizing radiation, the gene expression from the two timepoints can therefore be linked to the transcriptional regulation on the adverse outcome pathways to understand the effects of early toxicological events.

Individual neonates (less than 24h old) were removed from the main M7 media and placed in a plastic beaker (approx. 5-10 daphnids due to limited gamma beam width) containing 45ml of M7 medium and exposed to external cobalt-60 (8 Ci) for 4 days and 8 days. Daphnids were fed daily with concentrated *R. subcapitata*, and the media was renewed every two days. *Daphnia magna* were exposed to 7 dose rates of gamma radiation: 0 (control), 0.4, 1, 4, 10, 40, 100 mGy/h which correspond to group A, through to G at the FIGARO irradiation facility NMBU, As, Norway. The selected dose rates were sublethal (Gomes et al., 2018) and corresponding to the low exposure levels around the nuclear accident and contamination area at Chernobyl (Cardis and Hatch, 2011). While the pH of medium before and after the exposure of gamma radiation was carefully monitored with a WTW multiparameter portable meter MultiLine® Multi 3420 paired with a WTW SenTix® pH electrode with temperature sensor (Xylem Analytics, Weilheim, Germany), the samples were positioned at distances based on the measured dose-rates to water (DW) mentioned in previous study (Gomes et al., 2018, Song et al., 2020) using a nanoDot™ dosimeter (Landauer, Glenwood, USA).

2.1.3 RNA extraction

5 daphnids were pooled and stored in RNALater (Qiagen, Hilden, Germany) for each replicate, with 4 replicates per dose. The RNA was extracted using the RNeasy Mini Kit (Qiagen, Germany) following the manufacturer's protocol. Purity and integrity of the RNA was examined with a spectrophotometer Nanodrop® ND-1000 (Nanodrop Technologies, Wilmington, Delaware, USA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, California, USA). Intact RNA with clear peaks, high purity (A260/A280 > 1.8) and flat base lines as well as sufficient quantity (approximate 500ng/uL) were kept at -80°C.

2.1.4 Next Generation Sequencing and Library Preparation

Next-generation sequencing was performed on the BGISEQ-500 platform at the Beijing Genome Institute. The poly-A containing mRNA molecules were purified using poly-T oligo-attached magnetic beads, followed by the fragmentation of the mRNA using divalent cations in elevated temperatures. The cleaved RNA fragments were reverse transcribed into the first strands cDNA with random primers, and the second cDNA strands were formed with DNA polymerase I and RNase H. This process created a replacement strand containing dUTP in place of dTTP which quenched the amplification of second strands when producing the double-strand cDNA. As the synthesized cDNA fragments acquired the additional single 'A' base and subsequently ligated to the adapter, they were purified and enriched with PCR amplification. The PCR products were quantified by Qubit (Thermo Fisher, Waltham, USA) and pooled together to construct a single strand DNA circle (ssDNA circle) in the final cDNA library. Throughout the sequencing process, DNA nanoballs (DNBs) were constructed from the ssDNA circle with rolling circle replication (RCR) to boost the luminescent signals. The patterned nanoarrays were then packed with DNBs and the pair-end reads of 100bp were further read through by the combinatorial probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencer (Zhu et al., 2018).

The raw transcriptomics data generated was filtered for low quality reads (more than 20% of the bases quality < 10), adaptors, unknown bases (N bases > 5%), and mapped to a reference genome of *D. magna*.

2.1.5 Metabolite extraction

Frozen *D. magna* (10 pooled *D. magna* per replicate, 10 biological replicates for each dose) were added into a micro-centrifuged tube containing 1.35 mL of solvent consisting of methanol/chloroform/water of ratio 5:2:2 (v/w=9, µL/mg) for protein precipitation. The process of homogenisation was carried out for 1 minute in a TissueLyser JX-24 (Shanghai Jingxin Industrial Development Co., Ltd, China). The homogenates were stored at -20 °C for 24 hours and then centrifuged at 16,000g and 4 °C for 15 minutes. A volume of 1080 uL supernatant was drew into a new tube, followed by 1080 uL of methanol and topped up again with 540 uL of supernatant. 180 uL of the mixture was transferred into a GC vial filled with 10uL of amino acids mixture labelled with isotopes (0.1 mg/mL of L-Alanine-¹³C₃-¹⁵N-L-alanine, ¹³C₅-¹⁵N-L-valine, ¹³C₆-¹⁵N-L-

leucine and $^{13}\text{C}_6$ - ^{15}N -L-isoleucine). Next, the mixture dried under a gentle stream of nitrogen. 30 μL of 20mg/ml methoxamine hydrochloride in pyridine was added into the vial and vigorously vortexed for 30 secs. The vial was then kept at room temperature for 90 minutes. The mixture went through trimethylsilylation by adding in 30 μL of BSTFA (contained 1% TMCS) and was then derivatised at 70 °C for 1 hour.

2.1.6 GC-MS method and raw data preparation

The instrumental analysis was performed with an Agilent 7890A gas chromatograph linked to an Agilent 5975C inert MSD system (Agilent Technologies Inc., CA, USA). The separation of derivatives was done with an HP-5ms fused-silica capillary column (30 m \times 0.25 mm \times 0.25 μm ; Agilent J&W Scientific, Folsom, CA), with constant Helium gas (>99.99%, 1 mL/min) flow through the column. A 1 μL sample was injected in split mode (2:1) and the solvent delay period was 6 minutes. The oven temperature was set to 70 °C for 2 minutes, followed by an increase of 6 °C/min to 160 °C, continuing at 10 °C/min to 240 °C, and finally 20 °C/min to 300 °C, then held for 6 minutes sustaining 300 °C. The impact energy was 70 eV, full scan mode (m/z 50-600) was used for data collection, and the temperatures of the sampler injector, transfer line, and electron impact ion were respectively adjusted to 250 °C, 290 °C, and 230 °C. Pre-processing steps of raw GC-MS data for peak detection, picking, alignment, deconvolution etc. were based on a previously published protocol (Gao et al., 2010). The final output was exported as a peak table file and the data was normalized against the total peak intensity with an in-house bioinformatic pipeline from NIVA.

The resulting list of metabolites was matched with KEGG compound accession numbers for downstream analysis.

2.2 Data analysis and pipelines

This study, which created a data analysis workflow, starts below in section 2.2:







Steps of data analysis	Overview of software		
Data pre-processing 	DESeq2 • downloaded from Bioconductor • vst transformation	PCA • from package 'stats' included in R • <i>prcomp</i> for computing the PC (Principal Component)	
Identification of modules and DEGs 	DESeq2 • <i>DESeqDataSetFromMatrix</i> • <i>deseq</i> for DE analysis	WGCNA • downloaded from Bioconductor • <i>DeseqDataSetFromMatrix</i> for creating object that store input and intermediate values • <i>blockwiseModules</i> to identify modules • <i>factor</i> to create design matrix.	
Identification and detection of biologically meaningful modules 	GeneOverlapp • downloaded from Bioconductor • <i>newGeneOverlap</i> for Fisher exact test		
Conversion of identifiers 	BLAST2GO • paid commercial software • available for Windows system (7+) • Refseq -> GO ID • Refseq -> Entrez ID ncbi dataset • command line tool • run in Linux system • mma sequence -> gene sequence	OrthoVenn2 • online platform for orthologous gene comparisons using DIAMOND • Using OrthoMCL for gene clusterings KAAS • online platform • based on KEGG database • Refseq -> KO ID	
Detection of biologically meaningful modules and construction of transcriptional regulatory network 	ReactomePA • downloaded from Bioconductor • <i>enrichPathway</i> for identifying enriched pathway from modules clusterProfiler • downloaded from Bioconductor • <i>enricher</i> for GO overrepresentation analysis from modules	Bedops • command line tool downloaded from github • run in Linux system • extract upstream gene sequence AME • command line tool for motif enrichment analysis • run in Linux system	OrthoFinders • command line tool • software downloaded from github • detect orthologous sequence using DIAMOND with 'ultra-sensitive' mode NCBI Datasets • web service maintained by NCBI • useful in downloading small numbers of protein sequence
Integration of metabolomics and transcriptomics data 	limma • downloaded from Bioconductor • <i>lmFit</i> for estimating the logFC • <i>empiricalBayes</i> for estimating standard error • <i>makeContrast</i> for contrasting dose rates of interest • <i>contrast.fit</i> for final coefficients and standard errors • <i>topTable</i> for displaying results	BiNGO • Cytoscape plugin • GO analysis for DEACGs AutoAnnotate • Cytoscape plugin • Assigned new annotations for similar GO clusters EnrichmentMap • Cytoscape plugin • visualization of enriched GO and corresponding cluster of genes	Paintomics • online platform • integrating the DEACGs and DEMs for enriched pathway / chemical reaction analysis ReactomePA • <i>enrichPathway</i> for enriched pathway analysis for DEACGs

Figure 3: Overview of software in each step of data analysis. R commands are shown in italic.

2.2.1 Data pre-processing

Raw counts from the transcriptomics and metabolomics data were normalized using the function Variance stabilizing transformation (vst) included in the DESeq2 package to create a matrix of values with more consistent variance along the range of mean. For quick assessment on the overall similarity between samples, the transformed matrix was input into the principal component analysis (PCA) with `prcomp` function and visualised with PCA plot which were included in the next chapter.

2.2.2 Identification of modules and DEGs

As WGCNA works on normalized RNA-seq data, after variance-stabilizing transformation (vst function from DESeq2), datasets from 4 days (12,072 transcripts) and 8 days (11,921 transcripts) were piped into the function `blockwiseModules` in a single block manner, along with “signed network” and “signed TOMType” as parameters. The workstation in this study had access to 16GB of RAM, therefore it could handle up to 20,000 genes per block, as recommended.

Differential expression analysis with DESeq2

Unlike a standard WGCNA workflow which measures the Pearson correlation between gene expression and module eigengene, the negative binomial GLM from DESeq2 was used to model the differences in gene expression between conditions.

There were three input parameters in `DESeqDataSetFromMatrix`: the count matrix of the raw transcriptomics data, the metadata which documented the name of the samples and radiation dose rate in matrix form, and the design matrix. The output was piped into `DESeq` function for differential expression analysis. Adjusted p-value < 0.05 was used as the cut-off to select the differentially expressed genes.

Design matrix of X_j

Various dose rates are treated as different groups or different levels of factors by using the `factor` function for the organisation. Dose zero was specified as the base-level in the linear model.

$$\log_2(q_{ij}) = X_j \beta_i$$

- Measure the effects of dose rates by fitting a linear model

The model matrix consisted of only two groups in this model, control, `ctrl` (0 mGy/h) and treatment, `trt` (0.4, 1, 4, 10, 40 or 100 mGy/h). The hypothesis testing investigates if there was significant difference between the control and the treatment group:

$$H_0 : \beta_{ctrl} = \beta_{trt}$$

$$H_1 : \beta_{ctrl} \neq \beta_{trt}$$

- Measure dose-specific effects by linear combinations (Contrast)

To test if the combinations of variables have non-zero effects, the model matrix was modified to consist of 7 groups corresponding to various levels of dose rates. The contrast function was used to contrast coefficient of interest for the hypothesis testing:

$$H_0 : \beta_{trt0} = \beta_{trt1}$$

$$H_0 : \beta_{trt1} = \beta_{trt100}$$

$$H_1 : \beta_{trt0} \neq \beta_{trt1}$$

$$H_1 : \beta_{trt1} \neq \beta_{trt100}$$

where the different expression of gene i under different dose rates (0 vs 1 and 1 vs 100 mGy/h) were indicated.

2.2.3 Identification and detection of biologically meaningful modules

Module Membership cut off

The module membership (MM) measures the association between the gene expression and the module eigengene. It also depicted the intramodular connectivity because an intramodular hub gene that has a high absolute value of MM approaches 1. A cut off of $|MM| > 0.05$ was applied to all modules.

Fisher's Exact test

Every module created by WGCNA contains a unique topological characteristic. To identify modules that overlap with differential gene expressions with non-random association, DEGs from DESeq2 were loaded into the R package, GeneOverlap. The function `newGeneOverlap` was applied for Fisher's exact test (Shen, 2014). Modules that passed the statistical significance cut-off of p-value < 0.05 were considered to have a significant association.

2.2.4 Conversion of identifiers

BLAST2GO for Entrez ID mapping and GO annotation

Blast2GO (B2G) (v4.1) was used to conduct 'BLASTing', mapping and sequence annotation between *D. magna* and *D. melanogaster* with default settings (Conesa and Götzt, 2008). Out of 23570 transcripts with Refseq accession, 16221 were functionally annotated and mapped to Entrez ID. Entrez IDs which contained null values in the expression data such as base mean, p-value and adjusted p-value were removed. As multiple transcripts were assigned with the same Entrez ID, the remaining went through a duplication check with transcripts that bear the lower

adjusted p-value being taken into the expression data. After processing, 19512 sequences were successfully annotated with GO terms.

Ortholog detection

The detection of orthologous genes at the level of amino acids was carried out using OrthoVenn2 (<https://orthovenn2.bioinfotoolkits.net/>) with a default setting for the E-value of 1e-12, and an inflation value of 1.5 (Xu et al., 2019). The protein sequences of *D. magna* KIT (our study strain) were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_003990815.1/) whereas the sequences for *D. magna* xinb3, *D. pulex*, and *Drosophila melanogaster* were sourced from the Ensembl database connected to the website of Orthovenn2 (Pruitt et al., 2005). Firstly, the website also provided DIAMOND (v0.9.24), the embedded analysis tools used to perform all-vs-all sequence comparison and protein annotation. Afterwards, OrthoMCL (Li et al., 2003) was launched to allow the clustering of orthologous genes based on their conserved sequences.

Ortholog mapping: *Daphnia pulex* and *Drosophila melanogaster*

Out of the 23570 transcripts, 11921 expressed transcripts from *Daphnia magna* were converted to a gene sequence using the command line tool, ncbi datasets (Coordinators, 2015). 11868 gene sequences were uploaded to the KEGG Automatic Annotation Server (KAAS) (<https://www.genome.jp/kegg/kaas/>) to look for matching orthologs of *Daphnia pulex* and *Drosophila melanogaster* in the KEGG database with the single directional hits (SDH) setting. 4848 and 5420 KEGG Ortholog (KO) IDs were assigned to the corresponding sequences of *D. melanogaster* and *D. pulex*. The output of *D. pulex* produced more KO IDs than *D. melanogaster*, meaning *D. pulex* was the best choice at this stage. The KO IDs were mapped to the expressed transcripts through common gene sequence identifiers, followed by removing the transcripts without p-values or adjusted p-values.

2.2.5 Detection of biologically meaningful modules and construction of transcriptional regulatory network

Pathway Enrichment Analysis for significant modules

Due to the unavailability of high-quality pathway annotations for *Daphnia magna*, DNA-to-protein BLASTX was implemented in the program BLAST2GO conducting a translated search to map the annotations of *D. magna* to the gene and protein sequences of fruit fly *Drosophila Melanogaster* (Conesa and Götz, 2008). To identify underlying biological pathways associated with the effect of gamma radiation, genes from each module were assigned with an Entrez ID if they were homologs to *D. melanogaster*. The output was loaded into Reactome pathway analysis using ReactomePA (v1.38) and tested towards the curated pathways of *D. melanogaster*. A hypergeometric model was applied to determine if certain pathways were enriched with the function enrichPathway (Yu and He, 2016). Pathways found within a module containing genes with

very dissimilar biological functions, were ranked. The statistical cut-off for enriched pathway was p-value < 0.05.

GO analysis to characterize functionality of modules

Functional annotation with GO requires a gene set where each gene was given a predefined GO term. In this study, homologs between *D. magna* and *D. melanogaster* that were assigned with GO terms were input into the over-representation analysis. The R package clusterProfiler provides the `enricher` function for hypergeometric tests, chosen to conduct a statistical test with user customized annotation (Yu et al., 2012). GO terms which passed the cut-off of p value < 0.05 were considered significant.

Transcription factor (TF) motif enrichment analysis and ortholog findings

The upstream region of 2000bp from all coding sequences were extracted with the unix package, `bedops(v2.4.40)` (Neph et al., 2012a). The selected sequences were piped to the AME (Analysis of Motif Enrichment) algorithm, which use all different types of fly (*D. melanogaster*) motif databases Flybase (<https://flybase.org>) available for motif predictions. Genes that did not belong to the selected modules are used as background in the input. Over-represented motifs were searched for corresponding Flybase IDs. The protein sequences of Flybase ID were downloaded from NCBI and ortholog sequences between *D. magna* and *D. melanogaster* were found using DIAMOND(v0.9.24), with the setting 'ultra-sensitivity' from software package OrthoFinder v2.3.3 (Emms and Kelly, 2019).

2.2.6 Integration of metabolomics and transcriptomics data

Detecting differential metabolites with limma

PCA plots were first used to examine the quality of metabolome data before the differential expression analysis. The R package limma, was used to fit a linear model to the continuous data of metabolomic studies. The same design of model matrix for the linear model and linear combination from DEseq2 was used to call for differential metabolites in limma (Ritchie et al., 2015). The `lmFit` function was used to estimate the log fold changes and the `empiricalBayes` function was used to estimate and smooth the standard errors. Empirical Bayes is a modified t-statistic which makes effect estimation in differential expression analysis. To perform pairwise comparisons on dose rates of interest, the function `makeContrast` was employed to select groups of interest followed by the function `contrast.fit` used to generate final coefficients and standard errors. The result table was displayed by using `topTable` and the same criteria of adjusted p-value < 0.05 was used to select DEGs.

Differentially expressed and co-expressed genes (DEACGs)

Genes from significant modules based on the Fisher Exact test were combined into three groups: linear model, low dose responsive (0 vs 1 mGy/h), and high dose responsive (1 vs 100 mGy/h). The combination of unsupervised WGCNA and supervised DESeq2 included all relevant dose-rate

responsive genes to a greater extent and thus generated differentially expressed and co-expressed genes (DEACGs).

Paintomics for integrated transcriptomics and metabolomics analysis

The model organism chosen was *D. melanogaster* (because the KEGG database does not include *D. pulex*) and therefore the input identifiers were Entrez ID for gene expression data. For the input of metabolomics data, KEGG compound names were used as the identifier.

Reactome PA for DEACGs

DEACGs matched to the Entrez ID were filtered for duplicate identifiers by choosing the non-unique transcripts with the lowest adjusted p-value. Afterwards, the R package, ReactomePA, was used with the `enrichPathway` function to conduct pathway enrichment analysis on the transcriptomics data. By default, the pathway enrichment cut-off of adjusted p-value of < 0.05 was considered statistically significant in this study. The pathway network was built with gene names and top enriched pathways to demonstrate overlapping genes across pathways.

GO Enrichment Analysis for DEACGs

DEACGs from the three groups (low- and high dose rate responsive groups, and the linear model group) were individually loaded into a Cytoscape plugin (BiNGO v3.0.3; Cytoscape v3.9.0) to perform GO enrichment analysis (Maere et al., 2005, Shannon et al., 2003). GO terms with the Benjamin-Hochberg Corrected p-value (FDR) below 0.05, considered to be significantly enriched, were further piped into EnrichmentMap for visualisation (Merico et al., 2010, Isserlin et al., 2014). A GMT format file which served as the background set for all gene sets and corresponding GO terms was also added into the analysis. Mutually overlapping gene sets which formed clusters were input into AutoAnnotate to assign annotations to each cluster (Kucera et al., 2016).

Chapter 3 Results

This chapter can be divided into two sections: 3.1 and 3.2. The first part demonstrates the differences in the transcriptional regulatory networks from the 4 days and 8 days of gamma radiation exposure. The second part describes the integration of metabolomics and transcriptomics data, with the significant modules combined based on the design groups, then formed into DEACGs (differentially expressed and co-expressed genes). The beginning of each section introduces the methods used to assess the expression profiles of the genes and metabolites. The subsections describe the output from each step in the workflow. In short, section 3.1 includes the modules generated from WGCNA selected using the Fisher Exact test through overlapping the genes in every module and DEGs from DESeq2. The selected modules are termed as significant and processed with functionality characterisation. The transcriptional regulatory network was created with significant modules connecting to each other as nodes, while the number of edges was decided by the transcription factors and the motif binding sites available within and between the modules. Lastly, the similarity between the modules was compared and evaluated, followed by the AOP integration. In section 3.2, the significant modules were combined according to the design group and aligned with differential metabolites which were detected with limma for integrative pathway enrichment analysis. GO and Reactome Pathway analyses were conducted with DEACGs to provide insights beyond module specificity.

3.1 Identification of dose rate responsive co-expression modules and transcriptional regulatory networks

3.1.1 Initial assessment on gene expression profile with PCA and scatterplots

This workflow presents a simple diagnostic method using PCA plot, which is also a dimension reduction method in the beginning of this workflow to compare the sample similarity from 4 days and 8 days. PCA plot is a quick visualization tool which provides unsupervised information on directions that explain the most variabilities. After VST, the samples were visualised using a PCA plot, where the x-axis is the direction of the first principal component (PC1) which explained the most variance, and the y-axis is the second principal component which is perpendicular to PC1 and explained the variance of the data the second most.

The first assessment using a PCA plot showed that the clustering of samples is entwined with the exposure time (Figure 4A). Both transcriptomics data therefore cannot be combined into a single data analysis due to the clustering patterns which revealed the batch effects convolved with the

exposure time. Removal of the batch effects would have also eliminated the effect of time. Hence, the 4- and 8-days transcriptomics data were analysed individually. Clustering gene expression data from different samples did not detect any batch effects as shown in the PCA plots (Figure 4B and 4C). No distinct clustering pattern was observed corresponding to the 7 exposure dose rates indicating that the changes for each sample could be very small.

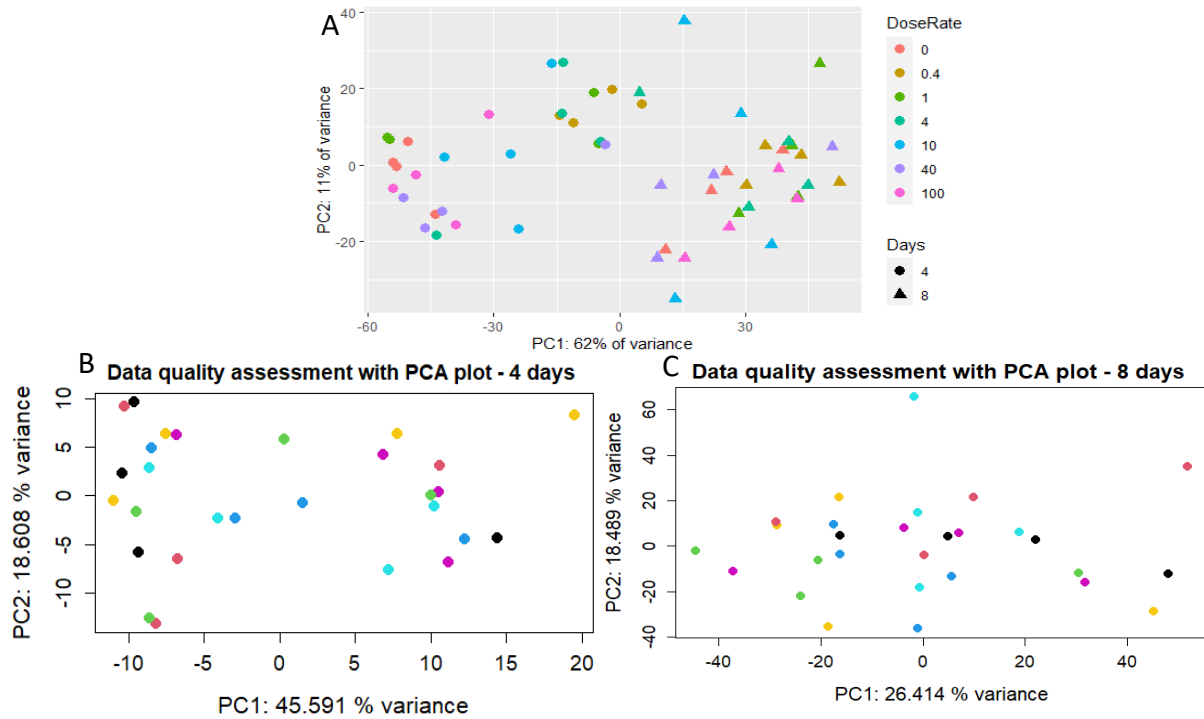


Figure 4: PCA plot showing the samples in a 2D plane spanned by the first two principal components which explained the most variance. **A:** The gathering of circle point on the left and triangle point on the right indicates that the batch effects were convolved with the exposure periods (4 days and 8 days). **B & C:** Sample similarity measure, no obvious clustering pattern observed.

The second diagnostic method used a scatterplot, where the x-axis represents the log fold changes of DEGs detected from 4 days, and the y-axis represents the same from 8 days. Both datasets were fit to a linear model using DESeq2 and the scatterplot below (Figure 4) compared the log fold changes of DEGs between 4 days and 8 days, and whether the numbers of genes were presented only in 4 days, 8 days or presented in both exposure periods. The second initial assessment (Figure 4) showed that the same genes that were downregulated from 4 days gamma radiation exposure were upregulated after 8 days of exposure. In the opposite directionality, the same genes that were up-regulated when exposed to a shorter radiation period became down-regulated when the exposure period increased. Only a small number of genes (red dots) maintained the same expression patterns in both datasets. The observed DEGs varied according to exposure period and shows that the transcriptomic responses of 8 days data do not follow the same expectation as the data from 4 days. As the regulatory magnitude and directionality of each gene varies according to the dose rate and exposure period, this study did not categorise each

differential gene expression into up and down regulation. Instead, the WGCNA approach was applied prior to the differential expression analysis to capture any monotonic and non-monotonic patterns of expression.

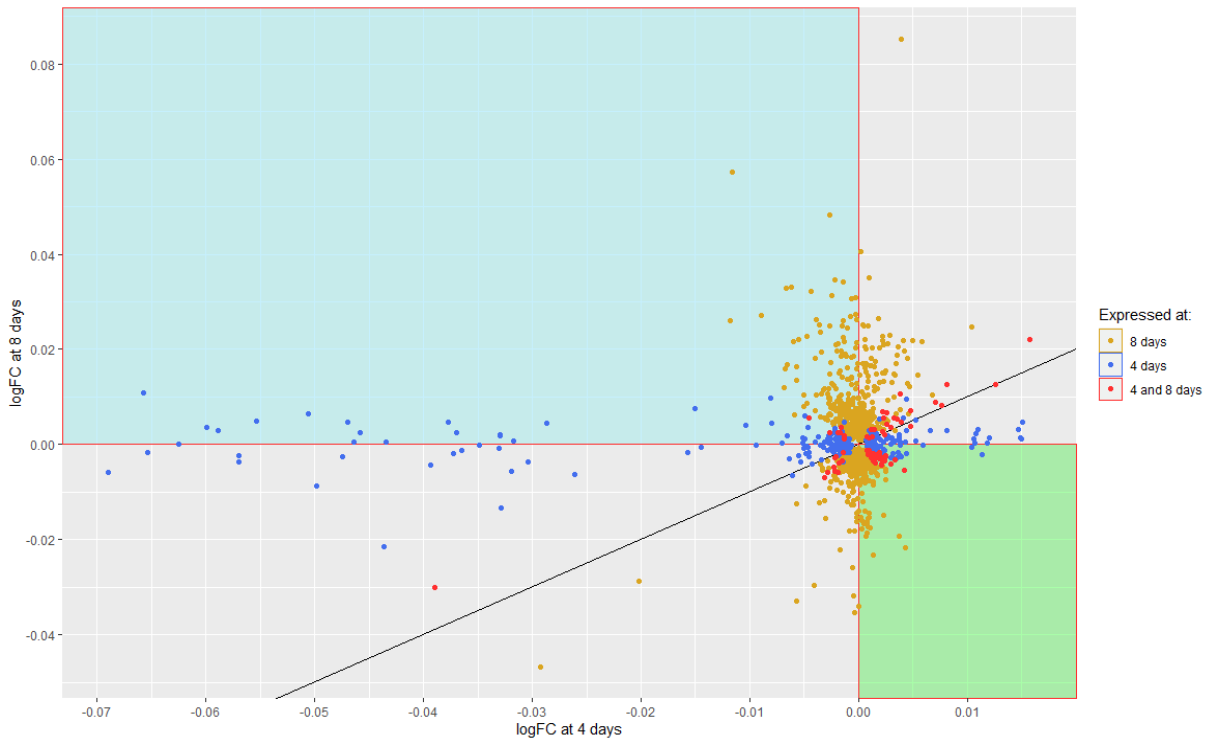
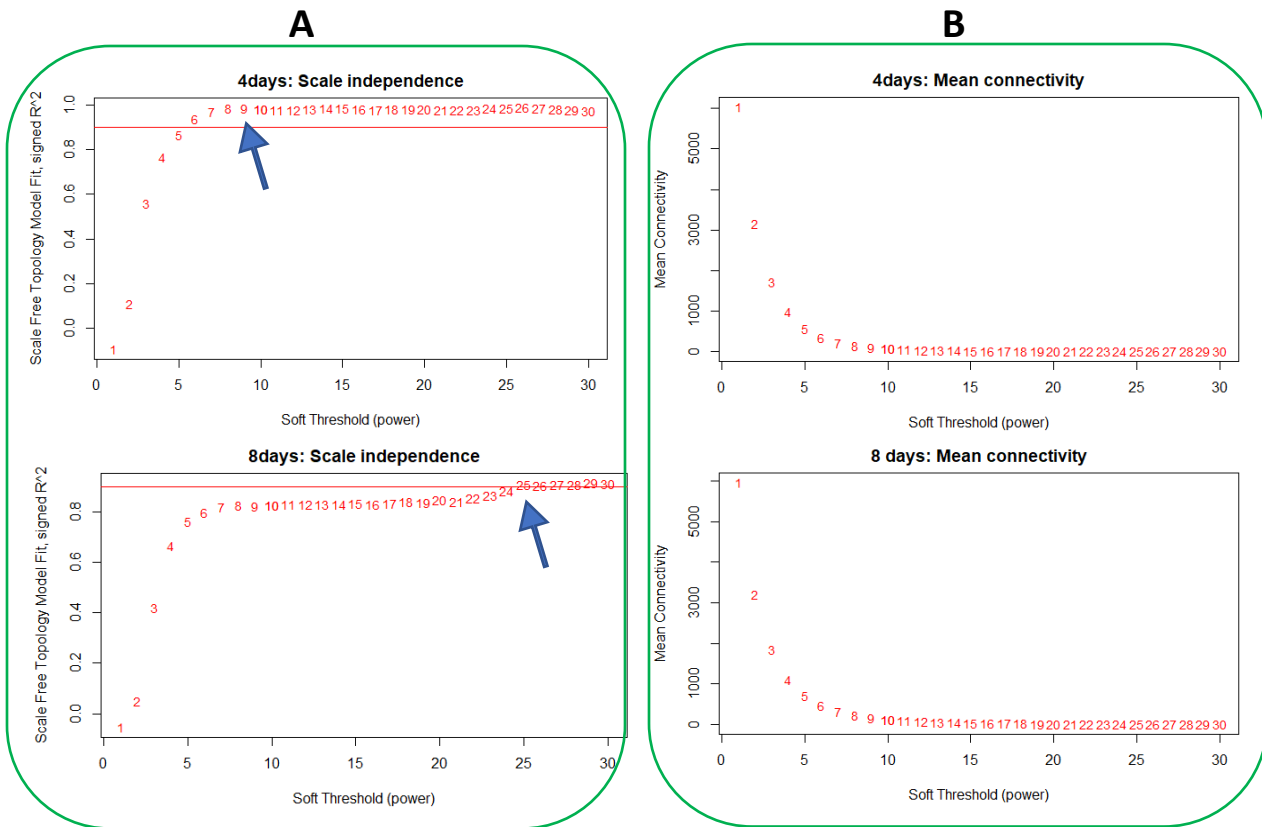


Figure 4: Scatterplot showing the log2 fold changes of the same genes exposed to 4 and 8 days of gamma radiation. Exclusive DEGs for 8 days are marked in yellow, exclusive DEGs for 4 days are marked in blue and DEGs present in both are marked in red.

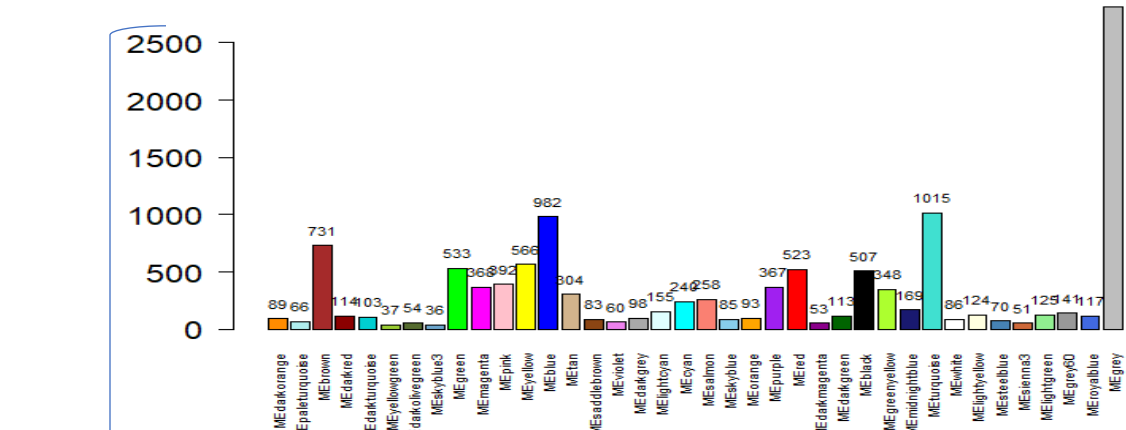
3.1.2 Creation of a co-expression network

A biological system is considered to behave like a molecular interaction network with scale-free characteristics because small numbers of genes, or nodes, have exceptionally high numbers of interaction compared to the majority. In random biological mutations, the removal of a small number of nodes/vertices does not alter the underlying fundamental structure easily because the chance of removing a highly connected node is very low, and the removal of peripheral nodes which have fewer connections usually does not affect the integrity of the entire network. To create a scale-free network which assimilates this property, the transcriptomics data from 4 days of gamma radiation exposure requires a soft threshold (β) of 9 to reach scale-free topology as R^2 was maximized at 0.98 and yielded a high mean connectivity of 94 (Figure 5). Meanwhile, $\beta = 25$ is chosen for the transcriptomics data of 8 days of gamma radiation exposure. R^2 reached 0.909 and yielded a mean connectivity of 9.87.

The proposed WGCNA approach yielded 38 modules among 12907 nodes and 36 modules among 11921 nodes from the 4 days and 8 days gamma radiation exposure transcriptome data. All modules contained at least 30 genes.



4days: Number of genes per module



8days: Number of genes per module

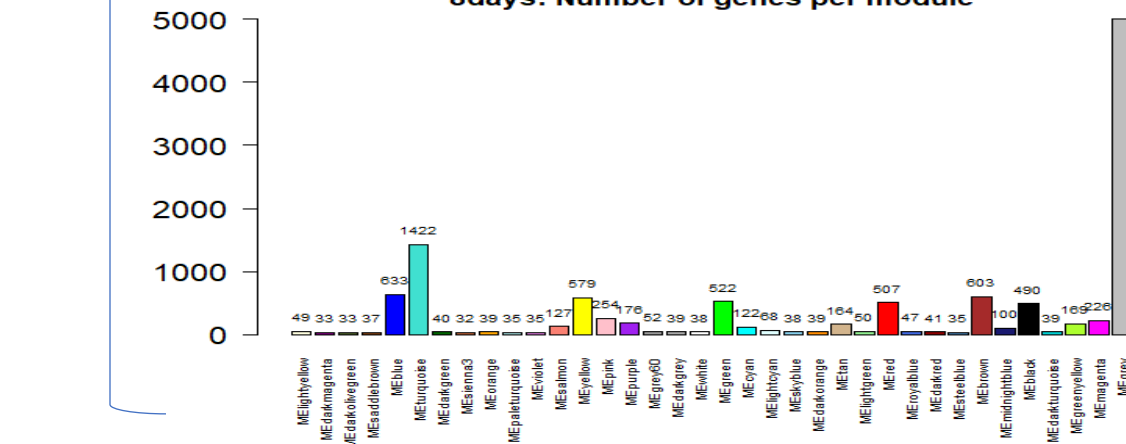


Figure 5: WGCNA. Diagnostic plots showing various beta fits to reach a scale-free topology network. Analysis of scale-free network topology using different soft-thresholding power on 4 days transcriptomics data. On the upper left, A shows the scale free fit index (y-axis) and the upper right (B) shows the mean connectivity (degree, y-axis). C shows the numbers of genes in every module, which was given an arbitrary colour.

3.1.3 Differential expression analysis

The numbers of genes involved in differential expression analysis were the same as in WGCNA, 12907 and 11921 for 4 days and 8 days data. The data from 4 days has fewer DEGs than the 8 day's data when fit to the linear model and assigned into linear combinations/contrasts (1 vs 100 mGy/h). Yet, 233 low dose rates responsive DEGs (Ctrl vs 1 mGy/h) from the 4 days dataset show a significant difference in expression which is more than the 8 days dataset which has only 49 DEGs. Table 1 summarizes all DEGs resulting from being fit to a linear model and contrast.

Table 1: The number of differentially expressed genes generated by DESeq2. Except for Ctrl vs 1 which is a low dose rate-responsive group, the data from 8 days has more DEGs than from 4 days in the high dose-rate responsive group (1 vs 100) and the linear model.

DESeq2		Number of DEGs	
		4 days	8 days
Linear model		312	1262
Linear combination /contrasts (mGy/h)	Ctrl vs 1	233	49
	1 vs 100	623	1153

3.1.4 Modules with non-random association with DEGS

To ensure the unbiasedness and retain all highly relevant genes, grey modules (grey bar in Figure 6: C, D, E, & F) which are formed by genes that could not be assigned to any module due to dissimilar co-expression, were included in the module membership filtration ($|MM| < 0.05$) and significance test. The intersection of modules and DEGs detected from DESeq2 selected 11 and 12 modules that passed the Fisher's Exact test for significance from 4 days- and 8 days- dataset (Table 2). Modules were termed as 'significant modules' if they passed the cut-off from Fisher Exact test of p value < 0.05 . Some significant modules from the 4 days data were found in both groups of linear combinations (marked with a £ sign), but such an observation was not found in the linear combinations of the 8 days data. This indicates that a shorter exposure period to gamma radiation tends to involve more genes from the same clusters and these genes are more likely to react in an opposite regulatory direction due to non-monotonic responses as the dose rate increases. Longer radiation exposure on the other hand activated different clusters of regulatory genes as the dose rate increased, suggesting more complicated functionality is involved.

Table 2: Summary of qualified (significant) modules having an overlap between WGCNA and DESeq2. Double asterisks (**) indicate modules that are exclusively found at the intersection of the specific design model ('linear model', 'Ctrl vs 1' or '1 vs 100') and WGCNA modules. Pounds sign (£) refers to those that are at the intersection of linear combinations (exists in both

'Ctrl vs 1' and '1 vs 100') and WGCNA. Modules with pound sign only found in 4 days data but not in the 8 days data.

Design model WGCNA \cap DESeq2		Modules with non-random association (p-value < 0.05)					
		4 days			8 days		
Linear model		Black** Pink	Blue Red	Brown Yellow**	Blue** Red Lightgreen**	Darkgreen Greenyellow Lightcyan	Green Lightcyan
Linear Combination (Contrast)	Low dose rate responsive (Ctrl vs 1)	Blue Red Darkturquoise**	Green [£] Lightcyan [£] Turquoise [£]	Pink Lightcyan [£] Turquoise	Black** Magenta**	Brown** Midnightblue**	Greenyellow
	High dose rate responsive (1 vs 100)	Blue Pink Darkturquoise [£]	Brown Lightcyan [£] Green [£]	Cyan** Red Green [£]	Darkgreen Grey60** Red	Green Lightcyan	Greenyellow

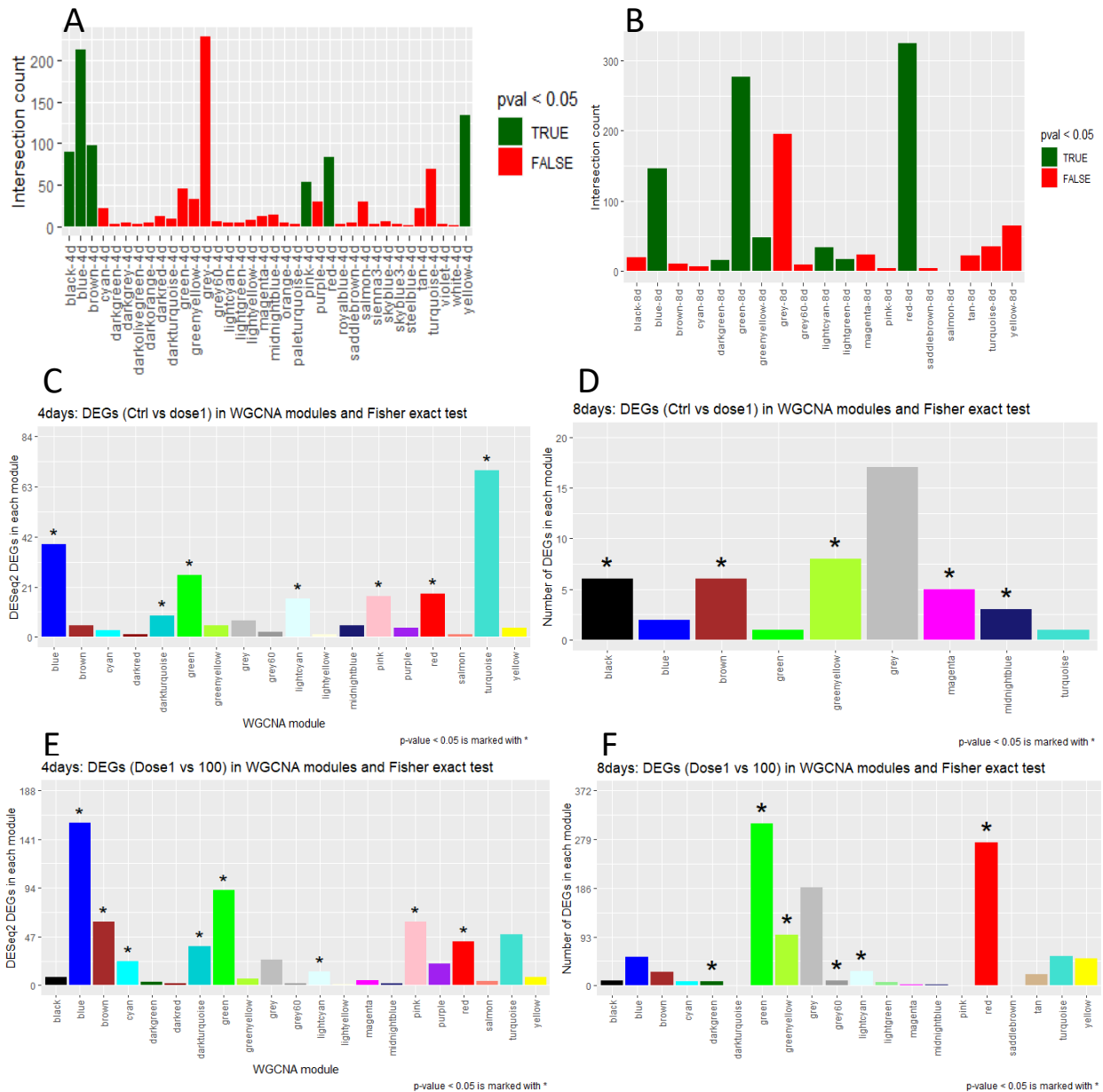
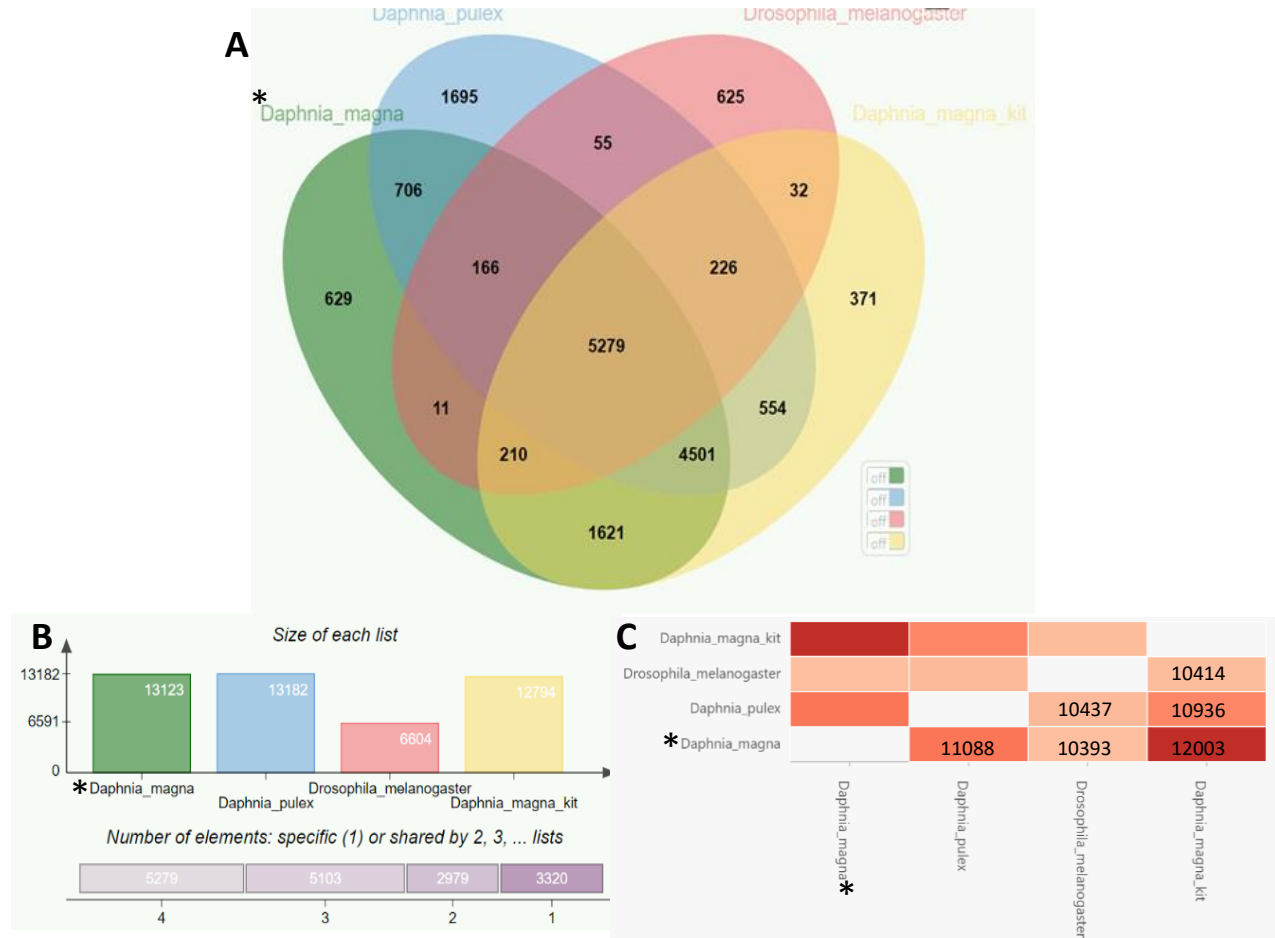


Figure 6: The bar plots show the association between WGCNA and DESeq2 after Fisher's Exact Test. **A & B**: Association of gene module with the linear model. Modules in red passed the p-value cut-off < 0.05. **C, D, E & F**: Association of gene module with the linear combination (contrast) model. Modules marked with asterisk (*) passed the significant p-value cut-off < 0.05. The 4 days data yielded 11 significant modules while the 8 days data yielded 12.

3.1.5 Whole proteome comparison and ortholog identification shared by *Drosophila* and *Daphnia*

The closest model organism to *Daphnia magna*, which has been extensively used in the study of evolution, toxicology, developmental biology, and genetics over hundreds of years is *Drosophila melanogaster* (Devineni and Heberlein, 2013, Campos et al., 2018). As the crustaceans have been proposed to be the sister group of monophyletic hexapods (insects), together they form the

Pancrustacea, *Daphnia* and *Drosophila* are thus likely to have shared related genes and developmental patterns throughout their evolution (Regier et al., 2005, Zrzavý and Štys, 1997). The functional annotation in this study relies heavily on mapping between different species; it is therefore important to get a quick glimpse on the homologous relationship between sequences before the mapping of gene identifiers for the functional enrichment analysis. In computational biology, inferring the orthology is an essential part to elucidate the evolutionary process, i.e., speciation or gene duplication.



**Daphnia magna* labelled with asterisk (*) sign refers to *Daphnia magna xinb3*, a different strain from our study species *D. magna KIT*.

Figure 7: Comparison of orthologues genes between different clones of *Daphnia magna*, *Daphnia pulex* and *Drosophila melanogaster*. **A**: Venn diagram showing the numbers of shared orthologous groups between *D. pulex*, *D. magna xinb3*, *D. magna KIT* and *D. melanogaster*. **B**: The bar graph above shows the numbers of protein clusters found in each species, while the bar plot below displays the number of orthologous clusters shared by 1, 2, 3 and 4 species. **C**: Pairwise heatmap with number of overlapping clusters between different pairs of species. The overlapping cluster numbers were indicated in the cells and the colour intensity followed the shared number of orthologous groups: the darker the colour, the more orthologs shared between species.

In Figure 7, 16681 clusters were shown in the Venn diagram, with 5279 orthologs shared in all 4 species, 13361 ortholog clusters shared by at least two species, and 3203 clusters containing orthologous genes which have only one copy in each species (single copy gene clusters). A total

of 3320 clusters uniquely belong to only one species. The clone *D. magna* KIT shared the most orthologous genes with *D. magna* xinb3, followed by *D. pulex*, and then *D. melanogaster*.

3.1.6 Mapping of transcripts to gene identifiers

The decision on ID conversion is not only determined by the model organism but also by the type of identifiers adopted by various database platforms. Entrez ID is probably the most popular identifier that supports multiple databases, but it is not available for *D. magna*. On the other hand, the KEGG database contains species from the same genus, *D. pulex*, which makes it a potential target of conversion for *D. magna*.

Overall, the gene sequence of *D. magna* KIT into *D. melanogaster* generated around 6470 unique matches and 4645 unique matches for Entrez ID and KO ID, whereas mapping with *D. pulex* generated 5189 matches for KO ID. The mapping with Entrez ID of *D. melanogaster* was able to retain more expression data and so it was chosen for functional enrichment analysis.

The clones *D. magna* xinb3 and *D. pulex* were undoubtedly the first and second choice target to map to due to being the same species, but the limited data available in the KEGG and Entrez database would naturally limit the applicability in functional annotation analysis.

Table 3: Mapping of identifiers from expressed transcripts of *D. magna* to Entrez IDs of *D. melanogaster* and to KO ID of *D. pulex*. Entrez ID was chosen for the mapping of ID from *D. magna* to *D. melanogaster* because it retained more transcripts than KO IDs from *D. pulex*.

ID conversion	Expressed transcripts	DEACGs	Conversion into Entrez ID (<i>D. melanogaster</i>)		KEGG Ortholog ID (<i>D. pulex</i>)	
			Mapped, non-duplicated, expressed transcripts	Mapped, non-duplicated DEACGs	Mapped, non-duplicated, expressed transcripts	Mapped, non-duplicated DEACGs
Low dose rate responsive (0 vs 1 mGy/h)	11921	1535	5214	767	2077	278
High dose rate responsive (1 vs 100 mGy/h)	11921	1352	4923	726	1933	287
Linear model	11921	1979	4922	1208	1932	487

The number of transcripts associated with KEGG Ortholog (KO) IDs was almost halved. This is a very common issue in mapping because each gene is associated with multiple transcript isoforms and the transcripts representing the isoform are not kept.

The number of transcripts was reduced further by removing those that share the same KO ID because software packages for enrichment analysis and visualization do not accept non-unique

identifiers. Transcripts which bore the lowest adjusted p-value were retained along with their KO ID, log fold change, p-value and adjusted p-value for the final expression data.

3.1.7 Pathway and GO Enrichment Analysis (Gene modules)

Reactome Pathway Analysis (PA) was implemented along with GO enrichment to integrate functional information with gene module identification. The addition of GO terms provides relevant pre-defined terms based on the functionality; meaning that the specific purposes which span across several pathways, usually not easily deciphered, can be revealed. This especially applies to pathways that are involved in multiple cellular functions. The top 20 enriched pathways and GO terms of all significant modules were displayed in bar and dot plots (S1 & S3 – page 92,97).

GO analysis demonstrated that 1436 and 1686 over-represented GO terms with p-value < 0.05 were found from the significant modules from 4 and 8 days of gamma radiation exposure, respectively. To compare the enriched GO terms between the 4- and 8-days data, the Venn diagram in Figure 8 shows the number of differences in biological processes, cellular components, and molecular functions. More than 99% of over-represented GO terms in the significant modules from the 4 days data also existed in the data from 8 days.

Mapping of *D. magna* into *D. melanogaster* for Reactome PA revealed 269 and 299 enriched pathways. Pathways related to RNA turnover, mitochondrial energy generation, cell cycle, and stress response were predominantly affected after 8 days of radiation exposure. Meanwhile, the data from 4 days shows that translation relevant pathways for initiation silencing and termination, ribosomal interaction, and formation were mostly affected. The difference between enriched Reactome pathways were correlated to biological and molecular functions annotated by GO.

However, it is noticeable that some exclusively affected cellular components from the 4 days data were distinctively opposite of those from the 8 days data. In the 4 days data, the affected photoreceptor outer segments are embedded with photoreceptor proteins which convert light into signals to trigger various biological processes. Some studies have proven that exposure to radiation damages the membrane of outer segments and leads to increased ROS production derived from the NAPDH oxidase (Nox) in the outer segment. It is therefore reasonable to suggest that the enriched pathway nonsense-mediated decay (NMD) which exists exclusively in the data from 4 days of exposure is employed to get rid of the build-up of the Nox protein and their isoforms (Roehlecke et al., 2013). Whereas in the 8 days data the affected cellular component, photoreceptor inner segments, contains mitochondria which is also responsible for high ROS production. Which again aligns with the top enriched Reactome pathway for TCA cycle, ATP synthesis, and electron transportation. Other cellular components such as the male pronucleus, female pronucleus (responsible for fertilization), and condensed nuclear chromosomes (responsible for chromosomal meiotic and nuclear mitotic process) also correlated well with the impact of radiation on the age and developmental stages of *D. magna*. A complete list of over-

represented GO and Reactome pathways can be found in supplemental documents (S1 & S3 – Page 92 & 97).

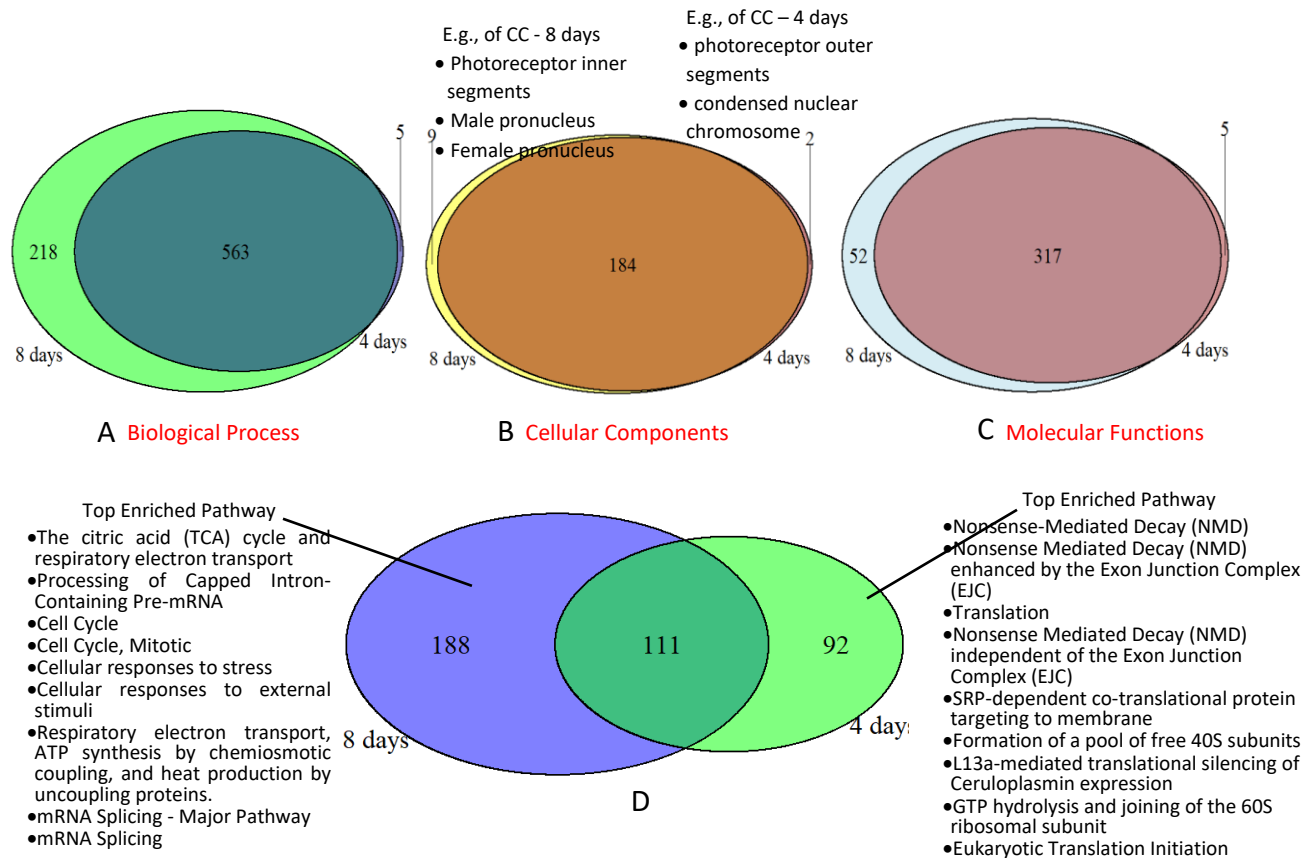


Figure 8: Venn diagram comparing GO terms and Reactome pathway between 4 days and 8 days data. The upper diagrams (8A,8B,8C) correspond to the GO domain: biological processes, cellular components, and molecular functions. Diagram 8D shows the difference of radiation affected pathways between 4 and 8 days of gamma radiation exposure arranged in a descending order according to the number of genes.

3.1.8 Transcriptional regulation by active transcription factors

The analysis of motif enrichment analysis (AME) identified 8 modules containing significant enrichment of motif binding sites in the promoter regions from both transcriptome data. The 59 TF-coding genes recognizing the motif sites of *D. melanogaster* were preceded with an ortholog search in the significant modules of *D. magna*. The putative ortholog output from OrthoFinders was narrowed down to 26 (4 days) and 30 (8days) TF candidate genes. A module usually contains multiple transcription factors but not every one of them will be activated. The TFs were considered activated and involved in transcriptional regulation if they belonged to significant modules and presented corresponding motif sites in the upstream of transcription start site (TSS). The enriched motif sites of significant modules (S1 – page 92) and TF encoding genes (Table 4

and Table 5) across different dose rates were combined; 17 and 28 active TFs from 4 days and 8 days data, respectively, were found. The transcriptional regulatory relationship between modules and corresponding TFs are described in Figure 9 and Figure 18. The expression profile of all activated TF encoding genes were documented in the Supplementary data (S2 & S3 - page 95 & 100).

Table 4: TF orthologs and their corresponding modules from the data of 4 days of gamma radiation exposure. Gene symbol and gene names followed the nomenclatures of *D. melanogaster* as documented in Flybase. TFs are considered activated if a corresponding enriched motif was found and the ortholog genes which encoded for the TFs were present in that module.

Transcript ID	Gene Symbol	Gene name	Module Name	Activated?
XM_032938013	<i>Lim3</i>	<i>Lim3</i>	black	Y
XM_032922018	<i>Sp1</i>	<i>Sp1</i>	black	N
XM_032939826	<i>NK7.1</i>	<i>NK7.1</i>	blue	Y
XM_032933675	<i>lola</i>	<i>longitudinals lacking</i>	blue	N
XM_032940896	<i>br</i>	<i>broad</i>	brown	Y
XM_032921876	<i>ovo</i>	<i>ovo</i>	brown	N
XM_032927221	<i>lola</i>	<i>longitudinals lacking</i>	cyan	N
XM_032925470	<i>lola</i>	<i>longitudinals lacking</i>	cyan	N
XM_032921859	<i>exex</i>	<i>extra-extra</i>	cyan	Y
XM_032935662	<i>ap</i>	<i>apterous</i>	green	Y
XM_032935992	<i>ken</i>	<i>ken and barbie</i>	lightcyan	Y
XM_032927636	<i>Blimp-1</i>	<i>mammalian B lymphocyte-induced maturation protein 1</i>	pink	Y
XM_032923201	<i>br</i>	<i>broad</i>	pink	Y
XM_032928081	<i>pnr</i>	<i>pannier</i>	pink	Y
XM_032925696	<i>Spps</i>	<i>Sp1-like factor for pairing sensitive-silencing</i>	red	N
XM_032943003	<i>HHEX</i>	<i>Hematopoietically expressed homeobox</i>	red	Y
XM_032922964	<i>Awh</i>	<i>Arrowhead</i>	turquoise	Y
XM_032927749	<i>Awh</i>	<i>Arrowhead</i>	turquoise	Y
XM_032933804	<i>Awh</i>	<i>Arrowhead</i>	turquoise	Y
XM_032936275	<i>lola</i>	<i>longitudinals lacking</i>	turquoise	N
XM_032936557	<i>ken</i>	<i>ken and barbie</i>	turquoise	Y
XM_032930893	<i>Lim1</i>	<i>LIM homeobox 1</i>	turquoise	Y
XM_032929238	<i>sr</i>	<i>stripe</i>	yellow	N
XM_032922650	<i>ap</i>	<i>apterous</i>	Yellow	Y

Table 5: TF orthologs and their corresponding modules from the data of 8 days of gamma radiation exposure. TFs are considered activated if a corresponding enriched motif was found and the ortholog genes which encoded for the TFs were present in that module.

Transcript ID	Gene Symbol	Gene name	Module Name	Activated?
XM_032925433.1	<i>br</i>	<i>broad</i>	black	Y
XM_032930893.1	<i>ap</i>	<i>apterous</i>	black	Y
XM_032922964.1	<i>Lmx1a</i>	<i>LIM homeobox transcription factor 1 alpha</i>	black	Y

XM_032927749.1	<i>Lmx1a</i>	<i>LIM homeobox transcription factor 1 alpha</i>	black	Y
XM_032926422.1	<i>Sp1</i>	<i>Sp1</i>	black	Y
XM_032920460.1	<i>ken</i>	<i>ken and barbie</i>	black	N
XM_032921220.1	<i>hbn</i>	<i>homeobrain</i>	black	Y
XM_032927420.1	<i>dati</i>	<i>datilografo</i>	blue	Y
XM_032936085.1	<i>dati</i>	<i>datilografo</i>	blue	Y
XM_032934852.1	<i>dati</i>	<i>datilografo</i>	blue	Y
XM_032943597.1	<i>CG7368</i>	<i>uncharacterized protein</i>	blue	Y
XM_032935044.1	<i>PHDP</i>	<i>Putative homeodomain protein</i>	brown	Y
XM_032929033.1	<i>ttk</i>	<i>tramtrack</i>	brown	Y
XM_032936275.1	<i>lola</i>	<i>longitudinals lacking</i>	brown	Y
XM_032923756.1	<i>lola</i>	<i>longitudinals lacking</i>	brown	Y
XM_032923740.1	<i>srp</i>	<i>serpent</i>	green	Y
XM_032930933.1	<i>Klf15</i>	<i>Kruppel-like factor 15</i>	green	Y
XM_032928317.1	<i>exd</i>	<i>extradenticle</i>	green	Y
XM_032924480.1	<i>br</i>	<i>broad</i>	green	Y
XM_032939826.1	<i>Antp</i>	<i>Antennapedia</i>	green	Y
XM_032922650.1	<i>ap</i>	<i>apterous</i>	green	Y
XM_032925696.1	<i>btd</i>	<i>buttonhead</i>	green	Y
XM_032926090.1	<i>Dbx</i>	<i>Dbx</i>	green	Y
XM_032937297.1	<i>opa</i>	<i>odd paired</i>	green	Y
XM_032938232.1	<i>E5</i>	<i>E5</i>	green	Y
XM_032934371.1	<i>ken</i>	<i>ken and barbie</i>	greenyellow	N
XM_032942013.1	<i>lbl</i>	<i>ladybird late</i>	lightcyan	Y
XM_032934613.1	<i>fru</i>	<i>fruitless</i>	magenta	Y
XM_032929670.1	<i>Abd-B</i>	<i>Abdominal B</i>	Red	Y
XM_032941531.1	<i>br</i>	<i>broad</i>	Red	Y

3.1.9 Transcriptional regulatory network between 4 days and 8 days exposure to gamma radiation

Despite being qualified as a significant module; **dark turquoise** did not contain any enriched motif sites or TF orthologs and is therefore absent from the network (Figure 9). Except **light cyan** and **dark turquoise**, the rest of the significant modules shared a specific regulatory target, **blue** module. The functional enrichment analysis of the **blue** module shows that it is involved in the pathways related to metabolism of RNA and translational regulation. Due to the crucial role of the **blue** module in the network, the affected pathway may explain the overall radiation impact in a shorter period of gamma radiation exposure. Interestingly, the **blue** module is heavily transcriptionally regulated by transcription factors from almost all the other significant modules and self-regulation, but provides neither positive nor negative transcriptional feedback regulation to the other modules.

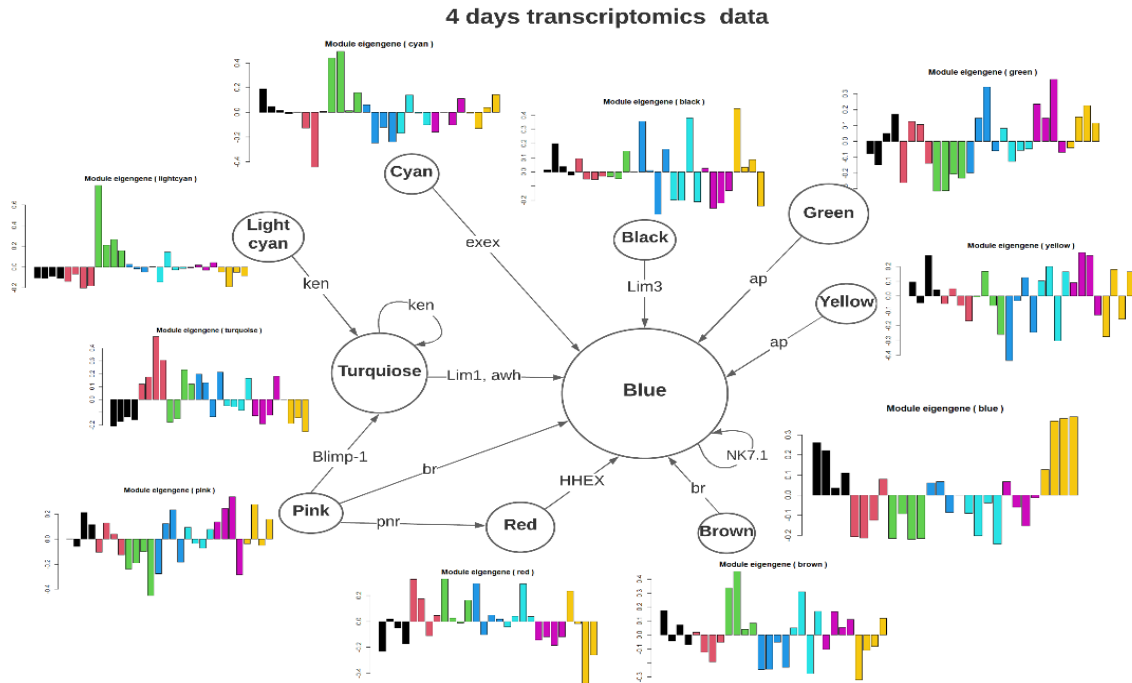


Figure 9: Network showing the changes of transcripts abundance and the eigengene expression across different dose rates. The size of the nodes corresponds to the number of enriched motifs. The pointing direction of the arrows indicate the regulation of TFs on the targeted node. The transcriptional regulatory direction is highly focused on the central module, blue, and shows that it is the key module as it contains the most motif binding sites.

In the 8 days transcriptional regulatory network, modules such as **lightgreen**, **midnightblue**, **darkgreen**, and **grey60** were excluded due to the absence of enriched TFs and motif binding sites. A more evenly distributed regulation between modules is shown in the network, with bidirectional transcriptional regulation observed between the biggest module, **green**, and the surrounding modules directly and indirectly such as: **magenta**, **red**, **black**, **brown**, **greenyellow**, and **blue** (Figure 10). The functionality of modules confirmed that longer exposure to gamma radiation affected various developmental processes to a larger extent; for example, axon guidance (**green**), signalling response from the immune system (**brown**), ribosomal biogenesis (**blue**), energy production (**black**), membrane trafficking (**red**), cell cycle (**red**) etc.

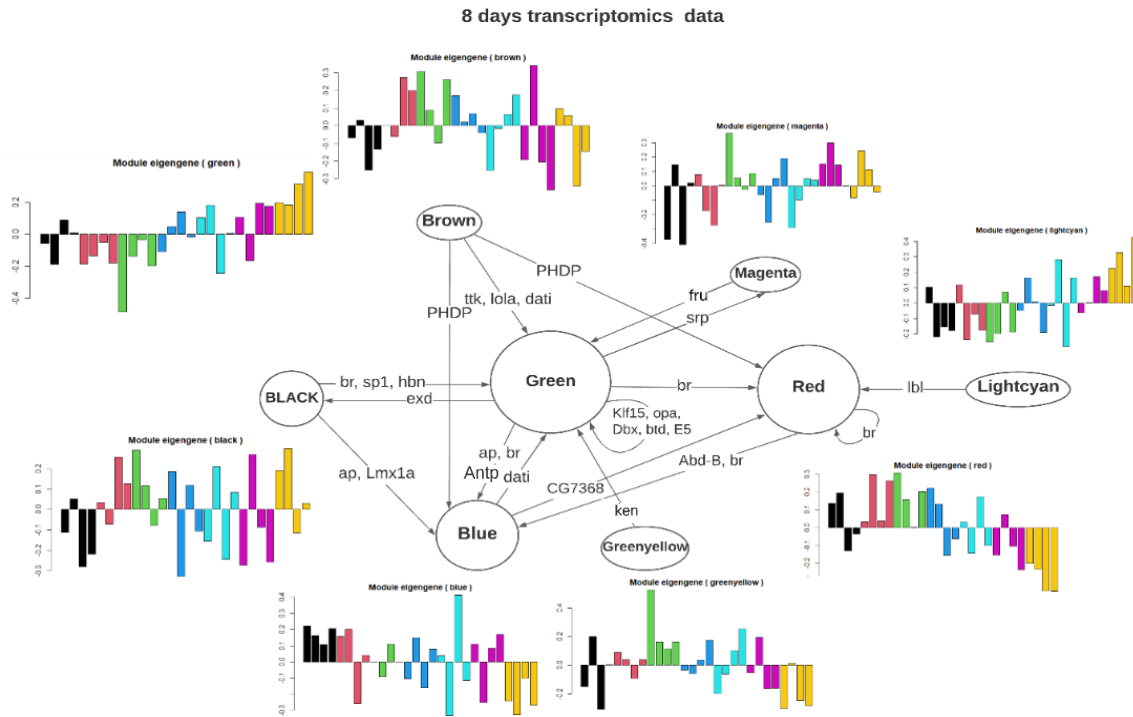


Figure 10: Network showing the changes of transcripts abundance and the eigengene expression across different dose rates. Size of the nodes are corresponding to the numbers of enriched motifs. The pointing direction of the arrows indicates the regulation of TF on the targeted node. The interactions between most modules were bidirectional.

3.1.10 Module similarity assessment to predict the outcome of 8 days of exposure to gamma radiation based on 4 days

Depending on the exposure length, any up-regulated genes could change to be down-regulated as the activation of different stress-defence mechanisms occur. To investigate whether the module eigengene expression generated using WGCNA aligns with previous deductions, a heatmap of module dissimilarity between the two exposure periods was plotted to demonstrate whether the eigengene expression of 4 days modules can be used to predict the eigengene expression of 8 days modules (Figure 11).

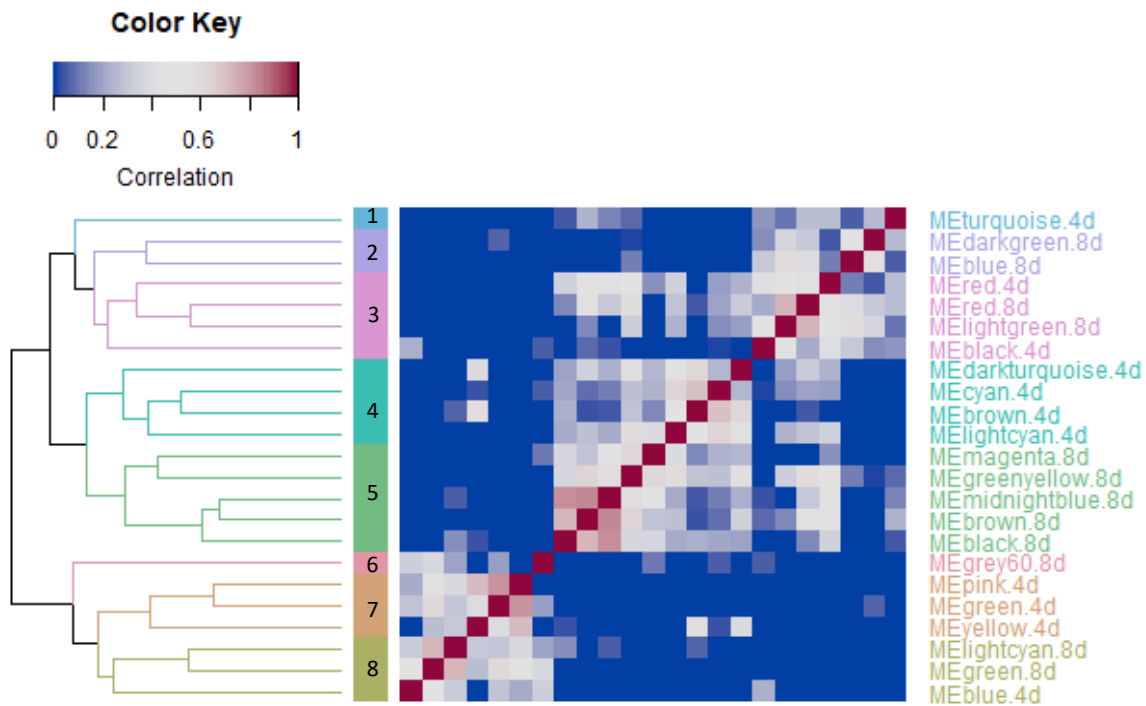


Figure 11: The eigengene expression of the 4 and 8 days module networks show an entirely different regulatory relationship. 4d and 8d in the end of every row name represents the radiation exposure period of each module.

The dendrogram initially branches into three clades followed by 8 branches which share a rather similar dissimilarity distance. The 8 branches represent the 8 clusters labelled with different colours. MEgrey60 from 8 days and MEturquoise from 4 days could be two outliers with very distinctive expressions when compared to the rest. Most clusters contain modules from the same period of exposure, except for cluster 3 and cluster 8. However, the branching splits of MEblack.4d was higher than MERed.4d of the same exposure period and within the same cluster. The same applies to MEblue.4d with a relatively higher dissimilarity than MElightcyan and MEgreen from the 8 days of exposure. The observation confirmed that the differences between module expressions from the same exposure period were mostly non-monotonic across different dose rates. As the modules of 4 days mostly did not blend into the same clusters with the modules of 8 days, the results again demonstrate that the gene expression at 4 days of exposure cannot be used to predict the expression at 8 days.

3.1.11 AOP Integration

The previous study (Song et al., 2020) proposed a hypothetical AOP to describe the impact of oxidative stressors on the reduced fecundity of aquatic life based on the detected GO terms. Unfortunately, key events like oxidative DNA damage, follicular atresia, mitochondrial

hyperpolarisation, and lipid oxidation that is central to the proposed AOP do not exist in GO terminology.

Integration of AOPs by matching the key words of key events with GO terms was demonstrated in Table 6 and Table 7. Key words such as “DNA damage”, “follicle cell”, and “lipid catabolic” were used in the search for key events that did not exist in the GO terminology. The modules consist of the enriched GO terms of key events labelled in green whereas the non-enriched terms are given a ‘tick’ to represent the presence of the gene. Toxicologically affected biological processes or signalling pathways that were discovered in other gamma radiation related studies (such as response to oxidative stress, cell cycle checkpoint and regulation, neurotransmitter signalling, neuron development, immune system process, inflammatory, autophagy, chitin development, and multicellular organismal development as well as lipid/protein/ion transport) have also been detected in both datasets disregarding the enrichment cut off of p-value < 0.05 (Song et al., 2020). The 4 days data shows that the enriched key events (green labels in Table 6) were distributed by 4 modules and most of the key events which were not enriched (ticks) were found in other modules. While in the 8 days data, key events were mostly found in the black module, more key events were found than in the 4 days data. However, there are three modules which do not contain any key events compared to 4 days which only has one module.

Table 6: Integration of significant modules with key events derived from AOP for the 4 days-transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a ‘tick’ to represent the presence of the gene in a specific significant module.

Key Events	Integration of significant modules										
	Blue	Pink	Red	Turquoise	Light cyan	Yellow	Brown	Green	Dark turquoise	Cyan	Black
Oxidative DNA damage	✓	-	-	-	-	-	✓	-	-	✓	-
Apoptosis	-	-	✓	✓	-	-	✓	✓	-	✓	✓
Follicular atresia	-	-	-	-	-	-	-	-	-	-	-
Mitochondrial hyperpolarisation	-	-	-	-	-	-	-	-	-	-	-
Oxidative phosphorylation	-	✓	-	✓	-	-	-	-	-	-	-
Mitochondrial ATP production	-	-	-	✓	-	-	-	-	-	-	-
Lipid peroxidation	-	-	-	-	-	-	-	-	-	-	-
Lipid storage	-	-	-	-	-	✓	-	-	-	-	-
Oogenesis	-	-	✓	-	-	-	-	-	-	-	-
Fatty acid oxidation	-	✓	-	-	-	-	-	-	✓	-	✓

Table 7: Integration of significant modules with key events derived from AOP for 8 days-transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they

are given a 'tick' to represent the presence of the gene in a specific significant module.

Key Events	Integration of significant modules											
	Black	Blue	Brown	Dark green	Green	Green yellow	Grey 60	Light cyan	Light green	Magenta	Midnight blue	Red
Oxidative DNA damage	✓	✓	✓	-	-	-	-	-	✓	-	-	✓
Apoptosis	✓	✓	-	-	-	✓	-	-	-	✓	-	-
Follicular atresia	-	✓	-	-	✓	-	-	-	-	-	-	-
Mitochondrial hyperpolarisation	-	-	-	-	-	-	-	-	-	-	-	-
Oxidative phosphorylation	✓	-	-	-	-	-	-	-	-	-	-	-
Mitochondrial ATP production	✓	-	-	-	-	-	-	-	-	-	-	-
Lipid peroxidation	-	-	-	-	-	-	-	-	-	-	-	-
Lipid storage	-	-	-	-	-	-	-	-	-	-	-	-
Oogenesis	✓	✓	-	-	✓	-	-	-	-	-	-	-
Fatty acid oxidation	✓	✓	✓	-	✓	✓	-	-	-	✓	✓	-

3.2 Multiomics integration of differentially expressed and co-expressed genes (DEACGS) with metabolite profiles

Supervised learning using DESeq2 revealed the differentially expressed genes with 1262 monotonic and 1202 non-monotonic responses from 8 days of gamma radiation exposure. While the gene expression reflects a genome wide response associated with different dose rates, it does not apply direct influence on the phenotypic changes. Due to factors such as epigenetic regulation, post-transcriptional and translational modification, inactivation of the substrates, and action of cofactors, there is a need to integrate the transcriptomes and metabolomes data to explain the observation and develop a more accurate mechanistic understanding of the phenotypes.

3.2.1 Differential expression analysis of metabolomics.

The PCA plot was made to examine the overall differences between the samples. No clustering patterns were observed, despite groups of samples being exposed to different dose rates (Figure 12 A). This indicates that the level of metabolite expression could be very small and so results in a small variation.

A total of 195 expressed metabolites were fit to a linear model and linear contrast, with 51, 93, and 123 of them differentially expressed in low-, high-dose responsive, and linear model groups, respectively. The abundance of differential metabolites and directionality of regulation is documented in Table 9. While there were 141 differentially expressed metabolites (DEMs) out of the 195 metabolites, the high dose responsive group shared almost half of the DEMs (62) with the linear model, corresponding to the findings of their overall module similarity (Figure 12B, C). Notably, the linear model group has the highest number of exclusive DEMs, despite the low dose rate responsive group having the most DEACGs. Interestingly, non-monotonic responses were observed with opposite trends in the two main DEM clades (labelled 1 and 2 in Figure 12 C) suggesting that two different sets of metabolites were produced in the low and high levels of dose rates.

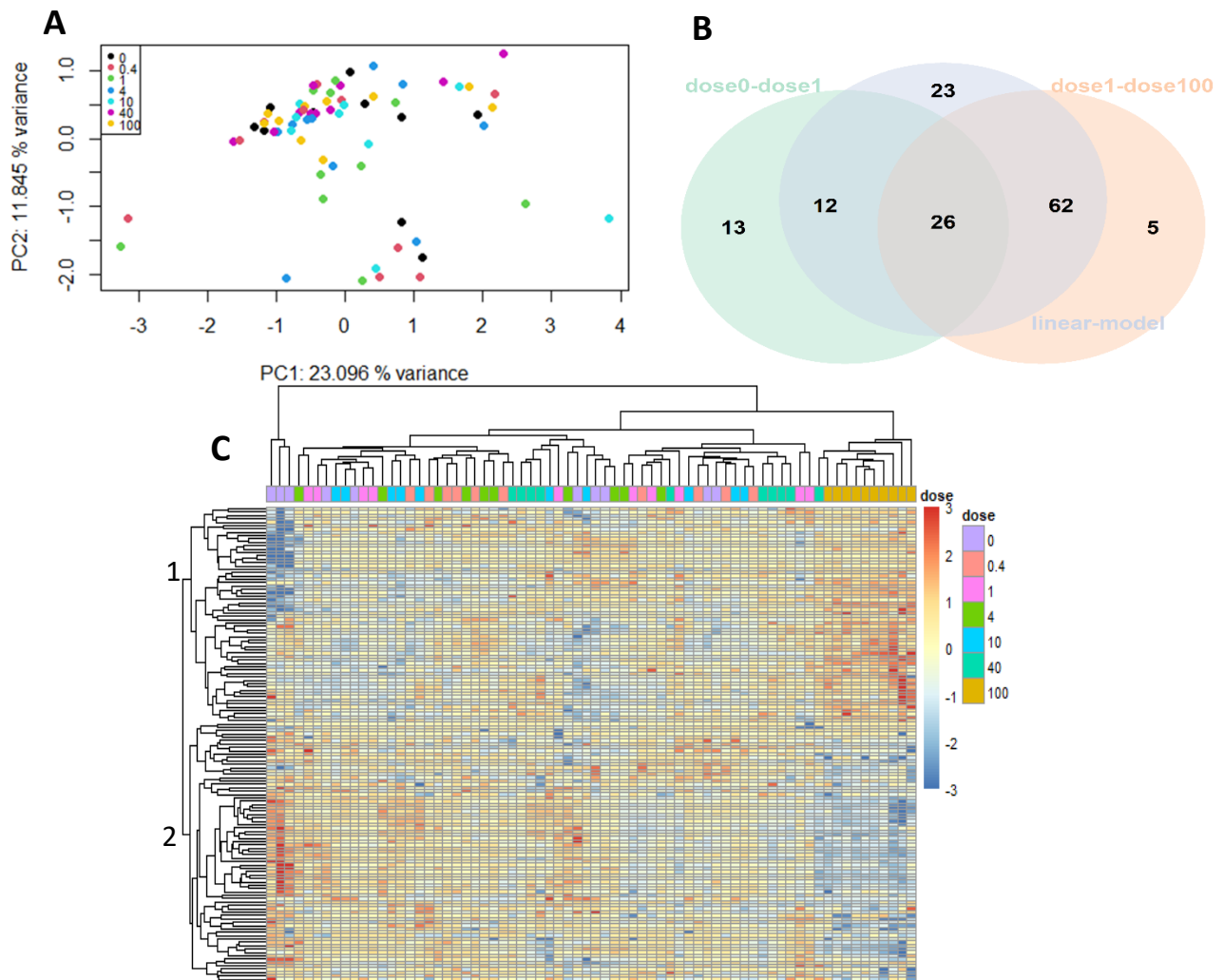


Figure 12: Plots show the initial analysis on metabolites abundance. **A:** PCA plot showed the sample of metabolites in 2D plane spanned by the first two principal components which explained the most variance. No clustering pattern observed indicates an extremely small difference between samples. **B:** Venn diagram comparing the numbers of DEMs shared and uniquely existing between low dose-responsive, high-dose responsive and linear model groups. **C:** Heatmap showing the gene expression of all metabolites and all samples; red to blue colour scale represents high to low gene expression and the colour of dose rates was

represented by the legend on the right.

Table 8: The number of differentially expressed metabolites from each group of the design model. The linear expression group has the highest number of metabolites, but it also shares about half of them (62 from Figure 12B) with the high dose rate responsive group.

Design model	No. of DEMs	Up-regulated DEMs	Down-regulated DEMs
Low dose rate responsive	51	23	28
High dose rate responsive	93	53	40
Linear expression	123	57	66

3.2.2 Generate DEACGs corresponding to the design models

To recap, Table 9 (A copy of Table 2) shows the overlapping of WGCNA modules and DEGs from DESeq2 (linear model and linear combination) narrowed down to 12 modules (blue, red, lightcyan, green, greenyellow, lightgreen, darkgreen, black, brown, midnightblue, magenta and grey60) to show a non-random association in the Fisher’s Exact test, meaning that they are significant modules.

Table 9: Significant modules from the previous chapter and their corresponding groups. Double asterisk (**) indicates modules that are exclusively found in the design model. The low dose rate responsive group has more exclusive modules than the linear model- and high dose rate responsive- groups.

Design model (WGCNA \cap DESeq2)		Modules with non-random association (p-value < 0.05) from 8 days gamma radiation exposure			
Linear model		Blue** Green	Darkgreen Greenyellow	Red Lightgreen**	Lightcyan
Contrast (mGy/h)	Low dose responsive (Ctrl vs 1)	Black** Greenyellow	Brown** Magenta**	Midnightblue**	
	High dose responsive (1 vs 100)	Darkgreen Greenyellow	Green Grey60**	Red Lightcyan	

The heatmap in Figure 13 demonstrates module eigengene expression across different dose rates. Samples from the lower dose rate groups (10 mGy/h and below) were not clustered according to the dose rates of exposure. Notably, modules from the low dose responsive group formed the middle clade (left dendrogram) and were mostly upregulated in the early increments of dose rate and downregulated in the mid dose range. Non-monotonic responses observed throughout the dose rates suggest genes within these modules are activated in early toxicological events. While the opposite expression direction was observed between the upper and lower clade (left dendrogram), significant modules from linear models largely co-exist in high dose responsive groups indicating that some modules from the linear model and high dose rates may have interacted antagonistically to each other.

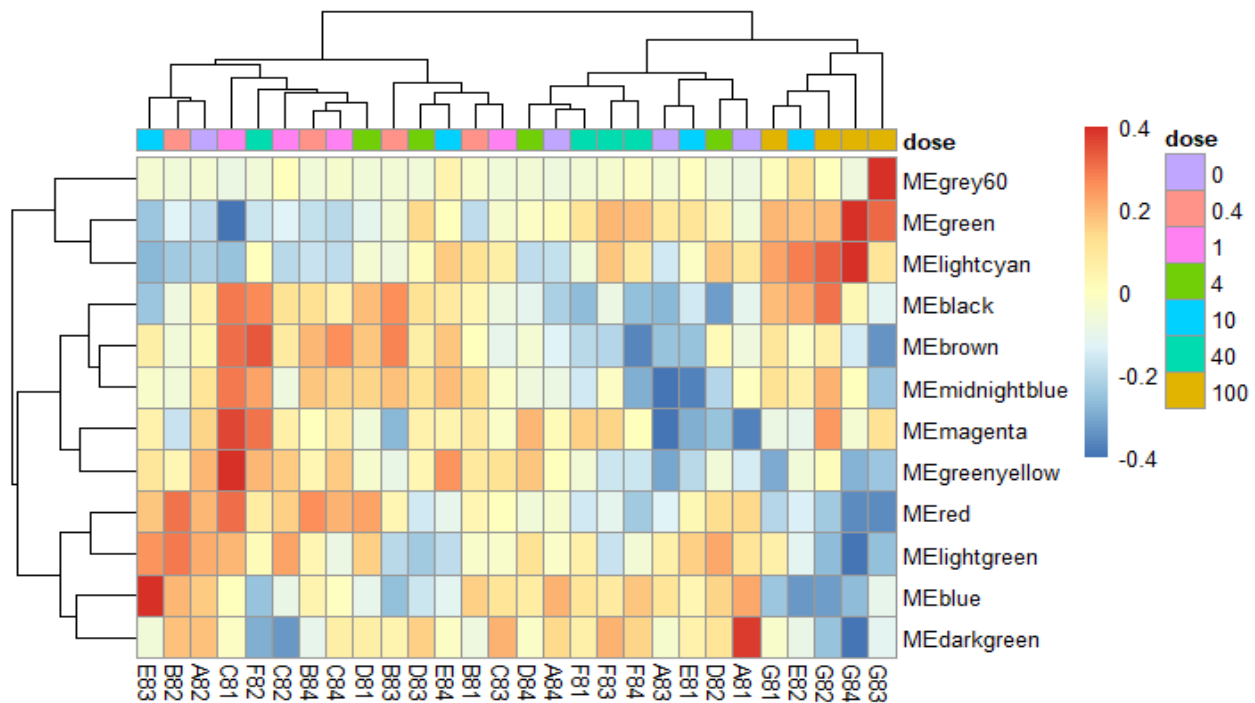


Figure 13: Heatmap showing gene expression profiles of all significant modules and all samples, red to blue colour scale represent high to low expression and the colour of dose rates was represented by the legend on the right.

Genes from the significant modules which qualified from the Fisher Exact test were combined into three groups: linear model, low dose responsive (0 vs 1 mGy/h), and high dose responsive (1 vs 100 mGy/h). The combination of unsupervised WGCNA and supervised DESeq2 detected all dose-rate-responsive genes to a greater extent and thus gave rise to more differentially expressed and co-expressed genes (DEACGs) being found.

3.2.3 Integrative pathway analysis (Paintomics3) (DEACGs and metabolites)

The pathway enrichment analysis was conducted using the KEGG database, one of the most comprehensive pathway collections in terms of molecular interactions with: manually curated maps categorised into metabolisms, genetic processes, environmental information processes, cellular processes, human diseases, drug developments, and organismal developments.

138 metabolites with KEGG compound IDs were converted to use their KEGG compound names, which resulted in 29, 52 and 74 DEMs matched with the KEGG Compound database. Approximately 50% of the expressed transcripts along with DEACGs were converted to the Entrez ID of *D. melanogaster* prior to integrative pathway enrichment analysis using Paintomics3. Based on the KEGG database, a total of 137 pathways were identified, with more than 60% being metabolic processes (predominated by carbohydrate, amino acid, lipid and glycan metabolism), 16.06 % and 9.49 % are genetic and environmental information processes, respectively (Figure

14). In terms of significant pathways, the low-dose responsive group surprisingly possessed 31, compared to the high dose responsive and linear model groups which had only 4 and 7 pathways.

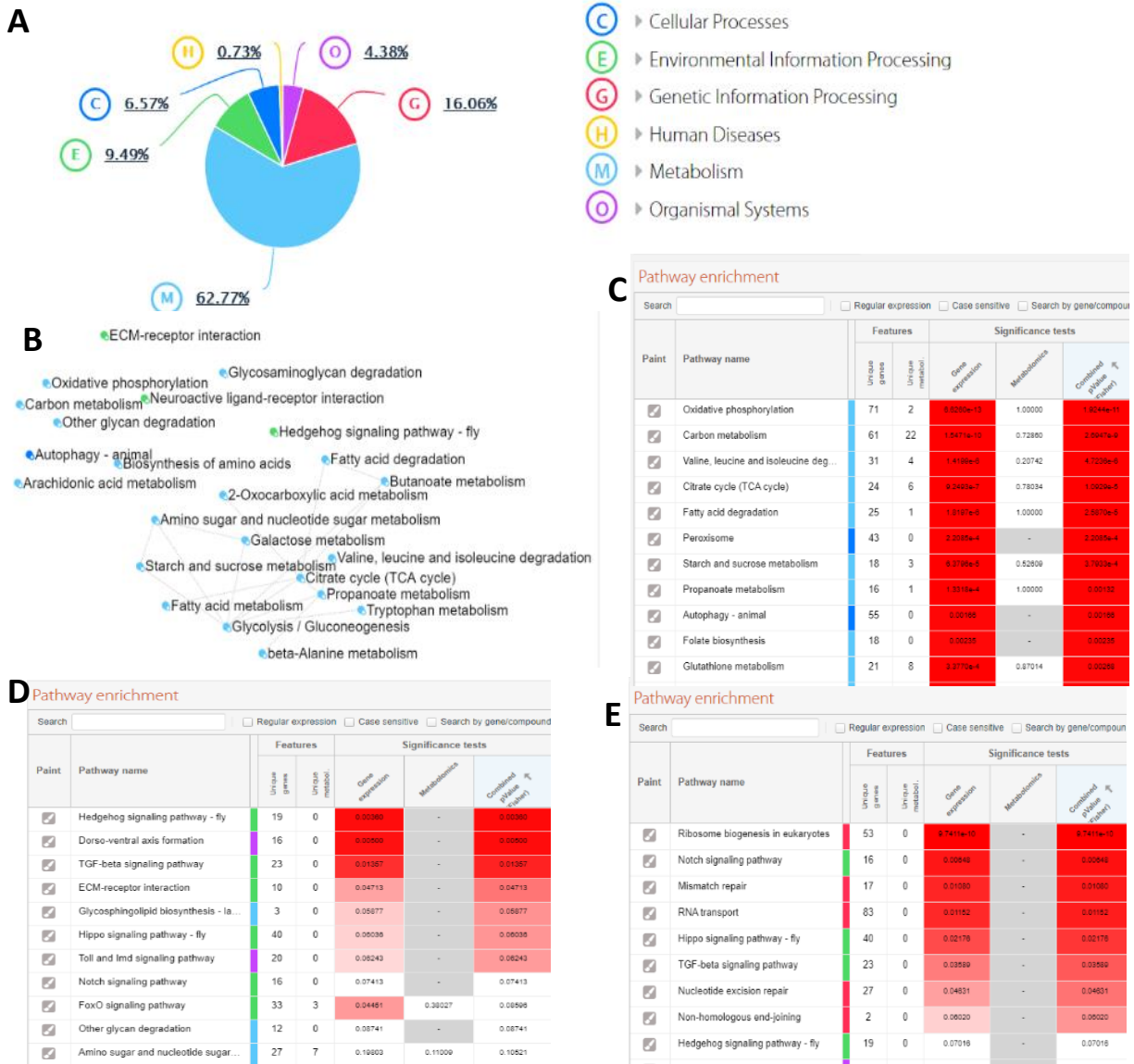


Figure 14: Parts of the output from Paintomics3. **A**: Pie chart demonstrating the pathway categories resulting from the overall transcriptomics and metabolites data. **B**: Pathway network from low-dose rate responsive group, the node represents the names of the pathway, and the edges show the shared features (DEACGs) between pathways. **C, D & E**: Pathway enrichment results from low dose-, high dose rate responsive group and linear model. Tables of enriched pathways are ordered in ascending order of P-value, the red colour intensity changes according to the level of enrichment/significance and the grey scale means no corresponding omics data was found in the pathway.

Table 10: Significant modules and the values in the bracket indicate the numbers of transcript paired with Entrez ID. Double asterisk (**) indicates modules that are exclusively found in the design model. The linear model- and high dose rate responsive-group share many modules, with the green module containing the highest number of genes. The blue module has the highest

number of genes overall and it is exclusive to the linear model group.

Design model		Significant modules (No. of transcripts matched with Entrez ID)		
Linear model		** Blue (516) ** Lightgreen (27) Green (396)	Darkgreen (28) Lightcyan (31)	Red (311) Greenyellow (111)
Contrast (mGy/h)	Low dose rate responsive (Ctrl vs 1)	** Black (369) ** Brown (427)	**Midnightblue (71) **Magenta (169)	Greenyellow(111)
	High dose rate responsive (1 vs 100)	Darkgreen (28) Red (311)	Green (396) ** Grey60 (13)	Greenyellow (111) Lightcyan (31)

In the low-dose responsive group, DEACGs (which comprised of black, greenyellow, brown, midnightblue, and magenta modules) were significantly enriched in oxidative phosphorylation, carbon metabolism, valine-leudine-isoleucine degradation, TCA cycles, fatty acid degradation, peroxisome, starch, and sucrose metabolism etc. 44 pathways from DEACGs alone were significantly enriched, suggesting genes from these modules were responsible for the adaptation and resistance in the early increment of gamma radiation dose rates by primarily altering their energy metabolism. However, none of the DEMs passed the p-value cutoff < 0.05, despite having several unique metabolites which were discovered in pathways such as carbon metabolism, ABC transporters, biosynthesis of amino acids, Aminoacyl-tRNA biosynthesis and so on. Significance tests of metabolomics revealed pathways like glycerophospholipid metabolism, histidine metabolism, glycine-serine-threonine metabolism, valine-leucine, and isoleucine biosynthesis and degradation, etc. bearing the smallest p-values of all, yet they were insignificant.

Meanwhile in the high-dose responsive group, only 4 enriched pathways were enriched by DEACGs, namely the Hedgehog signaling, dorso-ventral axis formation, TGF-beta signaling and ECM receptor interaction pathways. The FoxO signaling pathway was significantly enriched by gene expression (p=0.045) but lost its status after being combined with the p-value from the metabolomics data (p = 0.08). There were no unique metabolites or DEMs found in the 4 significant pathways. Top pathways bearing the smallest p values from the metabolomics data were involved in pyrimidine metabolism, phosphatidylinositol signaling system, amino sugar, and nucleotide sugar metabolism and so on. Nonetheless, the significant pathways from transcriptomics data further proposes that these modules in the later stages of dose rate increments were heavily involved in signal transduction, body formation and immune system development.

In Table 10, except for the grey60 module which consists of only 13 genes, the rest of the modules from the high-dose responsive group were shared with the linear model. Intuitively, significant pathways from this group were likely to also exist in the linear model. However, the linear model group consists of the largest module, blue, therefore significant pathways from the high-dose responsive group are mostly insignificant. A total of 7 significant pathways enriched in linear model group did not consist of any unique metabolites: ribosome biogenesis in eukaryotes, notch signaling pathway, mismatch repair, RNA transport, Hippo signaling pathway, TGF beta signaling pathway, and nucleotide excision repair.

Different from the low and high dose rate responsive groups, significant pathways enriched by DEACGs and metabolites from the linear model group suggest that changes in translation, DNA replication and repair as well as signal transduction were consistently and dominantly occurring throughout the exposure to gamma radiation at all increments of dose rate.

Although valuable insights were made incorporating biochemical perturbations into biological pathways, insufficient metabolites were identified and a lack of KEGG annotations in more than a quarter of metabolic compounds in the current study limited biological interpretation. In the KEGG database, the deposit of high-quality information such as cell, treatment, and species-specific pathways and metabolite information remains challenging. The absence of relevant information of *D. magna* in the KEGG database has forced the conversion of identifiers to Entrez ID of *D. melanogaster* leading to further lost data before the enrichment analysis.

3.2.4 Pathway and GO enrichment analysis (DEACGs)

3.2.4.1 Reactome Pathway Analysis

KEGG-based metabolomics pathway enrichment analysis showed that most metabolites were not significantly enriched, so the output of the transcriptomic pathway enrichment analysis were prioritised. In this part, all the genes (DEACGs) from significant modules were split into three different groups for transcriptomics pathway analysis using the Reactome database, another high quality manually curated database. Unlike the Paintomics3, Reactome pathway analysis does not take a customized background population into consideration. The major advantage of the Reactome database are the regular updates, including all curated entities which crosslink to other databases (Fabregat et al., 2017).

In Figure 15, the pathway output was different from the KEGG based pathway analysis, but they mostly belonged to the same categories and contributed to the same purpose. 39 pathways were significantly enriched by the DEACGs in the low-dose rate responsive group with the top pathways mainly being involved in energy metabolism: the citric acid (TCA) cycle, respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins, etc. The DEACGs input from this group was slightly up-regulated suggesting that the early impact of low-dose gamma radiation exposure resulted in a small increase in the gene expression and co-expression of the 39 pathways. On the other hand, significant pathways in the high-dose responsive group were involved in morphogenesis and the nervous system while those in the linear model group were mostly related to RNA metabolism. Despite sharing almost half of the genes, higher levels of gene up-regulation were observed in the pathways of the high dose responsive group whereas lower levels of up-regulation, and predominantly, down-regulated genes were observed in the significant pathways of the linear group. The findings agree with the output of KEGG based PA.

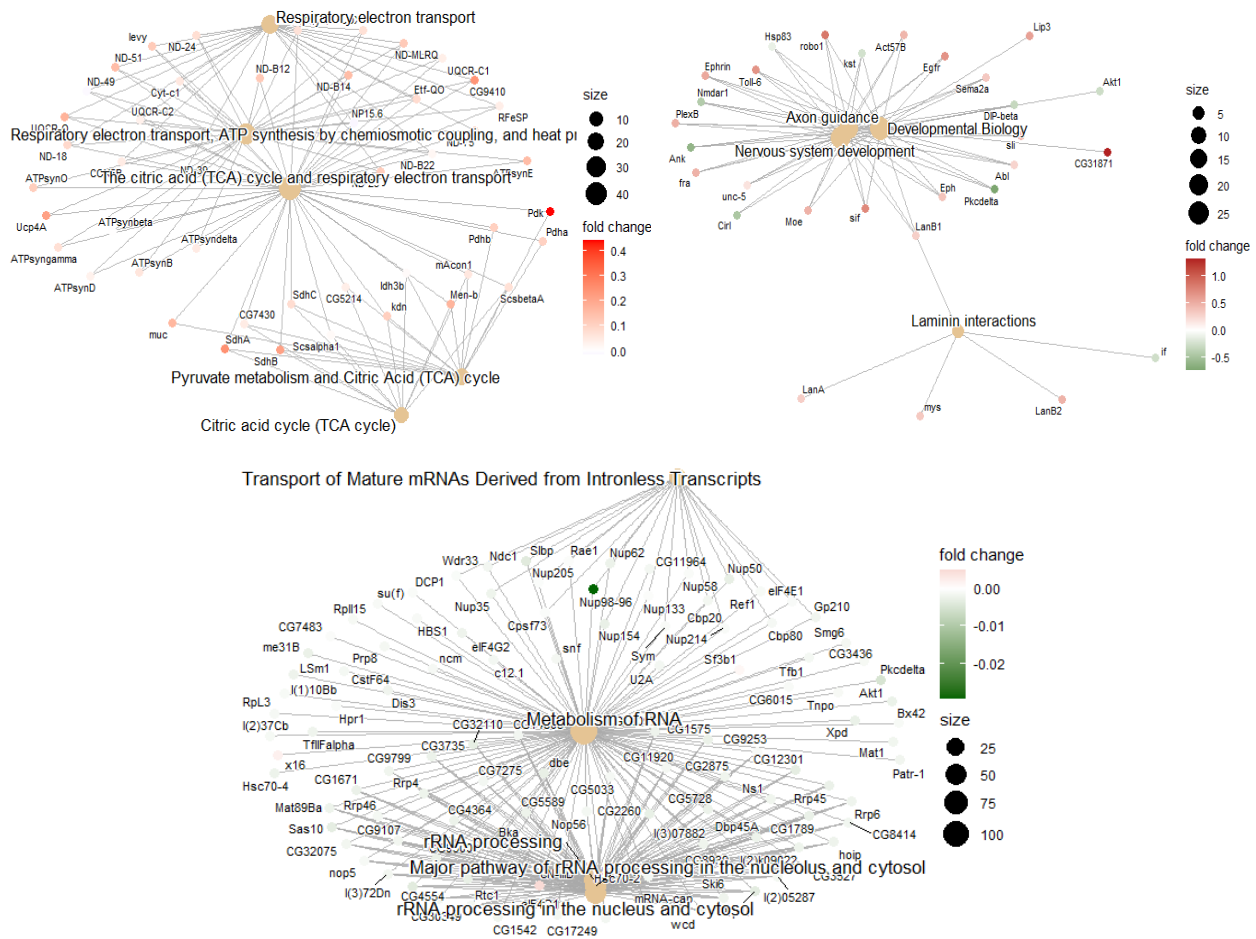


Figure 15: Network of top over-represented pathways and features genes associated with (A) low level of dose rates exposure, (B) high level of dose rates exposure and (C) linear response to gamma radiation. Nodes coloured in red, and green indicate whether the log fold change is positive or negative; the name of the node shown where the gene symbol corresponds to Entrez ID; the central node coloured in cream shows the name of the pathway with the size representing the number of genes involved in the pathway.

Table 11: The number of unique Entrez ID from each dose group and their composition to up- and down-regulation. The number of enriched pathways from Reactome PA is shown in the last column. The low-dose responsive group contains mostly up-regulated DEACGs in the enriched pathway, the high dose rate responsive group contains a similar amount of DEACGs in terms of the directionality, lastly the linear model consists of the most down-regulated DEACGs in the top enriched pathways.

Design model	Entrez (DEACGS) inputs in Reactome PA	Down regulated DEACGs	Up regulated DEACGs	No. of enriched Pathways (adj. p < 0.05)
Low-dose rate responsive	767	19	748	39
High dose rate responsive	726	353	373	37
Linear model	1208	836	372	70

3.2.4.2 The coherence of results between the gene sets annotated from GO- and pathway-analysis

The functionality of DEACGs was characterized by the overrepresentation of GO terms relating to biological processes. More than 97% of DEACGs exposed to early increments of dose rates (Figure 16, green labelled network) were downregulated (Table 11) indicating the disturbances of energy metabolism, induce of endoplasmic reticulum stress, lipid peroxidation, and catabolic metabolism. In the high dose rate responsive group, the enriched GO terms of up-regulated genes (such as immune response, neuron development, signaling process etc.) suggest high dose rates promote a critical inflammatory process and CNS injury which activates the immune system through signal transduction. Also, cell morphogenesis development proposes that the embryonic stem cells are predominantly affected because they are highly radiosensitive (Sanzari et al., 2013). The up-regulatory trends of genes suggest an irregularity such as uncontrollable cell division. As the high dose rate responsive DEACGs overlap with the linear expressive group, similar cluster annotations like cell morphogenesis development and response external biotic were also found in the linear group. The significantly down-regulated genes in the linear responsive group largely contributed to other clusters. Particularly in the rRNA regulation process and mitotic cycle, overrepresentation of relevant GO terms indicates a reduction of control over gene expression because ribosomal synthesis promotes cell growth and under normal conditions, it is under strict control to ensure a proper cell cycle, cell growth and proliferation (Chakraborty et al., 2011).

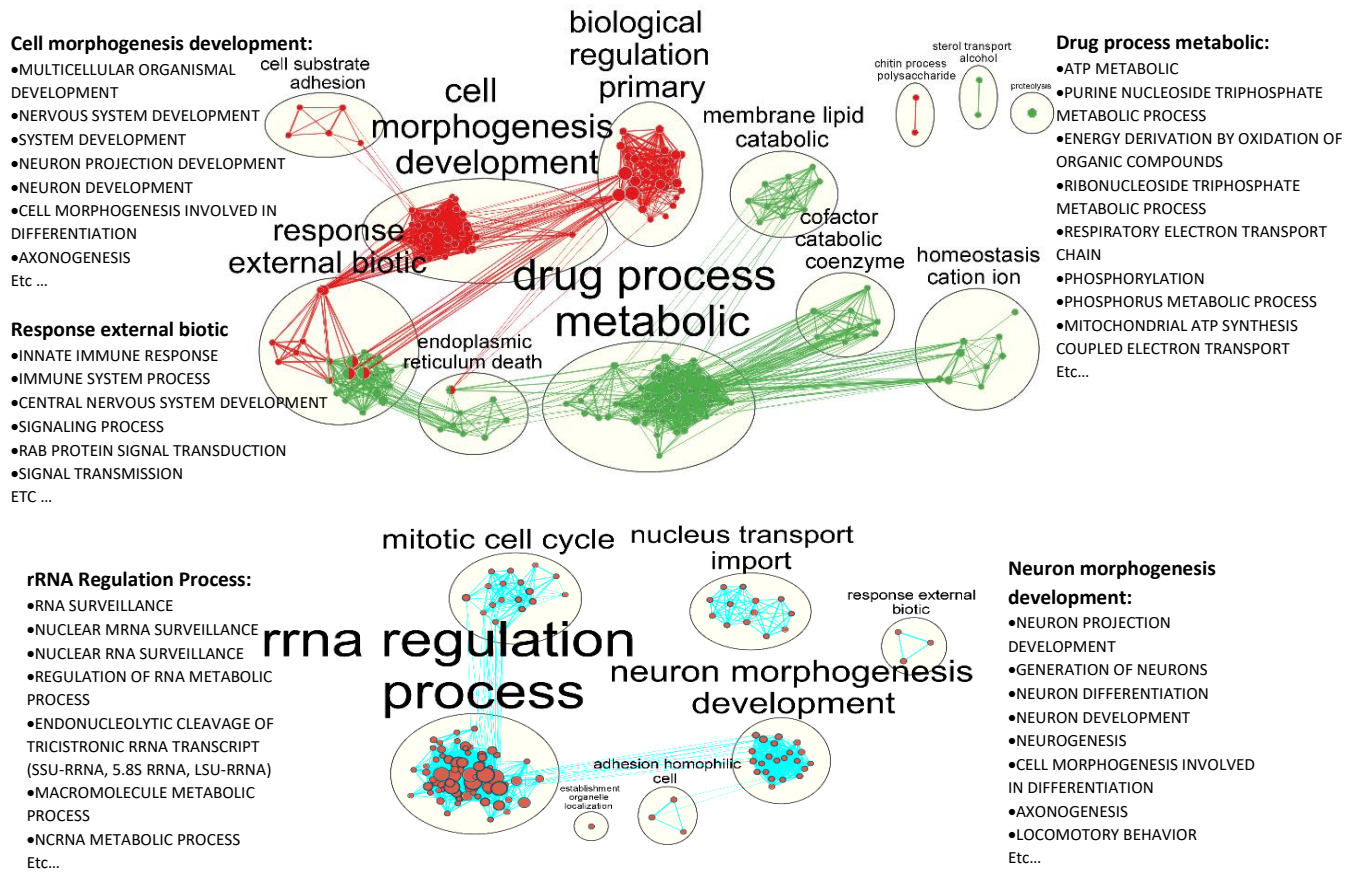


Figure 16: GO enrichment analysis made with the Cytoscape plugins BiNGO, EnrichmentMap and AutoAnnotate. Nodes represent the enriched GO term; node size corresponds to the number of genes and the thickness of edge depicts the number of overlapping genes

Chapter 4 Discussion

To recap, the first part of this proposed workflow created the regulatory network for short (4 days) and long (8 days) gamma radiation exposure periods by integrating supervised (DESeq2) and unsupervised learning (WGCNA), to reassemble the relationship within and between the modules and establish the differences between regulatory mechanisms activated based on the period of radiation exposure. The second part of this workflow integrated transcriptomics and metabolomics data with a focus on *D. magna* exposed to gamma radiation for 8 days using DEACGs generated from combining the significant modules and DEMs detected from limma, utilising only supervised methods. The characterisation of functionality based on the design groups of DEACGs and DEMs successfully unveiled the hidden mechanisms behind the key events of AOPs which is proposed to contribute to a reduction in fecundity.

In this chapter, section 4.1 discusses the technical challenges, limitations, and improvements that can be made in this workflow. Section 4.2 addresses the biological interpretations from the output of Part 1 of the workflow. Section 4.3 compares the coherence with existing AOPs, the performance of module integration with AOPs and potential future improvements. Section 4.4 covers the in-depth biological knowledge discovered using multi-omics. Lastly, section 4.5 explains the effects of gamma radiation on fecundity and potential further research.

4.1 Analysis and integration of data (technical discussion)

4.1.1 Non-conventional modules selection

Different from typical WGCNA proposed workflows which either correlate the module eigengene with conditional variables of interest or set a cut off value on the gene significance (GS) (correlating genes with variable of interest) and module membership (MM) (correlating genes with module eigengene) for module selection, this study utilised the intersection between DEGs and modules from WGCNA to discover important modules. This is because correlating the module eigengene with condition of interest uses Pearson correlation to measure only the linear relationships between the modules and traits of interest, meaning some transcriptional responses could be non-monotonic and could therefore result in a loss of information. Ideally, a monotonic dose response means that the increase of one variable results in relative changes/effects in other variables (Vandenberg et al., 2012). Whilst non-monotonic response refers to biphasic, U shaped and inverted U-shaped curves (Hong and Yang, 2017). Most biological responses are sigmoidal and non-linear. Measuring the values of GS and MM also makes the use of Pearson correlation, modules that pass the cut-off, again exclude the non-

monotonic expression, and lose some meaningful expression patterns whilst having extreme expression levels for some samples.

4.1.2 Software selections

Undeniably, there is still room for improvement in the software choices used in this study. The R packages DESeq2 and WGCNA are the most popular packages in omics analysis and therefore served as a blueprint to demonstrate the implementation of an efficient approach which incorporated supervised and unsupervised learning. While DESeq2 is known to be a very user-friendly software which also provides a very reasonable algorithm for DE analysis, it was also known for being more conservative when dealing with marginally expressed isoforms, therefore detecting fewer true positives (Khang and Lau, 2015, Mou et al., 2020). The ability for DESeq2 to detect small transcriptional changes from extremely low gamma radiation dose rates and non-monotonous pattern of expression at dose rate < 1 mGy/h is uncertain. On the other hand, module detection with WGCNA clustered co-expressed genes disregards any prior biological knowledge. Integrating the two methods covered up the drawbacks from each other and undoubtedly allowed for better exploitation of the high-dimensional data while retaining the biological sense. Better software algorithms can be employed in the future to replace the two current R packages for unsupervised learning and supervised learning.

4.1.3 Design model

In the past, a linear no-threshold model has been used in the risk assessment of radiation exposure assuming there was no threshold between radiation dose rates triggering cellular defence mechanisms. Such guidance is useful when the exposure dose rates are very high and acute, for example in the study of survivors, animals, and plants from the affected regions of atomic bombs. Recently, there is a growing interest in examining the health risks of long, chronic exposure to low dose rates of radiation. The applicability of a linear-no-threshold-model is challenging when the exposure dose rates are less than 100 mGy/h because the effects on cells are sublethal, a background influence, and cause the activation of natural repair. Sometimes, the adverse effects also require a longer period of exposure to overload the capacity of repair before causing real damage. As the effect of the lowest dose rates are variable due to animal species, age, radiation type etc., the experimental setup in this study has chosen to follow previous studies which investigated the effects of gamma radiation on the reproduction rate of *Daphnia magna*.

In this proposed workflow, the initial assessment on the expression profile showed that the radiation response of DEGs was non-monotonic because the direction and gradient of most DEGs in the low dose rate responsive- and high dose rate responsive groups were different. The response DEGs are sigmoidal, U shaped, inverted U shaped or biphasic. Therefore, apart from a linear model, linear combination was also implemented using software package, DESeq2, to demonstrate the dose dependent patterns. Since a dose rate of 1 mGy/h has been demonstrated

as the lowest observed effect level and employed a different mechanism affecting reproductive efficiency than high dose rates (100mGy/h), the comparison between control (background) vs 1 mGy/h and 1mGy/h vs 100mGy/h was performed.

One may argue that method is not without its weaknesses. Some DEGs that have a good fit to a linear model also belong to DEGs from the linear combinations. For example, a linear relationship can also mean there are significant changes between low doses (0 vs 1 mGy/h) and high doses (1 vs 100 mGy/h). However, this is not much of an issue because none of the eigengene expression profiles that overlap with the linear model show an exact linear relationship. As a result, there are likely very small numbers of DEGs that demonstrate this behaviour. The strength of this method is in discovering hidden non-monotonic responses of genes outweighed by their weakness by including genes that will not be picked up simply by using a GLM model. If one's interest is to focus only on the linear trend, it can be achieved in two ways following the workflow above: (i) select DEGs from a linear model that are not present in linear combination and look at their individual expression profiles; (ii) select DEGs that exist in the linear model, low and high dose group, and filter them according to the gradient of interest (logFC) from linear combination.

4.1.4 The conversion of identifiers and loss of information

Entrez records only document genomes that have been fully sequenced and those that are being actively used for research purposes (Wheeler et al., 2007, Maglott et al., 2005). It is probably the most widely used identifier for tracking other identifiers (RefSeq, GO consortium, Uniprot, Reactome, Ensembl etc.) which are also integrated in the Entrez system but named differently in other internal or external databases for gene-specific information. However, Entrez ID is not available for daphnids.

Meanwhile, the KEGG database has its own independent genome annotation system which contains two branches: KO (KEGG Orthology) assignment and KEGG mapping (Pathway, Brite and Module database) (Kanehisa et al., 2019, Kanehisa and Goto, 2000, Kanehisa et al., 2021a, Kanehisa et al., 2021b). The KEGG in-house annotation tools, BlastKOALA and KOALA or KAAS assigned KO IDs to genes corresponding to their own database, ID conversion between non-model and model organisms is smoother when KO IDs were adopted. KEGG annotation is available for *Daphnia pulex*, hence it was included in this workflow. On the other hand, metabolomics data were annotated using the KEGG compound database because the number of metabolites discovered in this study was fewer than 200, and KEGG compound ID are widely accepted by pathway visualisation tools. They also easily convert to other metabolites databases.

At least 50% of DEACGs and potential significant pathways were lost in the conversion of identifiers. Selection of the matched DEACGs and metabolites was based on lowest adjusted p-values for those sharing same identifiers with others, causing more differentially expressed features to be lost. Limited metabolomics annotation from mapping into mainstream databases led to a quarter of the data being lost and no differential metabolomics pathways being produced. Thus, current analysis did not take all transcripts, genes, and metabolites of *D. magna* into

consideration, but rather presents an unbiased quantitative approach to measure the significance of biological pathways for those that were mapped to their respective identifiers.

Functional and pathway enrichment analysis was conducted using multiple databases due to inconsistent nomenclature of pathways, accuracy of automatic pathway data curation, subsequent annotations, updates, and reconstruction. Manual curation of literature and pathway data is required to relate the observed differential metabolites with differential pathways. There are no specific guidelines when it comes to functional annotations, but with cross-referencing to multiple databases using popular and convertible identifiers, such as Entrez ID, end users of this workflow can easily retrieve the information of interest.

4.1.5 Significant modules in the network of transcriptional regulation

One gene module from the 4 days data and four gene modules from the 8 days data were classified as significant modules from the overlap between WGCNA and DESeq2. Reactome pathway enrichment and GO enrichment analysis confirmed their significance in response to oxidative stress. However, these modules were excluded from the network of transcriptional regulation because they contained neither enriched motif sites nor TF encoding orthologs.

This problem occurred because of the limitation of WGCNA in monitoring expression changes which are not gradual. Firstly, while WGCNA successfully clustered gene expression based on the co-expression of all samples, clustering methods neglected the effect of local co-expression among a subset of samples which could contribute to the activation of certain TFs (Neph et al., 2012b). The responsible genes which were co-expressed differently or oppositely for the same regulator were assigned into different modules. Using TF activity to bind specifically on motif sites requires ligand binding, but the activity profiles of TFs were not monitored in this study. This is also connected to our second point in which the combinatorial effect of proteins has not been taken into consideration. As one gene can only belong to one module, information regarding the combination of gene products such as TFs and hormones taking part in multiple pathways is again missed (Saelens et al., 2018). If the target genes of a TF lie within the absent modules, the expression of absent modules might be explainable by the expression of the TF (Marbach et al., 2012). One can thus deduce the regulatory relationship between genes and enhance the detection of modules. Methods like generalised singular value decomposition and bi-clustering have been proposed to overcome the limits of local co-expression and intersection of multiple modules because they allow the differential co-expression of genes between different samples. For bi-clustering, the prior classification of samples is not even required (Van Dam et al., 2018). Their flexibility in employing only a subset of genes to explain the variation of gene expression is very useful in TF regulation which is very context specific, yet these methods require a very high quality of data (Van Dam et al., 2018). This is because the theory behind these methods makes them very sensitive to outliers and affects the module eigengene expression.

Table 12: Numbers of genes in each module.

Module name	Darkturquoise-4	Lightgreen-8	Midnightblue-8	Darkgreen-8	Grey60-8
No. of gene	103	51	101	41	53

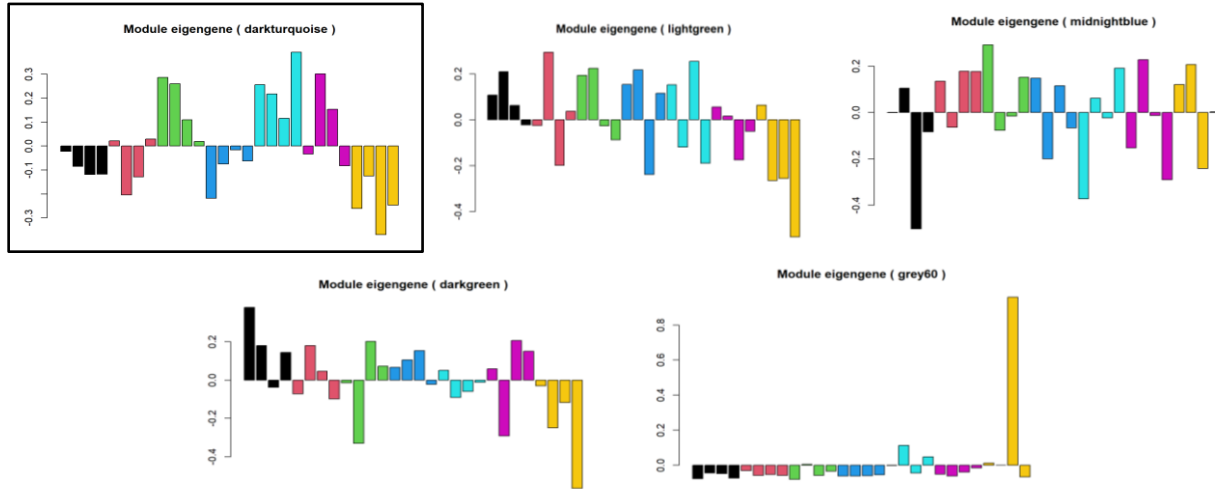


Figure 17: Significant modules that are excluded from the regulatory networks. The module with black border is from 4 days data while the rest are from 8 days data (top left).

4.1.6 Choices on motif database in AME

TF databases used in the AME analysis were from the fly databases available at MEME suite. The choice of databases determined the enriched TFs produced by finding a match between the input sequences to the motifs of interest. The TFs generated in this study were not directly related to radiation impact despite proposing the impact on phenotype appearances. The model organism, *D. melanogaster*, is not frequently used in studies investigating the effect of ionizing radiation compared to mammals and fish. Due to clinical concerns from existing medical treatments like radiotherapy, humans are the prioritized study target and relevant animals like mice and zebrafish have a closer biota to humans compared to flies. To improve the findings from this workflow, future study on the enrichment of TF binding sites can replace the motif databases with a more general choice for example “eukaryotes DNA”, “vertebrates (in silico and in vivo)” or “JASPAR CORE” (curated TF binding sites for eukaryotes) and “HUMAN AND MOUSE” to give results with different perspectives as mammalian and fish model species have undergone more extensive studies with ionizing radiation.

4.1.7 Confounding variables in the metabolomics data

The metabolomics data used in this study does not take the biomarkers of metabolites (blood, serum, interstitial fluid, etc.) and gender into consideration. The metabolomics workflow assumes all differences in all gene expression are simply due to the differences in treatment. Lacking this piece of information could result in bias in the data as metabolites from the samples which shared similar genetics are more likely to be treated as significant and produce a more confident result if they are the majority within each treatment. A minority of samples within the same block might be treated as outliers and the variance will be removed from normalization.

Besides, metabolites can be thought of as the downstream output from transcriptomic and proteomic activity. Detected differential metabolites may not explain the full picture of the transcriptomic changes despite being more sensitive and accurate to environmental and genetic changes than molecular observation. Future work beyond these two omics fields, such as including proteomic data, will improve the findings on the toxicological impact of low dose rate gamma radiation.

4.1.8 Conventional linear regression on the metabolomics data

It was decided to not include unsupervised machine learning with metabolomics data in this workflow. This workflow focuses on creating an efficient semi-supervised method through overlapping the linear combinations and linear model with unsupervised modules for highly dimensional data. While the author of WGCNA suggests sample size more than 30 is enough for the unsupervised learning, the availability on number of samples (n=70) with 10 replicates for each treatment and only 195 metabolites (features) from the metabolomics data, made the linear model sufficiently reliable when modelling the dose rate-dependent response (Langfelder, 2017). Even when the sample size is lower than 15, with clean data and strong differential signals, the output is likely not going to give insights which differs from what a conventional regression DE analysis could provide.

4.1.9 Creating DEACGs based on design model

Major weakness in the first part of this workflow is the exclusion of some significant modules from the transcriptional regulatory network. The selection on potential modules was based on non-random association (overlap) between the modules from WGCNA and DEGs from DEseq2 through the Fisher's Exact test. Some selected significant modules, despite proven functionally relevant to the impact of gamma radiation (by functional enrichment analysis), were however not included into the network of regulation. This problem was due to the choices on motif site database that were used in motif site enrichment analysis and the organisms used to look for orthologous TF encoding genes. As there are no motif database available for *D. magna*, motif databases of *D. melanogaster* and orthologs of TF encoding genes were used to find matching sequences for motif binding sites and the TF orthologs in *D. magna*. Besides, real time TF activities

are not monitored in this study. Novel TFs from *D. magna* are likely to be missed out in this case. Hence, the absence of TF motifs and TF encoding genes in the selected potential modules simply excluded the significant potential modules from the regulatory networks, but their importance should not be neglected despite having unknown regulatory contributions.

The integration of transcriptomics and metabolomics data hence began with a supervised learning method. To ensure the enrichment analysis produces statistically significance result, it required all significant modules and DEMS that correspond to the design models. The combination of modules produced DEACGs based on their design model group, resolving the issue of module exclusion, and yielding meaningful outputs in functional annotations.

4.2 Network based inter-modular transcriptional regulation

The transcriptional regulatory networks (Figure 9 and Figure 10) under chronic exposure to radiation at low dose rates suggests that different transcriptional regulatory mechanisms were responsible for the adverse outcomes, and contributions to the morphological and nervous development in *D. magna* in both study periods. The pattern of regulatory network and the module similarities at 8 days are different; therefore, the short-term exposure to gamma radiation yielded unpredictable results if the length of exposure increased. At 4 days, the transcriptional regulation between modules was highly centralised in the largest module, having the most genes, in which it possessed the most motif binding sites for being regulated by other modules and self-regulated at the same time. Meanwhile in 8 days, the transcriptional regulation between modules were not correlated to the numbers of genes. The regulation was more distributed between modules, most of them regulating and being regulated by each another. While the functional annotation of modules revealed that the impact of gamma radiation was different between 4 days and 8 days of exposure, the heatmap of modules similarity comparison (Figure 11) also supported the findings in which the clustering of all modules was based on their exposure period.

4.2.1 Networks of TF regulation after 4 days of radiation exposure

TF encoding genes usually possess multiple roles besides acting as a transcriptional regulator. The **Pink module** consists of TFs from the homeobox family and zinc finger protein family. The homeobox gene is known for encoding TFs that control the body plan and morphology of bilateral animals throughout evolutionary development whereas the proteins of the zinc family are more complex and serve different purposes in different taxa (Ferrier, 2016). The TFs of broad (*br*) and blimp-1 from the **pink** module belong to the C2H2 Zinc Finger family. Blimp-1 controls the differentiation of retinal cells and tracheal tissue. It also acts as a repressor controlling the

expression of mid-prepupal gene for the induction of ecdysone in (cultured salivary gland) which is responsible for metamorphosis, embryogenesis, progenitor cells and molting. Reduction in expression of *Blimp-1* is lethal to prepupal (Ferrier, 2016). On the other hand, the TF of *br* is transcriptionally activated when steroid hormone ecdysone is secreted. This explains the contrary expression profiles of the **pink** and **turquoise** modules.

Br is found in the photoreceptors of eyes and involved in eye disc morphogenesis. Recent studies, which investigated the effect of excessive sex hormones on non-reproductive organs, showed that over-proliferation and mis-differentiation of intestinal stem cells due to excessive ecdysone promotes gut dysplasia and tumorigenesis. The TF of *Br* is hence activated to suppress cell proliferation (Ahmed et al., 2020).

The gene of the third enriched TF, pannier (*pnr*) has been reported to express in dorsal mesoderm (Mandal et al., 2004). It is responsible for thoracic development, abdominal segmentation and head structure around the eyes (Ferrier, 2016). The presence of GATA TF pannier stimulates cardiac mesoderm to form part of the dorsal mesoderm and contributes to the lymph gland. Due to functional conservation among bilaterians, GATA TF pannier was also reported to trigger cell differentiation in cardiac cells, sensory organs, and the dorsal thoracic closure of embryos (Immarigeon et al., 2019). Previous studies also showed that the GATA TF of *pnr* requires the presence of LIM domain proteins to ensure proper specification of cardiac primordium in invertebrates. Interestingly, genes containing the LIM domain: *ap* (**green, yellow**), *Lim1* (**turquoise**), and *Awh* (**turquoise**) which are commonly expressed in the muscles, heart, brain, nervous system, lymph glands, early stage limb development, and the dorsal ventral axis of eyes in *D. melanogaster* were found and enriched in the 4 days TF regulatory networks (She et al., 2021). LIM domain proteins belong to a subfamily of the super class – homeobox. LIM domain proteins exist in the form of TFs or structural proteins, both can regulate cardiac and hematopoietic systems: *Lim* homeobox 1 (*Lim1*), *Lim* homeobox3 (*Lim3*) and *ap* are expressed exclusively in the different subtypes of motor-neurons along the ventral cord and interacted closely with the nervous system. LIM mRNAs were found in lymph glands, muscles systems, and the circulatory system while the TFs of LIM domains were involved in the regulation of *serpent* (*srp*), one of the three TF members of the GATA family which control the differentiation of gland, cardiac or pericardial cells. Other TF encoding genes like *awh*, *blimp-1* and *ap* were also present in the lymph gland, central nervous system (CNS) and brain of eukaryotes (She et al., 2021).

NK-like TFs also belong to the homeobox family with TFs such as *exex* (**cyan**), *NK7.1* (**red**) and *HHEX* (**blue**). Studies in development biology revealed that the NKL genes are primarily involved in neural development. Gene products of *exex* from **cyan** module were documented to negatively interact with the TFs of *Lim3* in neuronal differentiation. TFs of *exex* regulate the fate of cells within the motor neurons. Meanwhile, the TF role of *hhex* and *NK7.1* were reported to be varied across different taxa. In *D. melanogaster*, the expressions are observed in midgut primordia, gut section, brain, and CNS. However, research in vertebrates has shown their role as a

transcriptional repressor to inhibit cell hyperproliferation and other oncogenic activity (Marfil et al., 2015, Treffkorn and Mayer, 2019).

In addition, orthologs of genes encoded for activated TFs: *ap* (green, yellow), *br* (pink, brown), *ken* and *barbie* (light cyan, turquoise), *awh* (turquoise) were presented more than once, with only *br* demonstrating different isoforms. While the regulation of *ap*, *awh* and *br* are highly relevant to CNS, the *ken* and *barbie* (*ken*) encodes TFs that are responsible for the formation of genitalia and animalia in the juvenile stage.

4.2.2 Investigating a cyclical relationship in the 4 days regulatory network

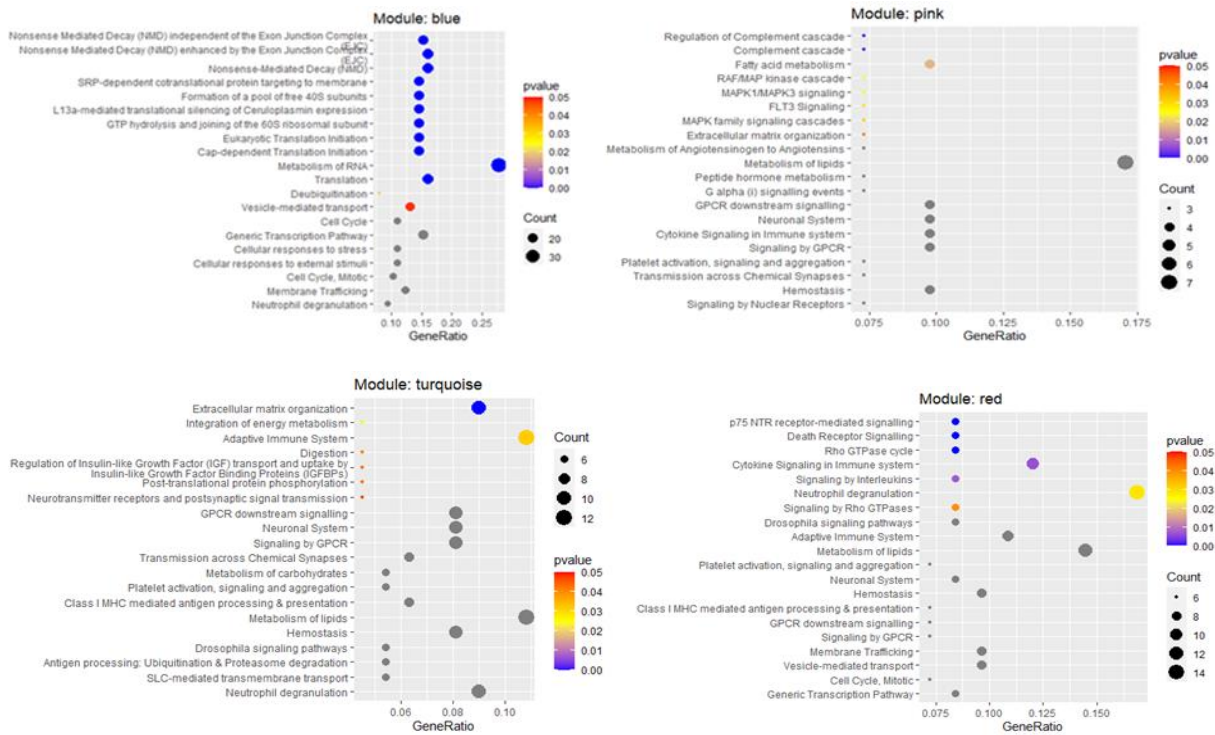


Figure 18: Top over-represented pathways from Reactome pathway analysis in blue-, pink-, turquoise- and red modules.

The **blue** module showed the most motif binding sites detected by AME yet contained no orthologs of TF encoding genes. The **blue** module also presents a distinctively different expression profile than the rest of the modules in the network below: the plummeting of gene expression upon exposure to the gamma radiation stayed rather consistent, until a sharp rise as the dose rate reached 100mGy/h. The **blue** module was self-regulated by *NK7.1* and regulated by at least one active TF from other significant modules in the network, except **lightcyan**.

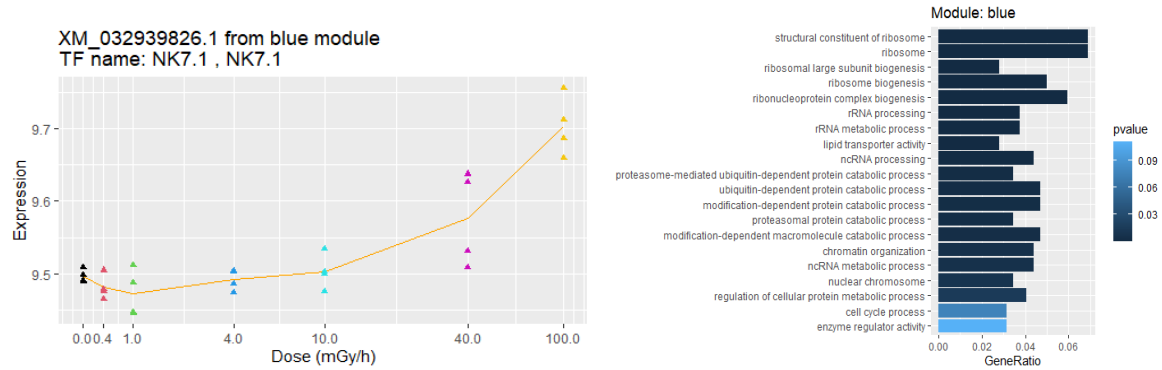


Figure 19: The gene expression of NK7.1 from the blue modules (left) and the output from GO analysis (right).

Turquoise was another self-regulated module by the TF *ken* within the same module and by the same TF from the **lightcyan** module on the same motif. The combined module eigengene expressions of lightcyan and turquoise from Table 2, which contains the differential expression of *ken* (Ctrl vs 1mGy/h) with the *ken* expression profile shown in Figure 20, displays that the initial increase of transcript abundance from the **lightcyan** module at 0.4 mGy/h triggered a drastic increase in the gene expression of the turquoise module. This could indicate the promotion of transcription by the active TFs of *ken*. Further increases of dose rate at 1 mGy/h spiked the transcript abundance of *ken* to the highest level, yet potentially triggered the deactivation of TFs and reduced the module expression of **turquoise**. Genes from the **turquoise** modules participated in a wide range of activities related to the resistance and adaptation to gamma radiation such as immune response, extracellular matrix organisation, digestion, signalling and growth development.

Despite sharing the same motif binding sites and orthologous relationship with *D. melanogaster*, *ken*-encoding genes from **turquoise** and **lightcyan** have an entirely opposite expression profile across all dose rates (Figure 20): “n” shaped vs “increasing” trend changes observed at low dose rates (0 to 1 mGy/h) and “raising” vs “decreasing” changes as the dose rate increases from 1 to 100 mGy/h. The observed TF activities of *ken* in *D. magna* could be functionally divergent from *Drosophila* and further study is required to explore the existence of *ken* isoforms where the TFs generated might act antagonistically towards each other. The **lightcyan** module is mainly involved in cellular defence and signalling pathways.

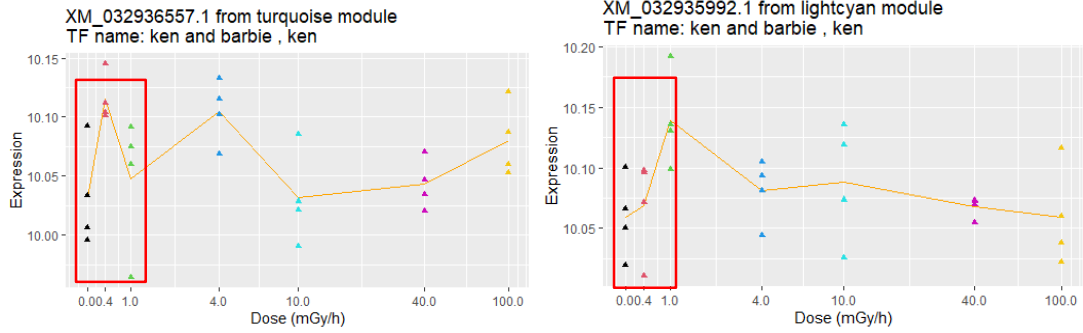


Figure 20: The gene expression profile of ken and barbie from the turquoise and lightcyan modules.

The TFs from the **turquoise** module are antagonistic to the **blue** module as shown by the TFs: *Lim1* and *Awh*. A similar expression change can also be seen in the **red** module, with *HHEX* as the active TF acting on the **blue** module. The **pink** module regulated the **blue** module directly with *br* and potentially exert an indirect effect on the **blue** module with TFs (*Blimp-1* and *pnr*) acting on the **turquoise** and **red** modules. Distinct opposing trends were observed in the expression profiles of the TF encoding genes of the **pink** module, from the **red** and **turquoise** (Ctrl vs 1 mGy/h) modules.

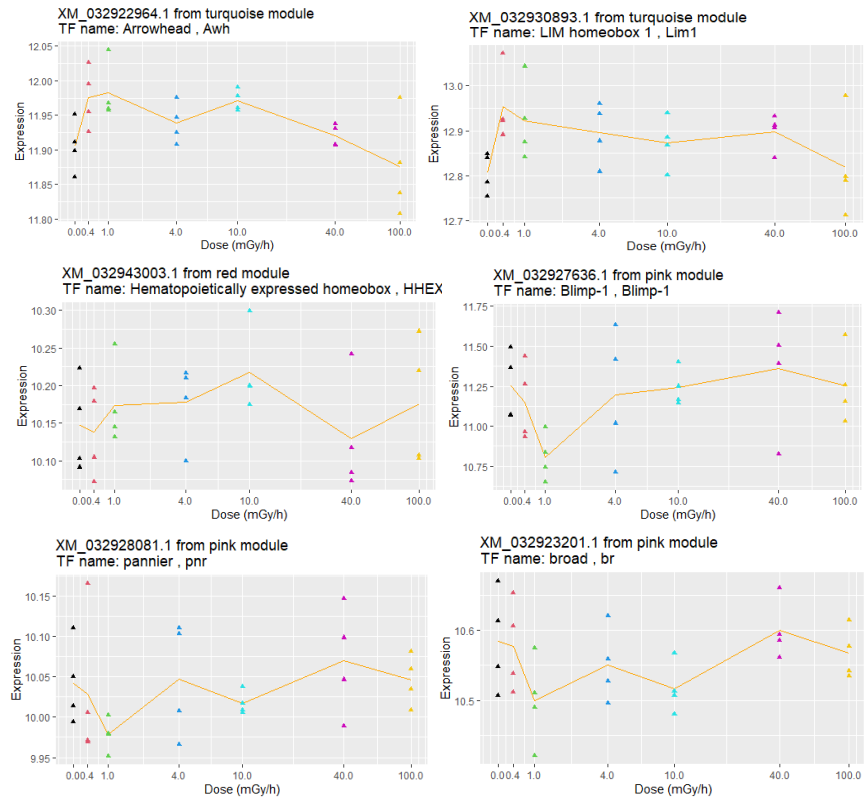


Figure 21: The expression profile of TF encoding genes from the turquoise-, red- and pink modules.

4.2.3 Module specific findings from a longer exposure to gamma radiation

In comparison to TFs from the 4 days network, *br* and *ap* were presented more often than the rest, with three different isoforms of *br* orthologs (I, J and L) and one isoform C for ortholog of *ap* in the 8 days network (Figure 10). The expression of *Br* in the neuron leads to the pathway of neuron maturation which links to certain levels of behavioural control in *Drosophila* (Li et al., 2004). While the network of 4 days showed that only isoform J of *br* was presented multiple times, except for isoform Q and P (Br Z1-Z4) which were found earlier. Unfortunately, few studies have been carried out on the detailed functionality of these isoforms as they were only discovered in the last decade.

The gene, *ap*, is most known for being involved with normal dorsal ventral and wing formation, but studies have found the expression of *ap* in many tissue types are related to reproduction, central nervous systems, brains and viability in *Drosophila*. Recently, a study focusing on the wing patterning of hemimetabolous insects refined that the role of *Apterous A (apA)* functions as both activator and repressor for wing size, patterning (sexual trait), bristle formation and potentially ventral development of eyespots, while *apB* serves as a backup of *apA* with minor defects if mutated (Liu et al., 2015, Prakash and Monteiro, 2018). The TFs of *apA* and *apB* exist exclusively in the network of 4 days but their morphological effect on *D. magna* needs further investigation. On the other hand, only isoform C of *ap* (*apC*) was found twice in the **green** and **black** modules' 8 days radiation exposure. *apC* is involved mainly in neuronal fasciculation and the ectopic expression of *apC* was reported to have minor defects on dorsal wing discs but caused a large reduction in juvenile hormones and uncoordinated movement (Lundgren et al., 1995). Despite sharing the same ortholog of *apC*, the observed difference in expression profiles at low dose rates (0 – 1mGy/h) could be explained by the spatial expression of *ap* on the two transcripts in the different tissues of *D. magna* (Figure 22).

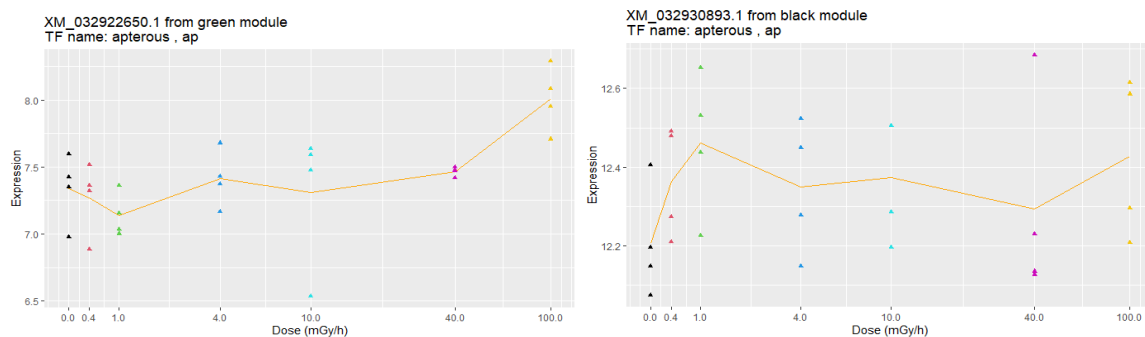


Figure 22: The expression profile of TF encoding genes from the green and black module.

In contrast to the network of 4 days data, the transcriptional regulation between modules from 8 days data is more evenly distributed. Forward and backward regulation were observed in most modules and the regulation between modules involved more than one target. However, modules

such as **lightgreen**, **darkgreen**, **grey60**, and **midnightblue** were not shown in this network because no TF orthologs or motif enrichment sites were found.

While the 4 days-network captured one of the three GATA ZF homolog, *pnr*, the 8 days-network presented another GATA ortholog *serpent* isoform E (*srpE*). The absence of *pnr* at the late stage of embryogenesis is expected, because previous studies have shown the repression of *pnr* was either mediated or directly controlled by *Abd-B*. The expression of *pnr* is specifically inhibited during the formation of spiracles.

No detailed functionality of *srpE* is documented, but the TF of *Srp* is best known for haematopoiesis by acting as an activator of an enhancer for metamorphosis in response to ecdysone in the final developmental stage of larval. Overall, prolonged radiation exposure activates more TFs from a variety of gene groups such as: Pair-like homeobox, Kruppel-like family (*KLF*), and TALE homeobox. TF-encoding genes like *br*, *Exd*, *Lmx1a*, *sp1*, *E5*, *lbl*, *btd*, *opa*, *klf15*, *hbn*, *CG7368*, *lola*, and *ttk* which were exclusive to the 8 days network are related to the growth of organs and body plans. Studies of other bilaterian with ectopic expression of these genes are accompanied with organ failure, cardiac defects, neurodegeneration, tumorigenesis, and metabolism disorders (Ahmed et al., 2020, Cheng et al., 2016, Chung et al., 2014, Qian and Bodmer, 2009, Dinges et al., 2017, Petersen et al., 2013). Longer exposure to gamma radiation will therefore compromise fitness and increase the susceptibility to diseases which potentially reduce the lifespan of the studies' subjects. Meanwhile, ectopic expression from other active TF-encoding genes like *lola*, *dati*, *fru*, *Lmx1a*, and *Abd-B* also implicate abnormalities in the brain, endocrine systems, reproductive organs, and nervous system development which can lead to irregular mating behaviour and a reduction of fecundity (Newell et al., 2016, Schinaman et al., 2014, Allbee et al., 2018).

These findings align with the age and development of experimental subjects, during which the female daphnia would become fertile. According to a previous study focused on fecundity in *D. magna*, most daphnids were transforming into an adult at this stage (8 days) and the embryo within was visible but not yet released. The integration of TF motif binding with module eigengene expression profiles in 4 and 8 days demonstrates a temporal pattern of expression, enabling a quick and comprehensive assessment on early toxicological events.

Reactome pathways and GO analysis from the largest module, **green**, showed a focus on the adult developmental process and signalling pathways (eg. Nervous system, neurogenesis and WNT signalling). Other highly interconnected modules in the network (**black**, **blue**, **brown**, and **red**) also revealed their significance in transcriptional regulation, ATP metabolism, cell cycle regulation, cell-to-cell signalling, reproduction, chitin synthesis, DNA damage checkpoints, DNA repair, and the immune system.

Overall, interesting insights were discovered from the exposure-period-specific modules and genes responsible for transcription factor encoding and other biological purposes. However, TFs that were linking modules in the regulatory network were considered activated as the transcripts

level of TF encoding genes correspond to the eigengene expression, but there is no way to confirm whether it is a negative or positive regulation. Real time TF activity monitoring could become the focus of future projects focusing on transcriptional regulation.

4.3 Limitation of GO-based module integration with AOPs and the new prospects

Gene ontology analysis is one of the most popular approaches for providing biological interpretations due to its flexibility in connecting any gene and gene products regardless of the upstream analysis procedure. In common practice, GO annotation relies heavily on hypergeometric tests to identify over or under-represented GO terms for a group of genes of interest. However, the hierarchical structure of GO in the form directed acyclic graphs (DAG) makes the interpretation challenging, especially when there are large numbers of GO terms (Manjang et al., 2020). The hypergeometric approach simply neglects the fact that GO terms are highly dependent on each other and possess a disadvantage in the integration of AOPs in this workflow.

Semantic measure which uses the frequency to measure the distance of GO terms has been proposed as a solution, but studies show that bias due to preference on biological applications will compromise the accuracy of the result (Mazandu and Mulder, 2014). Limiting a certain degree of GO connections in the enrichment analysis is another option but this method is very complicated and uncommon (Grossmann et al., 2007). Therefore, future work should focus on exploring existing automated tools that can further categorise GO terms into different GO levels (e.g., regular nodes, jump nodes or leaf nodes) (Manjang et al., 2020) to simplify structural information and for efficient DAG interpretation.

In this study, the integration of GO and AOPs was accomplished manually. Despite functionality assessment from Reactome PA and GO analysis discovering many toxic pathways and biological processes, Table 6 and Table 7 that matched the key events with the key words of GO term from Part 1, showed that many modules did not contain the expected key events. For example, the green module from the 8 days data consists mostly of genes and motif sites for TFs, like *sp1* (response to DNA damage, apoptosis, chromatin remodelling), which were not enriched with relevant GO terms of those cellular responses. As such, the automatic integration might need more specific names of key events or expansions on the types of key events to integrate massive numbers of GO terminology into the AOP.

In comparison to the previous study (Song et al., 2020) which discovered many enriched GO terms and turned it into the AOP, less enriched GO was found in this study. The differences in outputs were most likely caused by the inputs to the enrichment test, in which the previous study used all genes, whereas this proposed workflow used WGCNA to split the genes into different clusters before being piped into the enrichment individually. The enriched GO term is hence

module specific. The non-enriched GO terms in this study could simply just be ‘diluted’ because genes that possess the same functionality demonstrate a different expression pattern and are assigned into different modules. Therefore, to produce a result that is similar to the findings of Song et al. modules can be combined and turned into DEACGs, which is design-group specific for GO analysis in future studies, to reduce the effect of dilution. However, this method will neglect the advantages of module specific findings.

Tables below show the parent term or child term of GO terms that linked to the key events were assessed manually to demonstrate the potential of modules in the proposed AOP. For example, ‘lipid peroxidation’ was replaced with ‘lipid catabolic’ in the search, and child terms such as ‘glycolipid catabolic’ and ‘sphingolipid catabolic process’ were frequently found in many modules. Another example for mitochondrial hyperpolarisation was replaced with ‘synaptic transmission’ and ‘voltage-gated potassium channel’ due to the role of hyperpolarization-activated cyclic nucleotide-gated cationic (HCN) channels, ATP gated potassium channel, and the establishment of synaptic transmission in the inner membrane of mitochondria during polarisation. These GO terms were predominantly not enriched in individual modules, but they were involved in the pathways or processes that contribute to the key events. Such a replacement is reasonable yet labour-intensive, the reproducibility of result was compromised, and the existence of parent or child terms sometimes do not guarantee the occurrence of key events. Since a key event should be stringently specific, it is therefore discouraged to do so. Table 13, Table 14, and Table 15, below, depict the hidden potential (#) of each module discovered based on the alternative terms used in the search:

Table 13: Alternative keywords used in the search for Key events.

Key events	Alternative term (parent term /child term / other relevant)
Oxidative DNA damage	DNA damage
Apoptosis	Apoptotic process, apoptotic
Follicular atresia	Ovarian follicle cell, border follicle cell
Mitochondrial hyperpolarisation	synaptic transmission, potassium channel, voltage-gated potassium channel
Oxidative phosphorylation	oxidoreductase activity, alcohol dehydrogenase (NADP+) activity
Mitochondrial ATP production	Mitochondrial ATP synthesis, ATP generation from ADP, ATP biosynthesis, electron transport chain
Lipid peroxidation	Lipid catabolic
Lipid storage	Lipid localization
Oogenesis	Female gamete generation, ovarian follicle cell, gamete generation
Fatty acid oxidation	Fatty acid metabolic process

Table 14: Integration of significant modules with potential key events for the 4 days-transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a 'tick' to represent the presence of the gene. Hashtag (#) indicates that alternative terms have been used in the search for key events.

Key Events	Integration of significant modules										
	Blue	Pink	Red	Turquoise	Light cyan	Yellow	Brown	Green	Dark turquoise	Cyan	Black
Oxidative DNA damage	✓	-	#	-	#	#	✓	#	-	✓	#
Apoptosis	#	-	✓	#	-	#	✓	✓	#	✓	✓
Follicular atresia	-	-	-	-	-	-	-	-	-	-	-
Mitochondrial hyperpolarisation	#	#	#	#	-	#	#	#	-	-	#
Oxidative phosphorylation	#	✓	#	✓	-	#	#	#	#	#	#
Mitochondrial ATP production	#	#	#	✓	-	#	-	-	#	#	#
Lipid peroxidation	#	#	#	#	-	#	#	#	#	#	#
Lipid storage	#	-	#	#	-	✓	#	#	-	#	#
Oogenesis	-	-	✓	-	-	#	#	-	✓	-	-
Fatty acid oxidation	#	✓	#	#	-	#	#	#	✓	#	✓

Table 15: Integration of significant modules with potential key events for the 8 days-transcriptome data. Modules which consist of enriched GO terms in the key events are labelled in green whilst for the non-enriched GO term in the key events, they are given a 'tick' to represent the presence of the gene. Hashtag (#) indicates that alternative terms have been used in the search for key events.

Key Events	Integration of significant modules											
	Black	Blue	Brown	Dark green	Green	Green yellow	Grey 60	Light cyan	Light green	Magenta	Midnight blue	Red
Oxidative DNA damage	✓	✓	✓	-	#	#	#	#	✓	#	-	✓
Apoptosis	✓	✓	#	#	#	✓	#	#	-	✓	#	#
Follicular atresia	-	✓	-	-	✓	-	-	-	-	-	-	-
Mitochondrial hyperpolarisation	#	#	#	-	#	-	-	-	-	-	#	#
Oxidative phosphorylation	✓	#	#	#	#	#	-	-	#	#	#	#
Mitochondrial ATP production	✓	-	#	-	-	-	-	-	-	#	-	-
Lipid peroxidation	#	#	#	-	#	#	-	#	#	#	#	#
Lipid storage	#	#	✓	-	#	#	#	-	-	#	-	#
Oogenesis	✓	✓	-	-	✓	-	-	-	-	-	-	#
Fatty acid oxidation	✓	✓	✓	-	✓	✓	-	-	#	✓	✓	#

4.4 Summary toxicity pathways: Integrating the response of differential metabolites with differential transcriptomics pathways

Despite no statistically significant metabolomics pathways and integrated transcriptomics-metabolomics pathways being generated from Paintomics3, the output from KEGG, the Reactome pathway, and GO enrichment analysis, this study could still provide some useful insights using the expression of metabolites from the whole-body gamma radiation exposure of *D. magna*. A literature search on other model animals exposed to radiation was also performed to provide scientific evidence to elaborate the relationship between differential metabolites and the enriched transcriptomics pathways.

Note that the biological interpretations in the integration of metabolomics and transcriptomics shows that significantly enriched pathways found from the integrated pathway enrichment analysis and functional enrichment analysis (using DEACGs) using studies beyond crustacean, for example, human and mammals. Unfortunately, there is far less research on the impact of ionizing radiation on crustaceans, than there is on humans and mammals. This may give rise to questions of relevance because different species may not share commonalities in response to radiation. Hence, the findings are suggestions and should be used as a reference, more thorough study should be performed in the future to confirm the biological interpretation.

4.4.1 Low dose group

Activation of anaerobic- and aerobic-glycolysis results in energy disturbance

The unusual disturbance in energy homeostasis shows a significant depletion of DEMs such as 3-hydroxybutyric acid (-0.88 fold, $pval < 9.25E-05$), verbascose (-0.5 fold, $pval < 2.17E-07$), glycerol-2-phosphate (-0.17 fold, $pval < 0.003$), which are known as the alternate sources of energy commonly used in the state of ketosis, dehydration, or hypoxic environment. (Jorge and António, 2018, Bartmann et al., 2018, Hasikova et al., 2020). On the other hand, there were large numbers of significantly upregulated genes from the enriched pathway, oxidative phosphorylation. A gradual reduction in glucose levels in low dose irradiated cells suggest that there was a graduation shift from oxidative phosphorylation to glycolysis possibly due to an increased energy demand. The decrease in aspartic acid (-0.35 fold $pval < 0.007$) is an example of glucogenic amino acid indicating that non-carbohydrates products are converted into pyruvate to provide glucose for catabolic reactions (Kreamer et al., 2001). This could jeopardise the TCA cycle and explain the accumulation of pyruvic acid because tumor cells can obtain energy from glycolysis and carry out anaerobic respiration, rather than relying on the TCA cycle (Figure 23). The elevation of pyruvate (Figure 25: TCA cycle) indicates downregulated or impaired mitochondria pyruvate carriers which failed to the transport of pyruvate from cytosol to the mitochondrial matrix for oxidative

phosphorylation (Navas and Carnero, 2021). Thus glutaminolysis is presumably activated as an energy compensation strategy to support TCA when glycolytic carbon is not available (Yang et al., 2014, Navas and Carnero, 2021). Cancer cells usually store large amounts of glutamate and are made readily to convert to glutamine to channel into the TCA cycle for permanent functioning. On top of that, the accumulation of pyruvic acid probably resulted from aerobic glycolysis which is preferred by highly proliferating tumor cells. However, such mechanisms have also been found in immune cells to fuel inflammatory responses (Soto-Herederó et al., 2020). Pro-inflammatory cells such as macrophages and T-cells require large amounts of energy in very short periods of time. In the presence of oxygen, glucose was taken up by glycolysis to produce pyruvate, which was further fermented into NADH and 2 ATPs, instead of turning into lactate. Free energy from NADH was released through re-oxidation in the electron transport chain to create ~30 extra ATP molecules as one of the defense mechanisms in the immune system (Navas and Carnero, 2021). This also explains the up regulation of DEGs in oxidative phosphorylation (Figure 23) to support anaerobic glycolysis. The deposition of pyruvate in the presence of oxygen prevents the conversion into lactate.

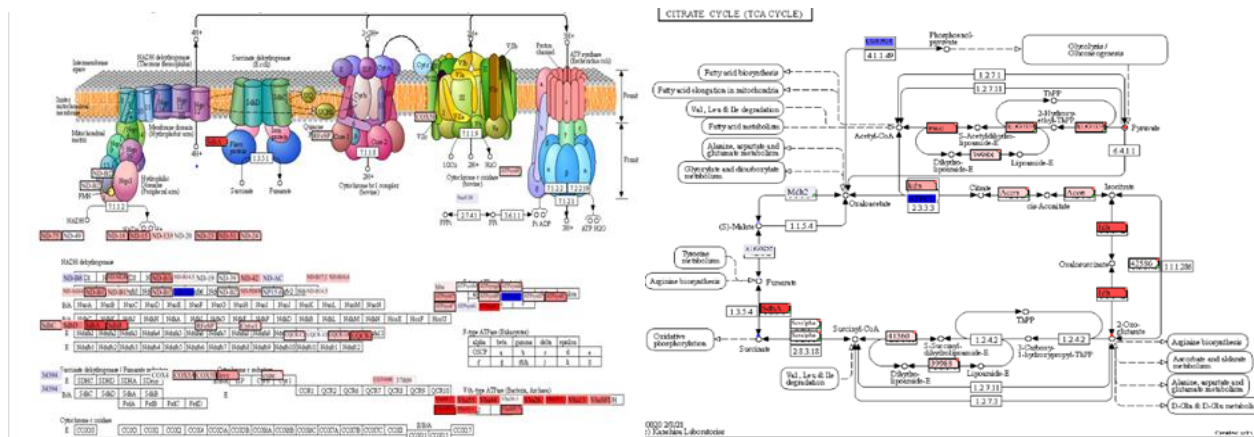


Figure 23: Top enriched pathways from Paintomics3: oxidative phosphorylation(left), TCA cycle(right). Map of the most significantly enriched pathway, blue shades represent down-regulated genes, red shades represent up-regulated genes, and the DEACGs are enclosed in a thicker border.

Disruption of the energy balance and amino acid biosynthesis in one carbon metabolism

The enrichment of the glutathione (GSH) synthesis pathway in carbon metabolism was reported to mediate the repair mechanism and particularly promote the survival of cancerous cells upon exposure to ionizing radiation, despite being an excellent defense against ROS build up and lipid peroxidation (Pujari et al., 2009, Estrela et al., 2006). Glutathione is made of amino acid glutamine, cysteine, and glycine. In this study, the key metabolites for GSH synthesis were differentially expressed. For instance, cysteine (-0.14 fold, pval < 0.001) was depleted, indicating a strong need, but glutamic acid (precursor of endogenous GSH), betaine (oxidative metabolites of choline which also participate in GSH metabolism), cystine (precursor of cysteine), pyroglutamic acid

(known as 2-oxoproline, 0.24 fold, p val < 0.001), methionine (0.11 fold, p val < 0.06), and betaine (0.17 fold, p val < 0.0003) were elevated. This could probably be explained by the findings in mammalian cells which do not synthesise and store cysteine; plenty of methionine and cysteine were reported as required to maintain the normal synthesis of GSH. GSH is the most prevalent low-molecular weight thiol found to counteract ROS mediated production and protect cysteines from over-oxidation (Ulrich and Jakob, 2019). However, insufficient GSH due to a shortage of cysteine triggers ferroptosis. Despite no metabolite trace of GSH being found, the constituents suggest an increase in GSH synthesis is one of the early defenses induced by ionizing radiation (Pujari et al., 2009, Kim et al., 2003, Wang et al., 1997). However, the accumulation of GSH is also another major source of energy to replace glucose in rapidly dividing cells, and could also promote fatty acid beta oxidation (Aledo, 2004, Golla et al., 2017). Irregularity in methionine cycle from the one-carbon metabolism was reported contribute to tumor growth, necrosis and brain damage. In Figure 24, the overall up-regulation of most DEACGs (depletion of glucose and alternative macromolecules, elevation of the precursor of GSH, up-regulated enriched carbon metabolism, fatty acid metabolism, amino sugars, and nucleotide sugar metabolism corresponding to the de novo fatty acids biosynthesis), is the usual sign known for uncontrollable cell proliferation (Shuvalov et al., 2017).

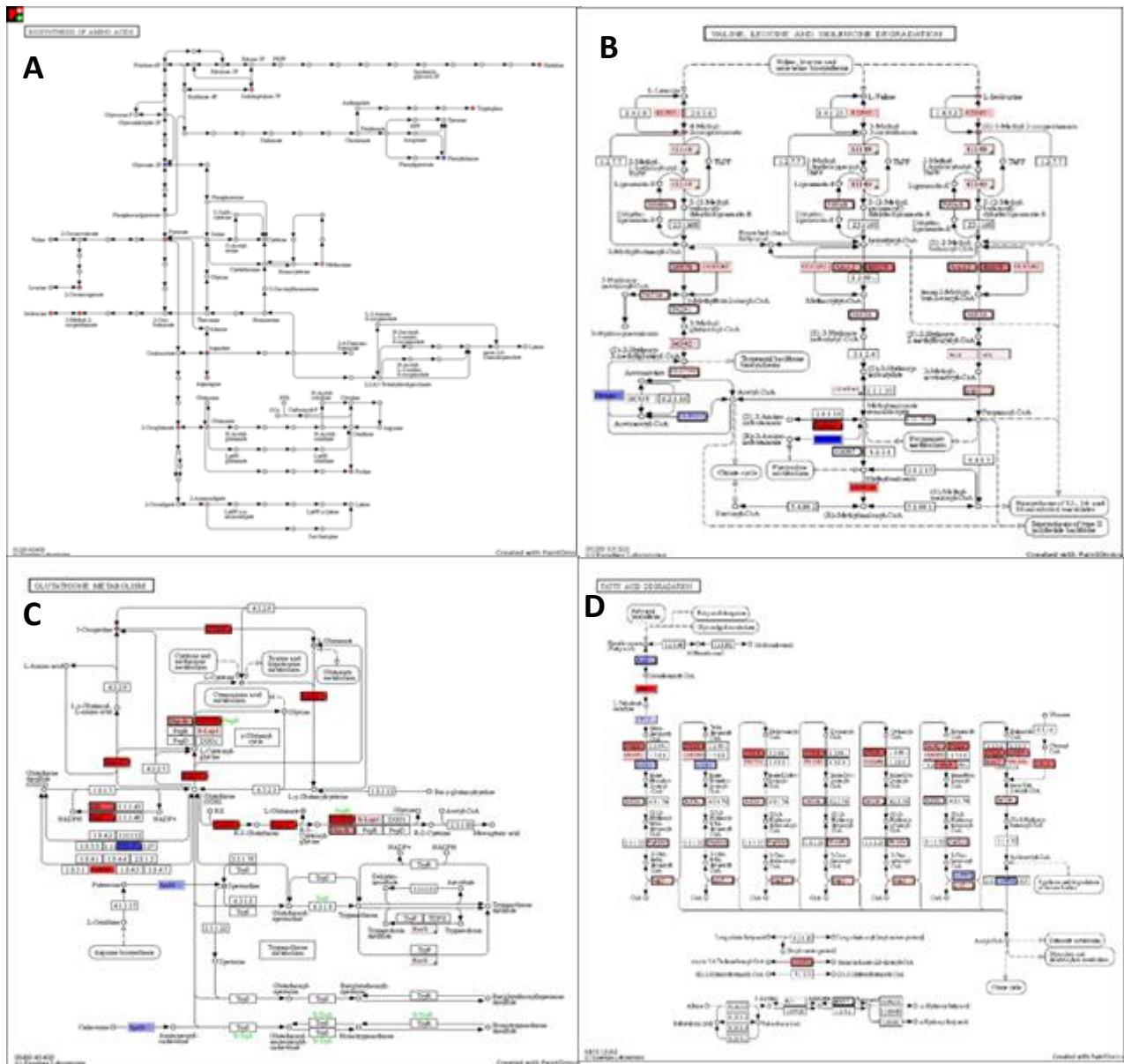


Figure 24: Enriched transcriptomics and metabolomics pathways from Paintomics3 output. **A:** Biosynthesis of amino acid. **B:** Valine, leucine and isoleucine degradation. **C:** glutathione metabolism. **D:** fatty acid biosynthesis.

Resistance and damages

In Figure 25, the activation of the immune system was identified as a TP by changes in some metabolites: the elevation of histidine (0.6 fold, pval < 0.0002), precursor of histamine - an inflammatory agent in immune system; tryptophan (0.2 fold, pval < 0.001), and riboflavin (0.3 fold, pval < 6.39E-05). Their increase is known to suppress the free radical oxygen species. Allantoin (-0.5 fold, pval < 0.007), an antioxidant which is produced through purine metabolism,

was depleted; gamma-Linolenic acid (0.32 fold, pval < 0.005) was accumulated to increase the chance of apoptosis in irradiated cells (Irani et al., 2018, Antal et al., 2015).

On the other hand, the level of deoxy carnitine (0.16 fold, pval < 0.012), the precursor of carnitine which aids the conversion of ATP to ADP in mitochondrial oxidation, was elevated. Essential metabolites such xanthine (0.3 fold, pval < 0.001), and 2-Deoxy-D-ribose 5-phosphate (0.3 fold, pval < 8.32E-05) were elevated, which shows evidence of base deamination from DNA and RNA strands that can contribute to mutagenic lesions. This is likely the result of nucleic acid damage caused by autophagy or sclerosis (Golla et al., 2017, Chaurasia et al., 2016).

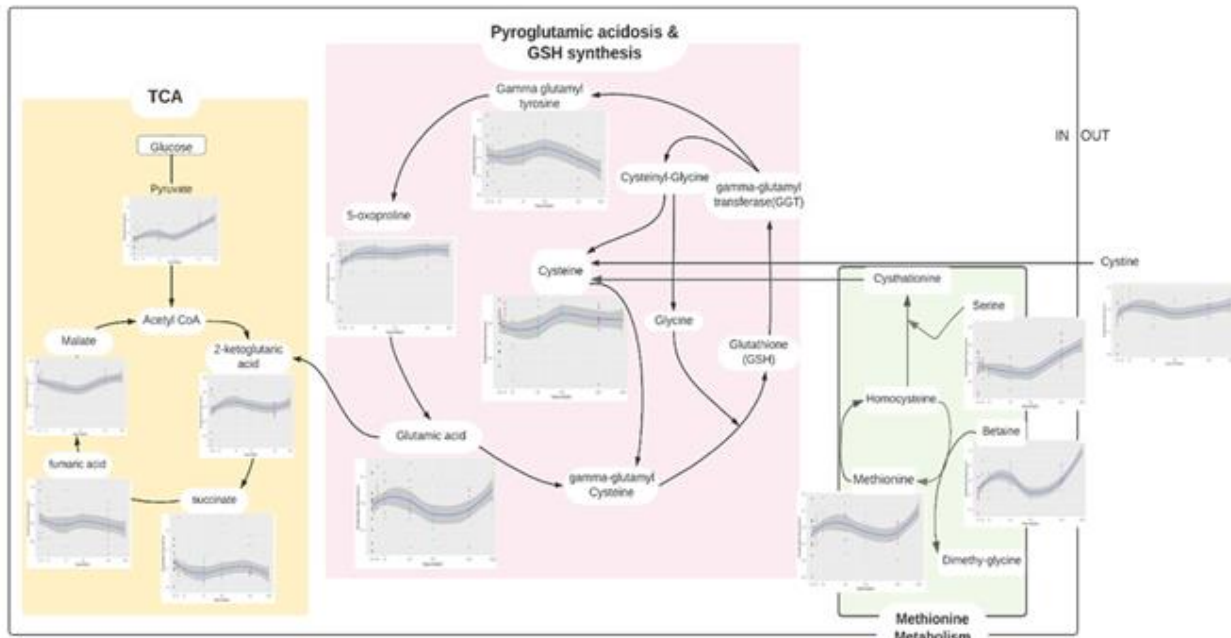


Figure 25: The relationship between Glutathione (GSH) synthesis, Tricarboxylic acid cycle (TCA) and methionine metabolism as the critical alteration in carbon metabolism.

4.4.2 High dose rate responsive group

Metabolic reprogramming revealed an adaptation strategy on de novo synthesis and salvaging of purine and pyrimidine from endogenous mechanisms

Similar findings in gluconeogenesis show a significant depletion in glucose (-0.4 fold, pval < 0.012), fructose (-0.08, p val < 0.003) along with verbascose (-0.5 fold, p < 5.4E-10). This suggests highly active glycolysis for expressed pathways like glycosphingolipid synthesis, and also nucleotide metabolism (Wang et al., 2009). However, the most representative DEMs in the high dose rate responsive group were structural components from nucleic acids. Significant depletions were found in nitrogenous bases sugar cytosine and uracil by at least 0.2fold (pval < 0.015), and in nucleosides such as cytidine, xanthosine by at least 0.07 fold (pval < 0.001) and nucleotides for

example GMP (0.07, p val < 0.0001), UDP (0.4, p val < 0.013) as well as AMP which is also the most depleted metabolite (0.8 fold, p val < 0.01). Proliferating cells in high dose rates require a substantially larger amount of substrate for nucleotide production and energy to support the metabolism than those in the low dose rate group.

Purine and pyrimidine synthesis starting from glycolysis and their salvaging processes are regulated by the phosphatidylinositol 3-kinase signaling (Akt/PI3K) pathway through controlling the amount of phosphoribosylpyrophosphate (PRPP) (Wang et al., 2009). The underlying mechanism includes vigorous glucose uptake and conversion of citrate to acetyl-CoA through ATP citrate lyase for synthesis of fatty acid. Extracellular purine molecules were discovered as the main receptor ligands of oncogenic cells and the purigenic signaling on the surface receptors was reported controlling the speed of metastasis. Depletion of glucose, AMP and GMP seems to correlate with the increase of uridine (0.58 fold, p val < 6.38E-07) and creation of large amounts of ATP. Previous studies have also reported both oxidative and non-oxidative sources are involved in building ribose-5-phosphate for purine synthesis (in abnormal conditions), in agreement with the elevation of ribose-5-phosphate by 0.17 fold (p val < 0.0005) (Wang et al., 2009). ATP plays a crucial role in energy metabolism and acts as a neuro-signal for intercellular communication in CNS (Haydon, 2012). The unusual increase in GABA, by 0.4 fold (p val < 0.007), indicates more neurotransmitters released from the glia cells into the microenvironment made available for scavenging to meet the needs of neuroactivity (Haydon, 2012). On the other hand, pyrimidine metabolites such as thymidine, thymine, uracil (-0.27 fold, p val < 0.013), dCMP, cytosine (-0.21, p val < 0.015), cytidine (-0.07, p val < 0.007), dUMP, and uridine (0.59 fold, p val < 6.38E-07) secreted by non tumour entities (see Figure 25) have also been proposed as oncogenic metabolites (Siddiqui and Ceppi, 2020). The top metabolomic pathway, pyrimidines metabolism for pyrimidine synthesis and scavenging is hence likely to enhance radio-resistance. Since more than 80% of DEACGs in over-represented pathways showed up-regulatory expression, a great amount of ATP generated is likely to supply the underlying abnormal cellular activity like the cellular communication and metastasis.

Another expressed pathway, TGF- β signaling, is known for its dual role for inhibiting cell growth in early carcinogenesis and being pro-oncogenic in tumor development and facilitating cell invasion in the late stage. Secretion of TGF- β stored in the extracellular matrix (ECM) is activated once the straining from injured fibrotic cells is detected (Hinz, 2015). Downstream metabolomics expression was therefore associated with the gene expression of the TGF- β pathway and extracellular matrix (ECM) receptor interaction. Activated TGF- β requires specific ECM escort proteins to bind to the receptor protein on the cell surface to promote cell signaling (Todorovic and Rifkin, 2012). The differential expression of both pathways likely indicates that the rapid proliferation of tumor cells creates vascular and inflammatory stress which leads to remodelling and un-terminated ECM signals on healthy cells. Profibrotic ECM in the extracellular environment could also mislead the nearby healthy cells into fibrogenesis (Shimbori et al., 2013, G. Gritsenko et al., 2012).

Under irradiated situations, signaling from TGF- β received by the receptor tyrosine kinase (RTK) will activate the PI3K/AKT signaling pathway (Zhang et al., 2013). PI3K signaling is involved in the regulation of glucose metabolism, cell growth, proliferation, protein translation and mediating

cell cycles in G1 and the transition of G1 to S phase (DNA synthesis) (Wang et al., 2009). Activation of PI3K/AKT prevents the TGF-beta induced pro apoptotic response. The interplay of the two pathways enhances cancer survival and proliferation. On top of that, in both the linear and high dose rate responsive groups, another expressed pathway *forkhead* transcription factor pathway (FOXO) is reported to produce FOXO substrate to facilitate TGF-beta induced growth inhibition under normal circumstances. However, phosphorylation of FOXO from PI3K/AKT truncated the localization on the nuclear and contributes to the uncontrollable cell growth (Zhang et al., 2013). Furthermore, activated PI3K/AKT regulates the level of PRPP to synthesize and salvage purine and pyrimidine in the intercellular and extracellular environment (Wang et al., 2009, Fumagalli et al., 2017). All metabolites involved in the metabolism of purine and pyrimidine as shown in the Figure 26 were differentially regulated as the dose rates went from 1 to 100 mGy/h. Therefore, the activity of the PI3K/AKT pathway was correlated to the high signaling activity between the stimuli of the extracellular environment into the intracellular signaling network, which is coherent with the findings of GO analysis for the high dose rate group (Zhang et al., 2013).

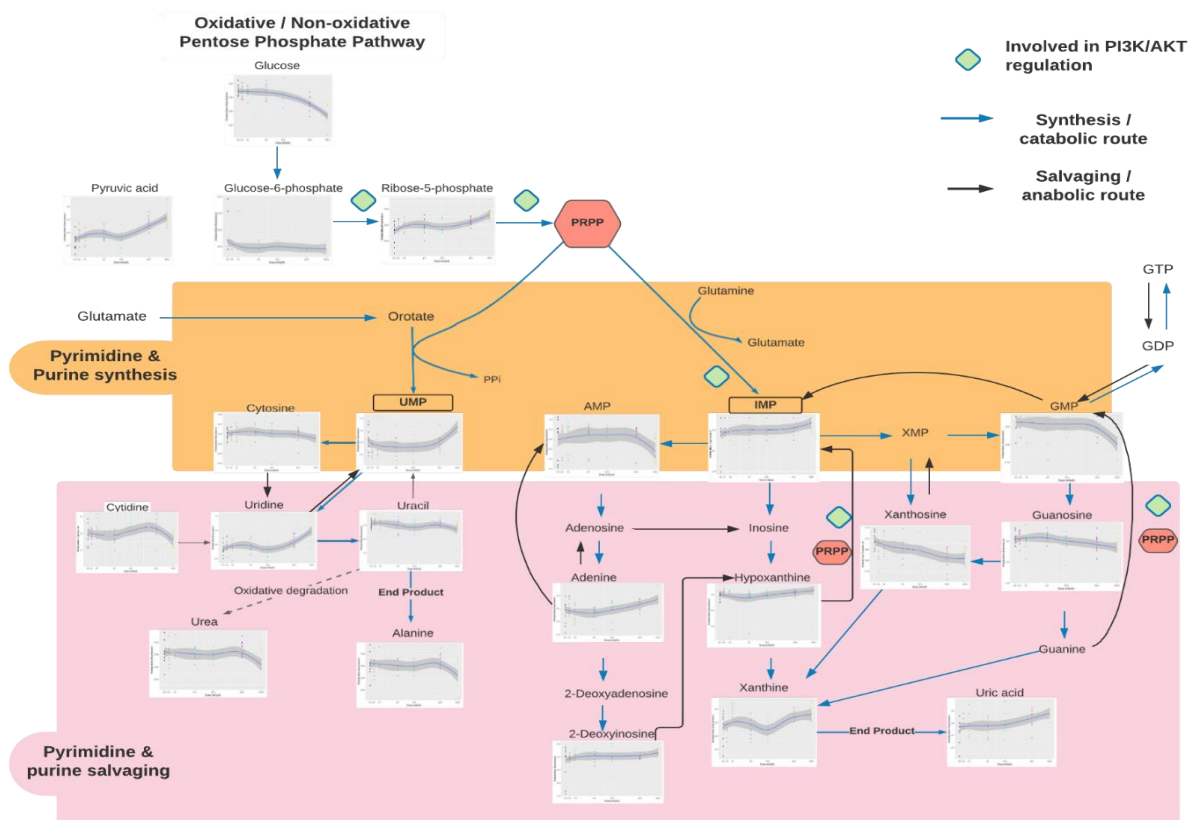


Figure 26: Schematic of purine and pyrimidine metabolism regulated by PI3K/Akt signaling. The level of phosphoribosylpyrophosphate (PRPP) is the key substrate regulated by PI3K/Akt.

Other metabolites, which pointed towards the morphological impacts include depletion of level of myo-Inositol (-0.2, pval <2.18E-0.6), have been associated with chronic hepatic disease and hypoxic encephalopathy; and the accumulation of the derivatives, myo-Inositol-1-phosphate which was reported found in the testes and brains of mammals, is involved in phospholipid

biosynthesis and the controlling of the metabolic flux of inositol (Lackey et al., 2003, Pittner and Hoffmann-Ostenhof, 1976, Chhetri, 2019).

Finally, the low dose rate group had 13 types of lysophospholipids (LysoPC, LysoPE, LysoPA...) that were significantly depleted after the radiation exposure, whilst 20 types were found in the high dose rate group. Lysophospholipids, reported as the by-product of radiation-induced lipid peroxidation in the process of ferroptosis (iron dependent cell death), were however decreased in this study indicating that resistance against ferroptosis was established through the elevation of oleic acid (0.12 fold, p val < 0.001) and GSH (Ubellacker et al., 2020). In fact, lysophospholipids contributed to cell growth and oleic acid secreted by the lymph was reported to protect against the metastasizing of cancerous cells. Lysophospholipids were excluded in pathway analysis due to no documentation being found in the KEGG compound database (Ye et al., 2020).

4.4.3 Linear model group

Fitting a linear model to the metabolic data resulted in differential metabolites (DMs) with very small log fold changes (-0.006 to 0.006) due to non-monotonic expression of the DMs. The top expressed transcriptomics pathways in the linear model group were very similar to the top pathways from the high-dose responsive group but with RNA metabolism (GO analysis), ribosome biogenesis (Reactome PA) and ribosome biogenesis in eukaryotes (KEGG) topping the overall pathways from different databases. Hence, the linear model provides a different perspective describing that the metabolism of RNA was conducted in a linear trend.

miRNA promotes metastasis of the malignant cells with EMT

miRNA is known for regulating the gene expression (mRNA) through mRNA cleavage and deadenylation (Eulalio et al., 2009). Recently, miRNA was reported to act as a positive feedback regulator for TGF-beta/SMAD, by hindering the negative regulator protein, SMAD and PTEN (Zhang et al., 2013). Consequently, PI3K/AKT signaling was hyperactivated, leading to the metastatic transition of epithelial-to-mesenchymal (EMT). EMT occurred during embryonic development and tissue regeneration. Cancer EMT expressed in the form of hybrid cells (a mix between epithelial and mesenchymal) allow it to migrate collectively and invade more aggressively than other cancer cells (Jolly et al., 2015). As tumor tissues established enhanced stiffness through extracellular matrix remodelling (significantly over-represented pathway in this study), miRNA was also reported to be upregulated to control cell matrix interaction for the invasion of cancerous cells. For example, TGF- β -induced miRNA increased the integrin signaling and fortified the ECM for malignant mammary epithelial tissues (Taylor et al., 2013).

Impaired ribosomal biosynthesis is a defense mechanism against the proliferation of oncogenic cells

Apart from DNA repair and replication in the high dose rate group, the massive depletion of nucleotides metabolites in the linear model group can be linked to the massive production of ribosomes. Ribosomes are essential in carrying the free amino acids for protein translation. The

pressure on nucleotide deficiency was greater in oncogenic cells leading to a higher demand for ribosomal biogenesis in tumorigenesis. Besides, ribosomal biogenesis was also reported to play a crucial role controlling the progression of cell cycle, cell division, cell growth and triggering of the EMT (Derenzini et al., 2017, Prakash et al., 2019). Recent studies have pointed out that impaired ribosome biogenesis acts as the first barrier against chromosomal DNA instability by triggering the activation of p53-mediated cell cycle checkpoints (Pelletier et al., 2020). Induced impaired ribosomal biogenesis can be achieved through lowering the level of guanine, and the effects can be enhanced with the gradual inhibition of inosine-5'-monophosphate dehydrogenase (IMPH) (Pelletier et al., 2020). IMPH can speed up the synthesis of GMP for nucleotide deficiency. The oxidation of IMPH produces xanthosine monophosphate (XMP) which can be converted to GMP (Huang et al., 2018). In the linear group, inosine-5-monophosphate (IMP) was slightly elevated and exclusive to the group. The increase of IMP indicates the reduction of IMPH produced and suggests an ongoing IMPH inhibition in the high dose rates responsive group. Hence, the significant depletion in xanthosine observed could lower the amount of XMP and amination of XMP to GMP (Figure 26). As such, the differential depletion of GMP and guanosine depicts a low amount of GTP available and exaggerates replicative stress in the nucleotide pool of oncogenes.

4.5 Contributions and future prospect: Integrating multi-omics revealed an altered nature, prioritizing survivorship over reproduction

The dose rates used in this study were set according to a previous study which examined the toxicity pathways of gamma radiation leading to a reduction in fecundity (Song et al., 2020). Dose rate as low as 1 mGy/h was known as the benchmark response to cause reproduction delay but a high dose rate of 100 mGy/h sped up the cycle of reproduction with a compromised brood size. An intermediate dose rate was reported to not contain significant changes from the output of 1 mGy/h.

4.5.1 The energy consumption of DNA repair enzymes and regaining homeostasis

Differential metabolites from the low-dose group showed a significant expression when the dose rates proceeded from 0 to 1 mGy/h. High energy demands in this group suggest that the cellular system prioritised DNA lesion repair over reproduction for survival, since the delay in production is not associated with a decreased number of progenies. Combined with the output from functional enrichment analysis (Figure 15A), production of antioxidant enzymes, and the disruption of endogenous thiols (e.g., deficiency in GSH) slowed down the rate of cell division in normal cells also contributing to the delay. This study proposes that the detention in the cell cycle

for phases such as G1, and S for DNA synthesis is due to inefficient energy distribution, a lack of nutrients and building blocks (Blakely et al., 1989). Future work focusing on examining the cell cycle analysis can provide novel insights on the genotoxicity of gamma radiation on *D. magna*.

4.5.2 Perturbances from cell cycle arrest and the maintenance of genomic stability

Surprisingly, the intermediate dose rate from 1 to 10 mGy/h however found obvious opposite trends to the low-dose group (see metabolites expression in Figure 25) and such intermediate changes were not observed from the differential metabolites of the high-dose group either. The observed changes revealed some sort of adaptation could probably be built up at the intermediate dose which required different levels of metabolites compared to the low dose group. Since there was no significant difference in the reproduction output from 1 mGy/h, after combining the expression profile of all module eigengenes and differential metabolites from the low dose rate group, this study proposes that the delay of reproduction at this stage was similar to the low dose group. This is most likely because the detention was due to investigation from various checkpoints within the cell cycle and the process of repair taking longer than in the lower dose group (Blakely et al., 1989). Surviving cell cycles at this stage are likely to be more sensitive to ROS damage in DNA and cause the same sort of delay to repopulate the germinal cells (De Felice et al., 2019). Future integrated analysis with metabolomics should probably include an intermediate dose rate of 1 to 40 mGy/h as a separate dose responsive group of interest.

4.5.3 Accelerated cellular metabolism

As the dose rate reaches 100 mGy/h, the differential metabolites in the top differential pathways demonstrate more drastic changes than 40 to 100 mGy/h while the eigengene expressions showed more consistent changes from 1 to 100 mGy/h. Most of the metabolites which were differentially expressed in this group were not differentially expressed in the low dose group. The GO analysis on cellular components (See Figure 8: Cellular Components) for 8 days *D. magna* found GO terms such as female and male pronuclei, along with neuron and morphological development being enriched in the pathway enrichment analysis indicating the alteration of reproduction strategy from parthenogenesis into sexual reproduction when the environmental condition is not favorable (Hiruta and Tochinai, 2012, Hiruta and Tochinai, 2014). Male pronuclei is required to activate the female pronuclei from the resting stage. Mammal cells were highly radiosensitive at M phase in the cell cycle and were reported to be the most susceptible to radiation induced cell death (Onozato et al., 2017, Hafer et al., 2010). The reduction in brood size is likely caused by an induced permanent degeneration in nurse cells, either due to damage that is beyond repair, or a lack of raw entity supply due to alterations of the biological priority in survival. Notably, the significantly enriched Hippo signaling pathway, which controls the size of organs through apoptosis especially in stem and progenitor cells, is an evolutionary conserved signaling pathway found in *Drosophila* and mammals. While the inhibition of key genes (hpo and wts) were found to encourage tissue overgrowth and prevent apoptosis, these genes were up-regulated in this study and probably explain the reduction in brood size (Snigdha et al., 2019).

Besides, other studies also discovered that dysregulated hippo signaling contributes to unsuccessful mitotic exit and cell cycle defects, future study could focus on evaluating the impact of defective Hippo signaling on the follicle cells (de Sousa et al., 2018).

Meanwhile, the acceleration in the reproductive cycle proposed is due to abnormal cell cycle progression. In the study on human and other mammalian tissues, rapidly dividing cells regulated by the cell cycle-dependent radioprotection capacity, established from the early radiation exposure, were reported to be very resistant to radiation induced apoptosis (Hafer et al., 2010). Previous studies in humans and yeast have concluded that the apoptotic resistance in highly dividing cells are stronger than slowly dividing cells whilst remaining protectable by antioxidants. Such protection was reported absent in cells from the stationary phase (Hafer et al., 2010). In agreement with the findings from this group, excess demands on sugar and amino acids can be presumed to fulfil the urge of cell cycle progression as a defence mechanism against apoptosis. As the response to stress stimuli and activation of DNA damage repair are modulated by the cell cycle, the survival strategy has thus prioritised cell proliferation over repairing genome abbreviations. The accelerated reproductive cycle is one of the observed phenotypic changes. While faster cell cycles can undoubtedly facilitate DNA damage repair, the study target or the progeny could possibly suffer from compromised health and shorter lifespans.

The progression of cell cycle is highly relevant to the expression of DEGs, transcriptional regulation and binding of activators/ co-activators to TFs at motif sites and the level of differential metabolites. Future study should include monitoring of the cell cycle progression with AOP to explain the adverse effects of radiation.

Chapter 5 Summary

The workflow in this chapter provides an efficient pipeline which successfully captures the temporal and dose dependent expression of genes and their association with the level of metabolites. The first part of this workflow talked about the detection of co-expression modules which are biologically meaningful and how they were formed into a regulatory network, based on the transcriptional regulation between modules. The aim was to determine if the transcriptional regulation from the 4 days of gamma radiation exposure can be used to predict the adverse outcomes of the 8 days of radiation exposure. Every module was represented by a node and the edges were formed when transcription factors (TF) and their motif sites were found within or between modules. The selection of potential modules was based on non-random association (overlap) between the modules from WGCNA and DEGs from DEseq2 through Fisher's Exact test. The functionality of significant modules was verified with functional enrichment analysis and the functional annotations were also mapped into the proposed Adverse Outcome Pathways (AOP) of *D. magna*. The analysis also proved that the impact of gamma radiation from a shorter period of exposure cannot be used to predict the adverse outcomes at 8 days of radiation exposure due to dissimilarity in the module expression profiles and transcriptional regulation. There is a weakness due to the availability of motif databases which causes the exclusion of significant modules from the transcriptional regulatory network. Improvements in terms of computational algorithms has been discussed and relevant solutions in terms of integration with metabolomics has also been demonstrated in the second part of this workflow.

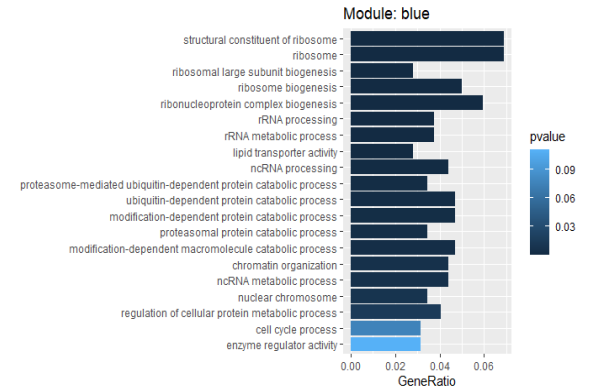
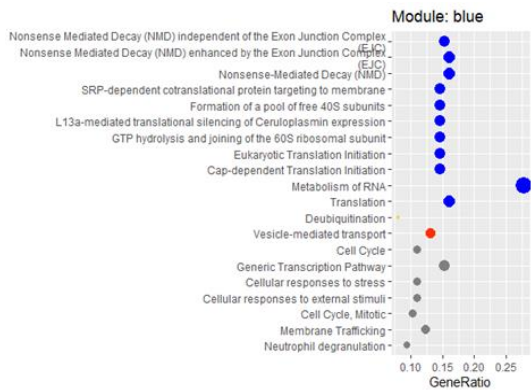
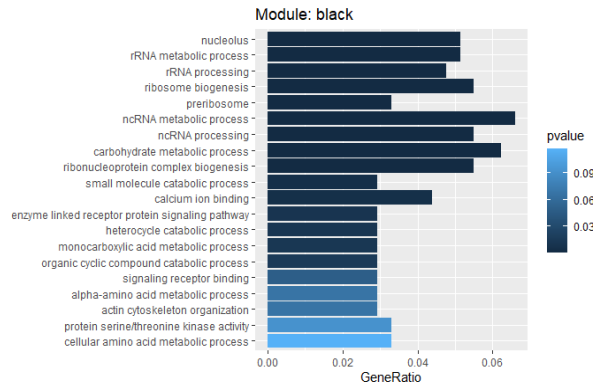
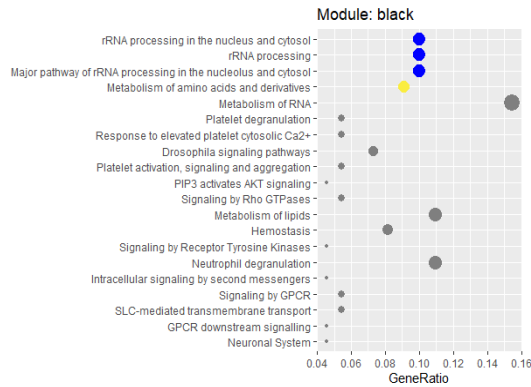
The second part took advantage of significant modules generated from the previous chapter with an attempt to solve the problems and bias introduced by the exclusion of significant modules. The aim was to explore the differences of underlying mechanisms in inefficient reproduction due to varying chronic dose rates shown in the preliminary findings of a previous study. A supervised learning model was used to analyse the 195 metabolites from 70 samples. Using the same design model from the transcriptomics study: linear model and linear combinations (contrast), the design model groups for metabolomics were low dose rate responsive, high dose rate responsive, and linear model. Significant modules from the 8 days transcriptomics data were overlapped with DEGs that were modelled from the same design, significant modules from the same group were combined and the genes were termed differentially expressed and co-expressed genes (DEACGs). This combination took all the significant modules into consideration for the integrated analysis with metabolomics data. However, the comprehensiveness of the output was compromised due to the loss of data in mapping the identifiers from *D. magna* to *D. melanogaster*. Other issues such as insufficient metabolites identified has also lowered the power of the metabolomics data in the Fisher probability combined test, having contained no enriched metabolomics pathways. Therefore, some manual curation and literature research on the relationship of enriched pathways and key metabolites was performed. This workflow also performed other functionality enrichment analysis with different databases to provide a different perspective using DEACGs.

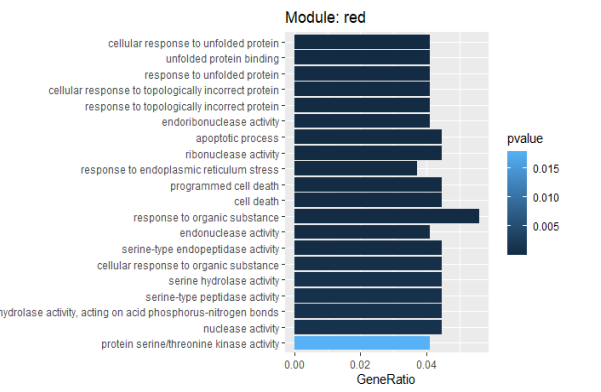
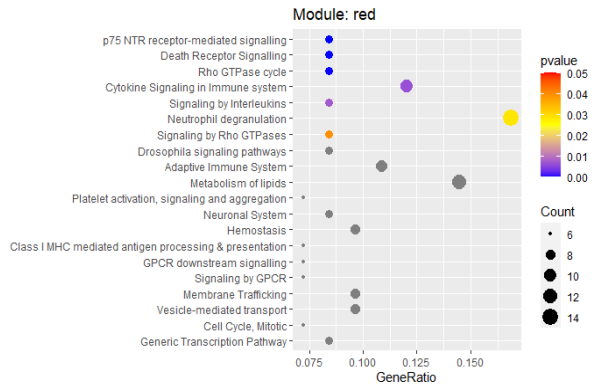
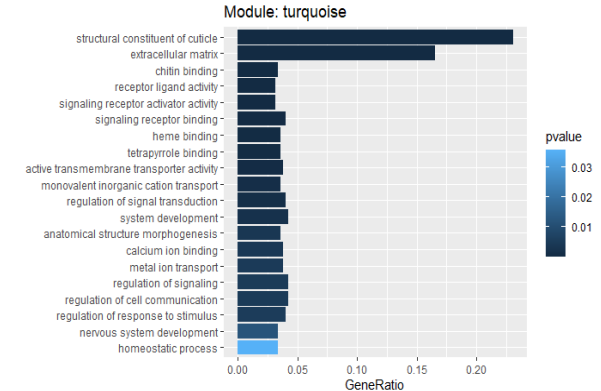
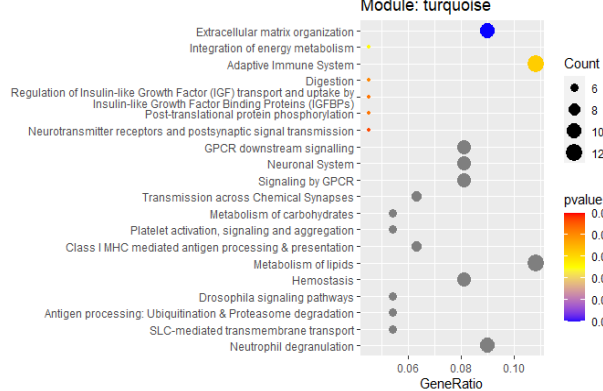
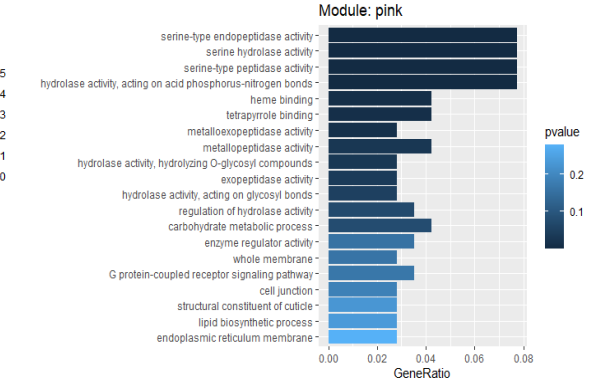
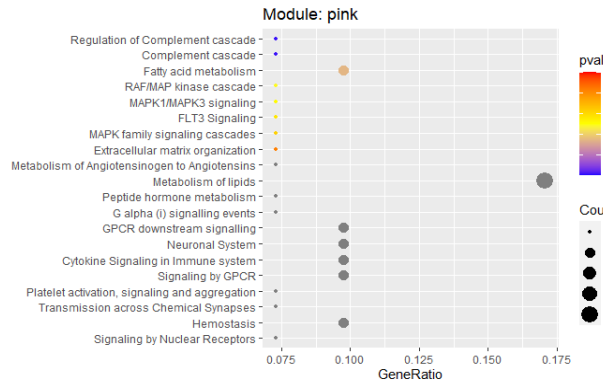
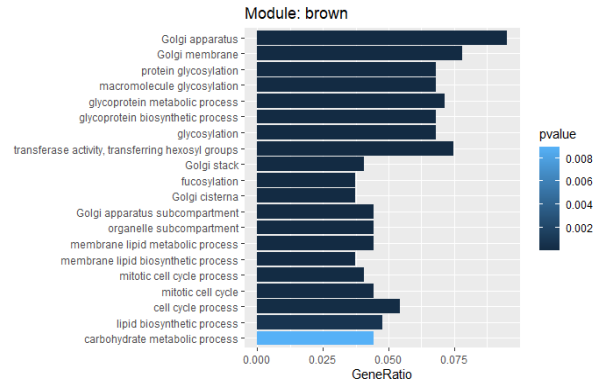
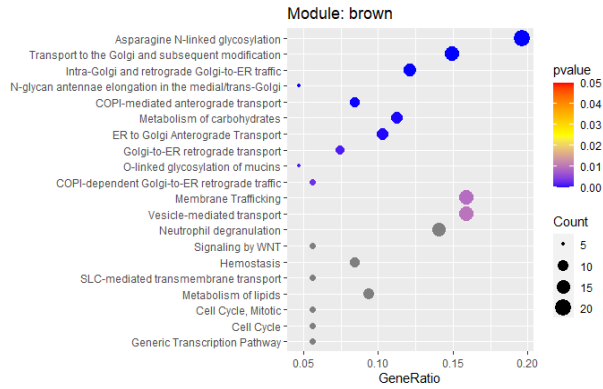
The final output matched the previous reported findings and provided a new insight explaining the underlying mechanism for the observed reproductive differences induced by different gamma dose rates which has not been described in previous studies. Combining the information from metabolome and transcriptome data, new insights suggest that the alteration to the cell cycle contributes to the varying reduction of fecundity under the effect of different dose rates of radiation.

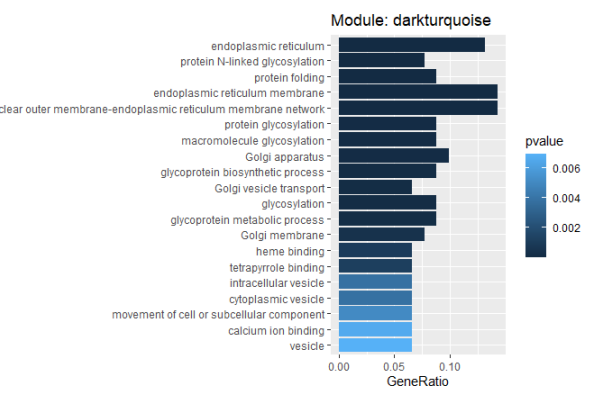
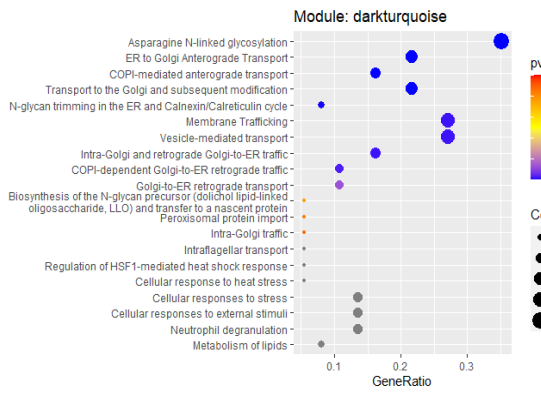
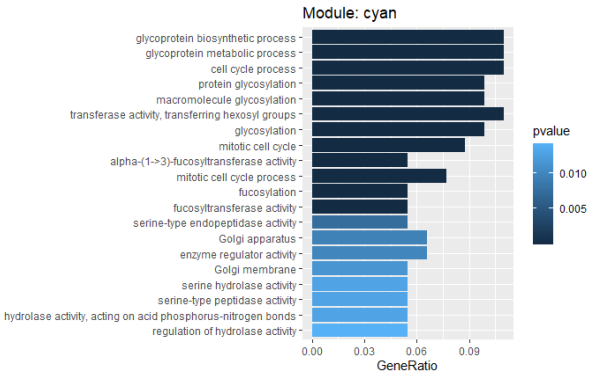
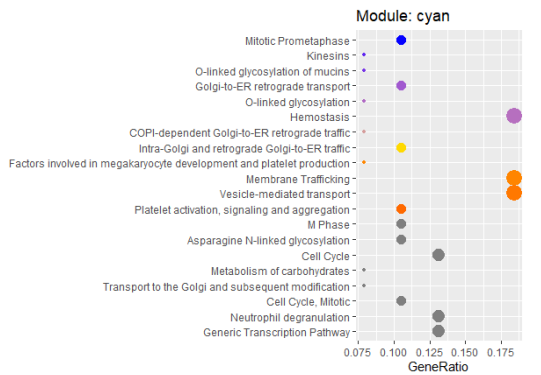
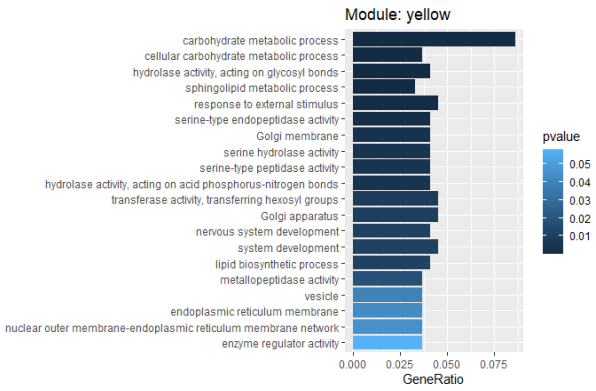
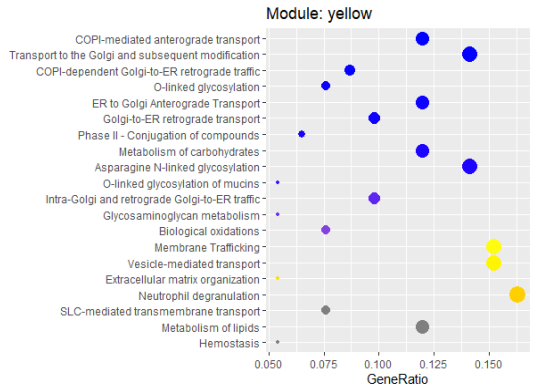
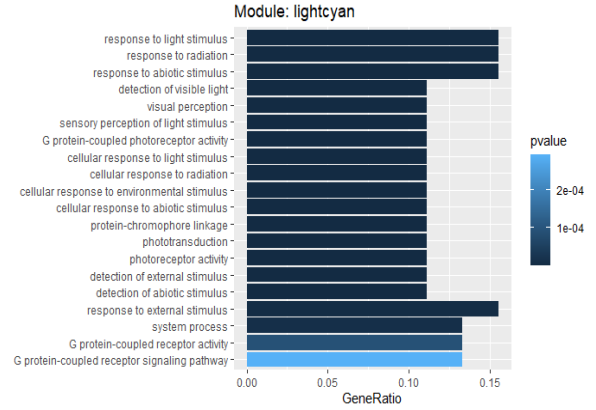
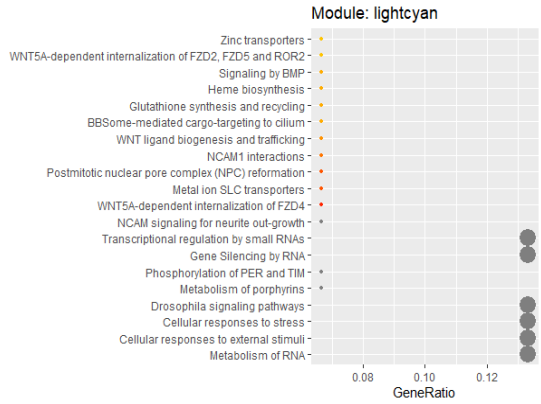
Chapter 6 Supplementary Data

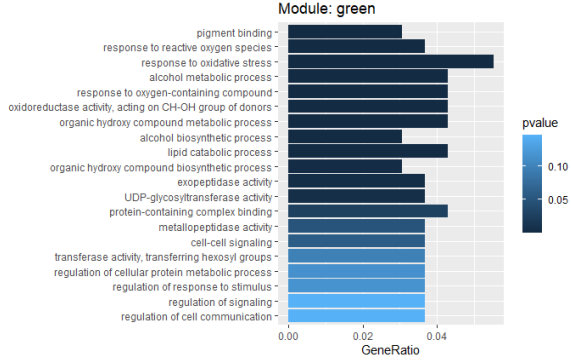
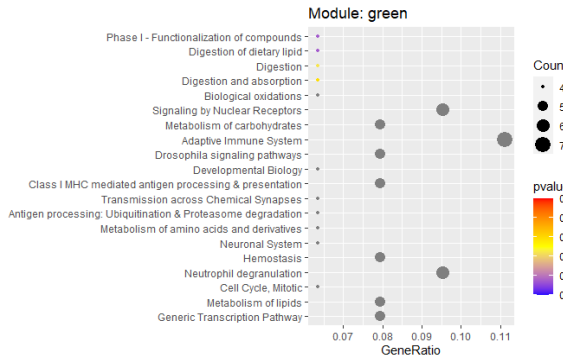
S1 Functional Enrichment – modules from 4 days data

The top 20 enriched pathway (left) and GO terms (right) are arranged according to the order of significance on the y-axis. Gene ratio is the number of genes related to the pathways or GO term divided by the total number of pathways in the module.

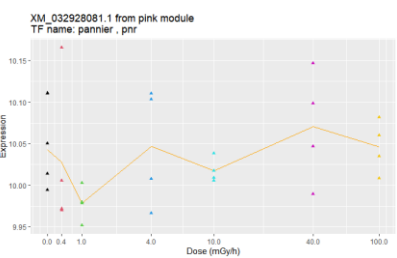
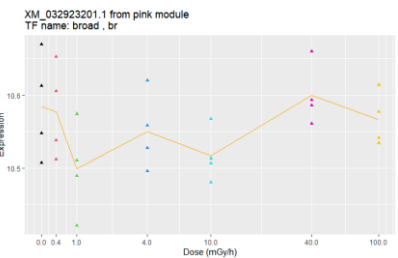
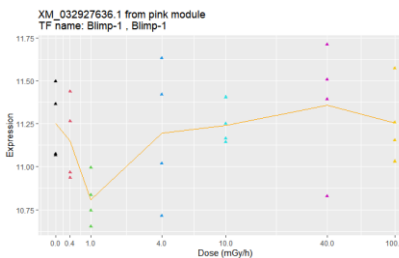
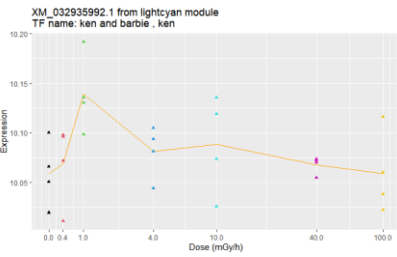
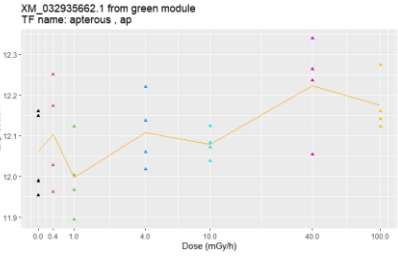
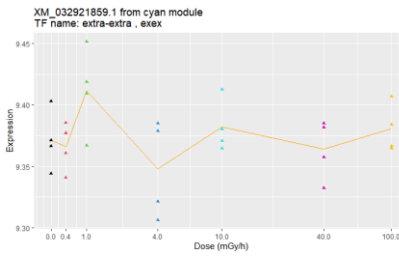
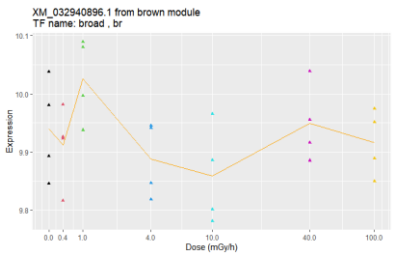
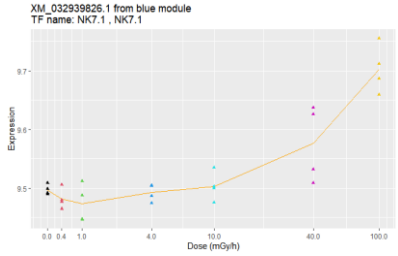
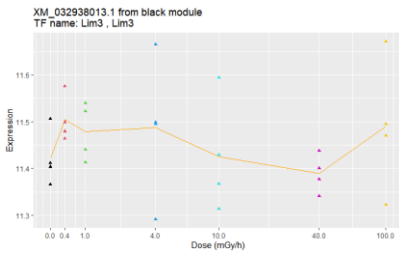


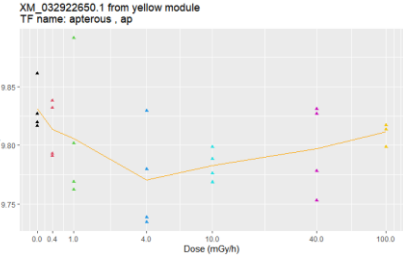
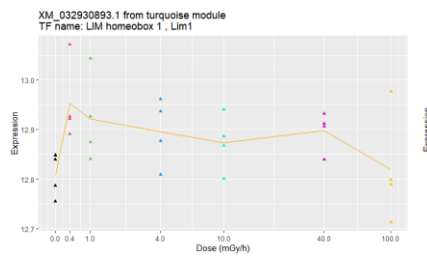
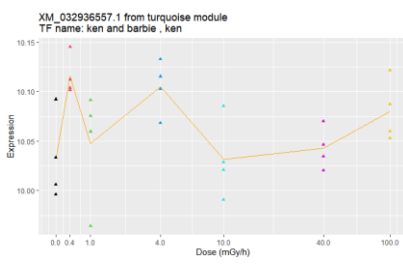
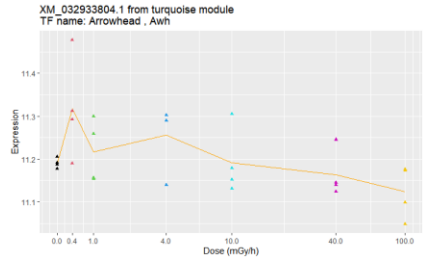
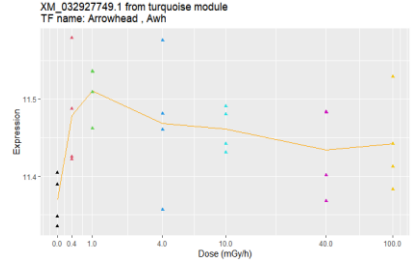
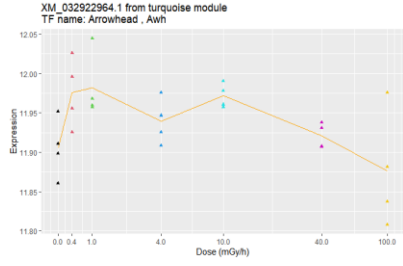
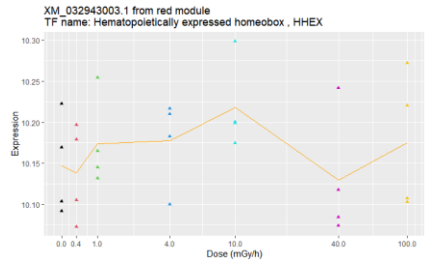




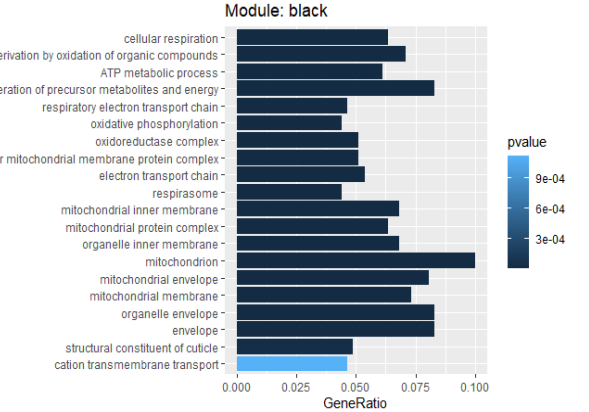
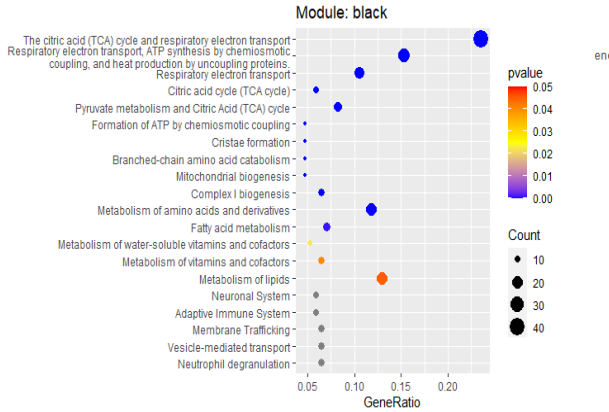
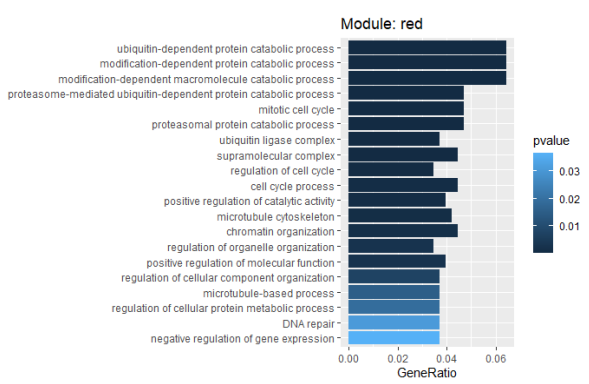
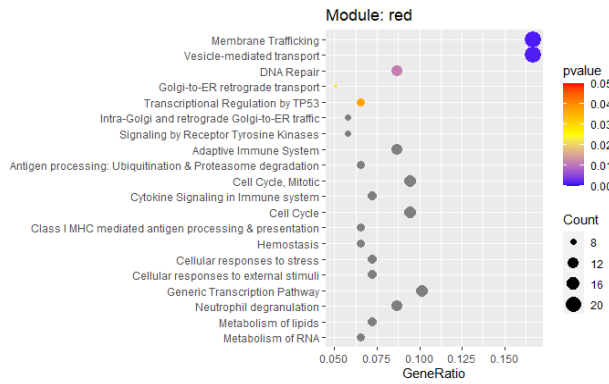
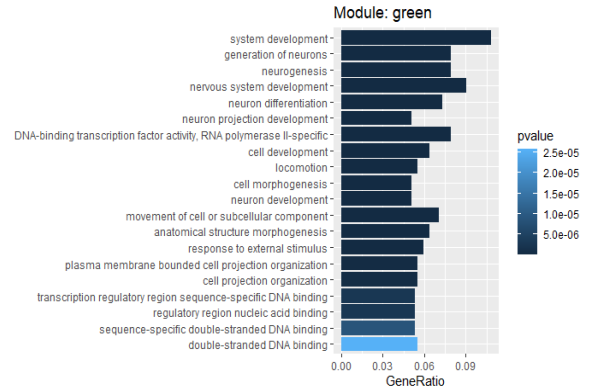
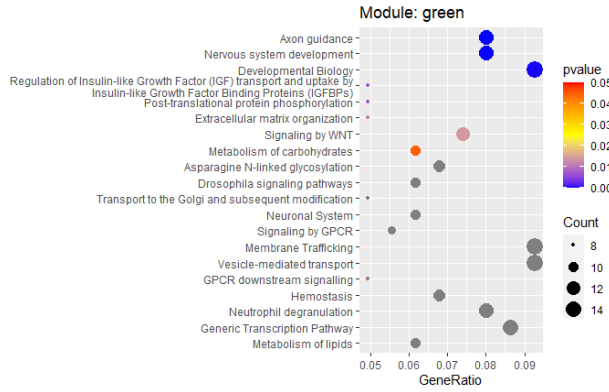


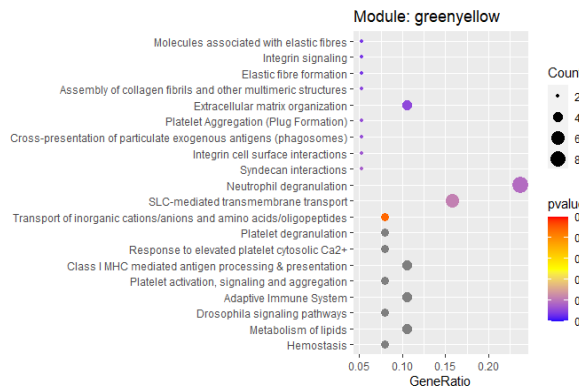
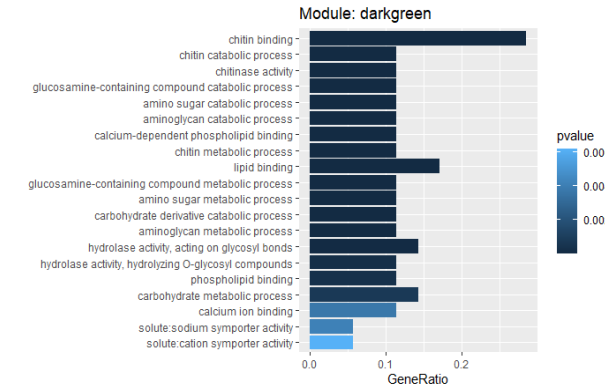
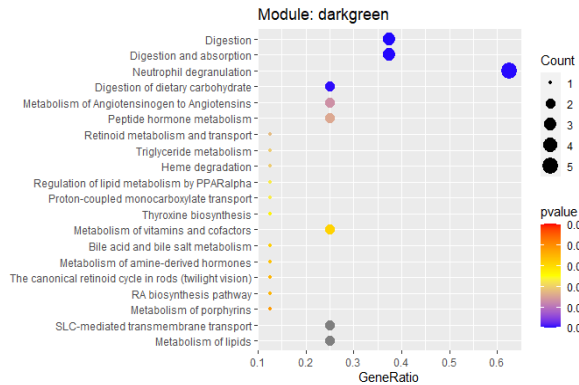
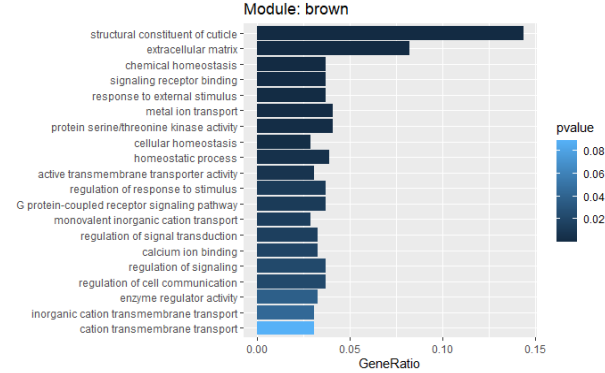
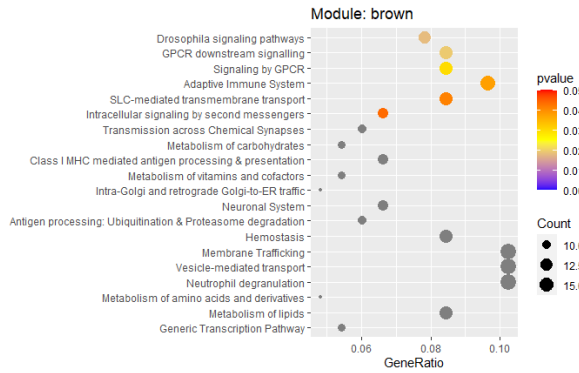
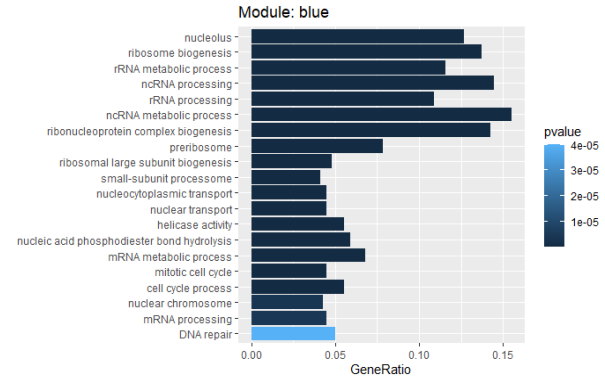
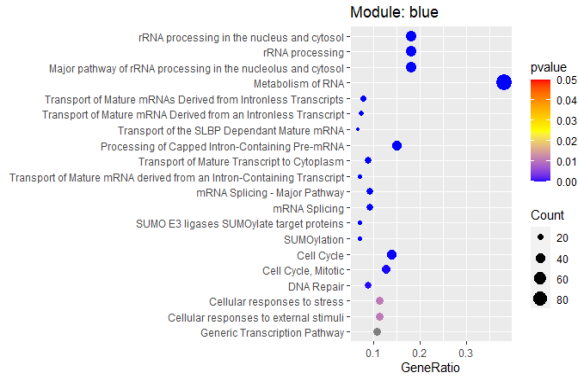
S2 TF gene expression profiles – 4days

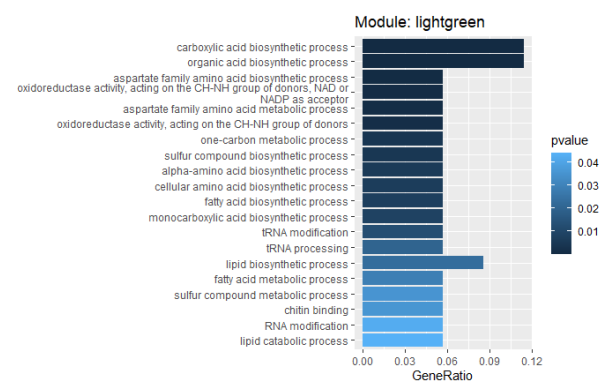
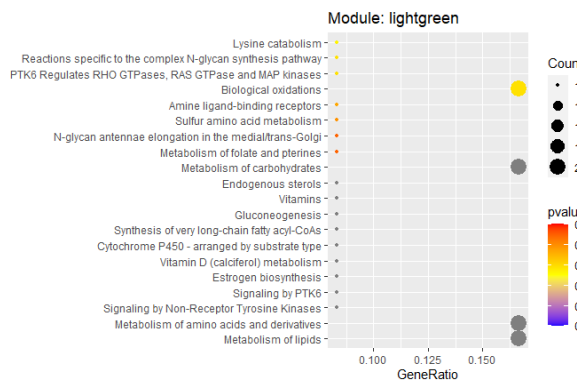
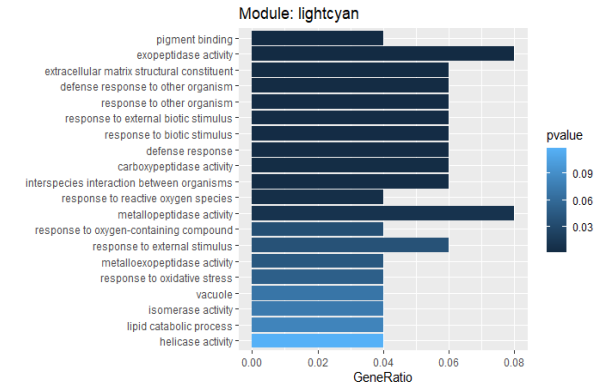
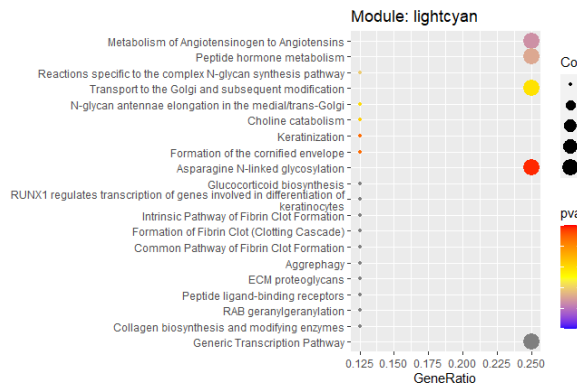
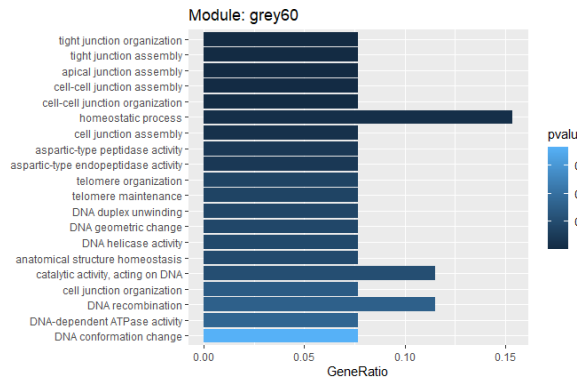
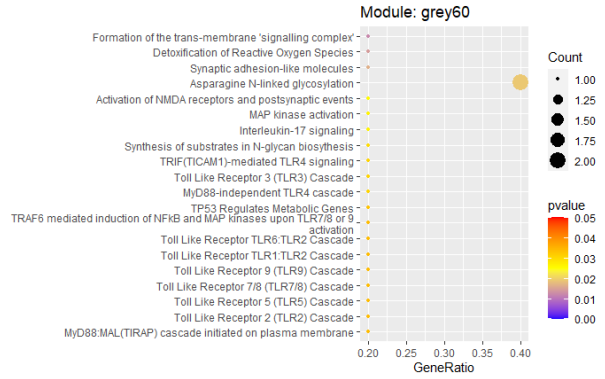
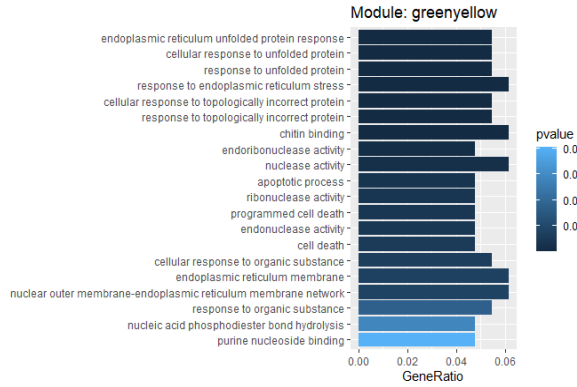


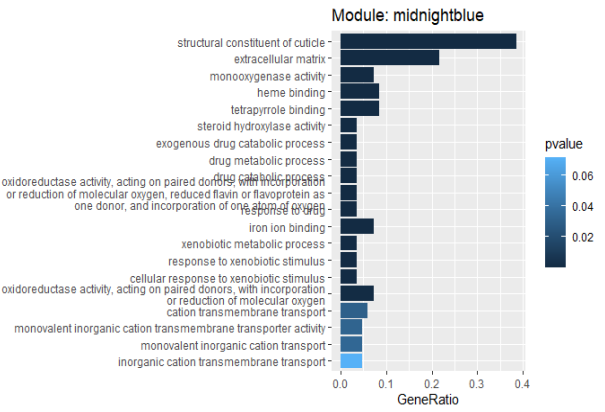
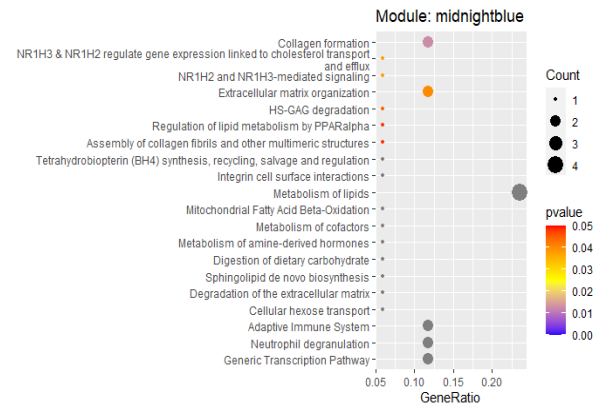
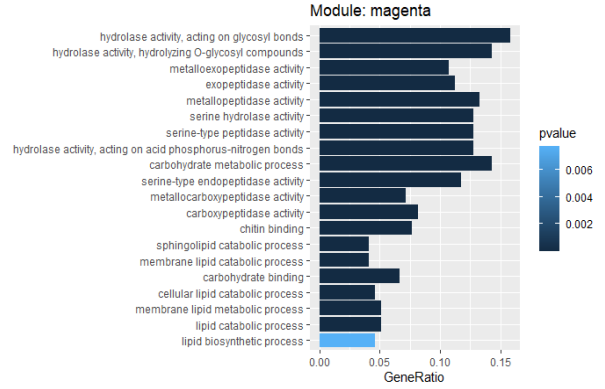
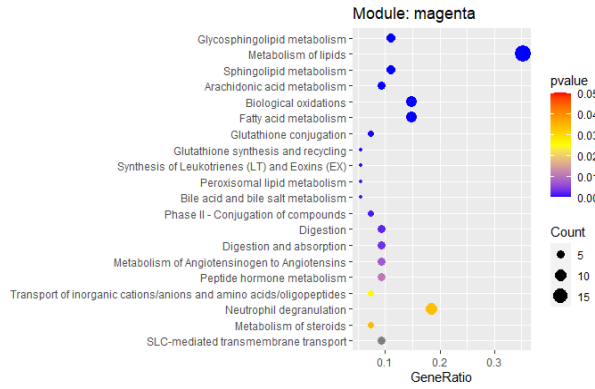


S3 Functional Enrichment - modules from 8 days data

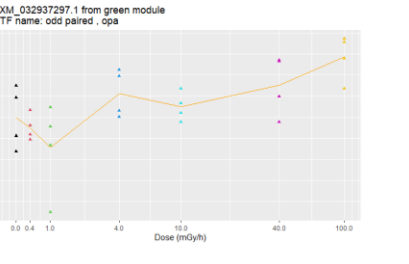
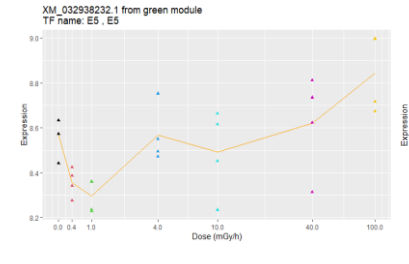
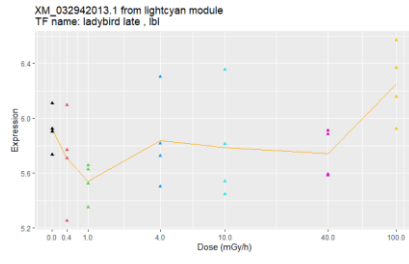
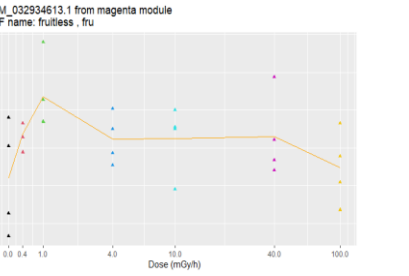
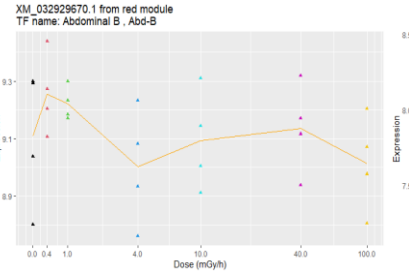
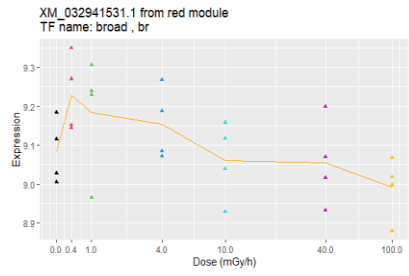




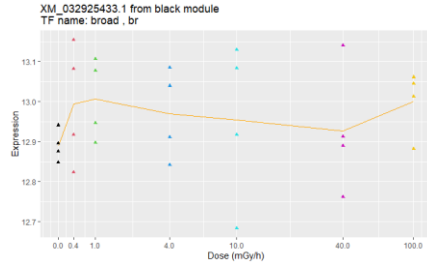
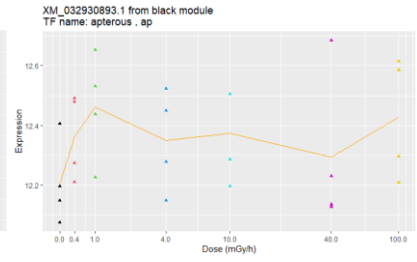
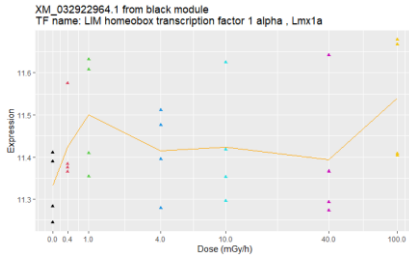
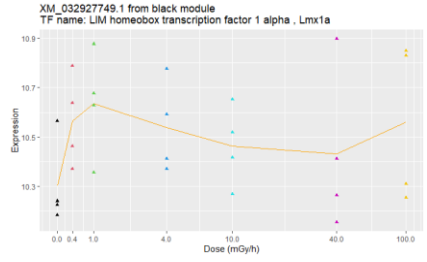




S4 TF expression profiles – 8days







References

- AHMED, S. M. H., MALDERA, J. A., KRUNIC, D., PAIVA-SILVA, G. O., PÉNALVA, C., TELEMAN, A. A. & EDGAR, B. A. 2020. Fitness trade-offs incurred by ovary-to-gut steroid signalling in *Drosophila*. *Nature*, 584, 415-419.
- ALEDO, J. C. 2004. Glutamine breakdown in rapidly dividing cells: waste or investment? *Bioessays*, 26, 778-785.
- ALLBEE, A. W., RINCON-LIMAS, D. E. & BITEAU, B. 2018. Lmx1a is required for the development of the ovarian stem cell niche in *Drosophila*. *Development*, 145, dev163394.
- ALPEN, E. L. 1998. Stochastic Effects — Genetic Effects of Ionizing Radiation. *Radiation biophysics*. Academic press.
- ANTAL, O., PÉTER, M., HACKLER JR, L., MÁN, I., SZEBENI, G., AYAYDIN, F., HIDEGHÉTY, K., VIGH, L., KITAJKA, K. & BALOGH, G. 2015. Lipidomic analysis reveals a radiosensitizing role of gamma-linolenic acid in glioma cells. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1851, 1271-1282.
- BARTMANN, C., RAMAN, S. R. J., FLÖTER, J., SCHULZE, A., BAHLKE, K., WILLINGSTORFER, J., STRUNZ, M., WÖCKEL, A., KLEMENT, R. J. & KAPP, M. 2018. Beta-hydroxybutyrate (3-OHB) can influence the energetic phenotype of breast cancer cells, but does not impact their proliferation and the response to chemotherapy or radiation. *Cancer & metabolism*, 6, 1-19.
- BILLEN, D. 1990. Spontaneous DNA damage and its significance for the "negligible dose" controversy in radiation protection. *Radiation research*, 124, 242-245.
- BLAKELY, E., CHANG, P., LOMMEL, L., BJORNSTAD, K., DIXON, M., TOBIAS, C., KUMAR, K. & BLAKELY, W. F. 1989. Cell-cycle radiation response: role of intracellular factors. *Advances in Space Research*, 9, 177-186.
- CARDIS, E. & HATCH, M. 2011. The Chernobyl accident—an epidemiological perspective. *Clinical Oncology*, 23, 251-260.
- CHAKRABORTY, A., UECHI, T. & KENMOCHI, N. 2011. Guarding the 'translation apparatus': defective ribosome biogenesis and the p53 signaling pathway. *Wiley Interdisciplinary Reviews: RNA*, 2, 507-522.
- CHAURASIA, M., BHATT, A. N., DAS, A., DWARAKANATH, B. S. & SHARMA, K. 2016. Radiation-induced autophagy: mechanisms and consequences. *Free radical research*, 50, 273-290.
- CHENG, S.-C., SCICLUNA, B. P., ARTS, R. J. W., GRESNIGT, M. S., LACHMANDAS, E., GIAMARELLOS-BOURBOULIS, E. J., KOX, M., MANJERI, G. R., WAGENAARS, J. A. L. & CREMER, O. L. 2016. Broad defects in the energy metabolism of leukocytes underlie immunoparalysis in sepsis. *Nature immunology*, 17, 406-413.
- CHHETRI, D. R. 2019. Myo-Inositol and its derivatives: Their emerging role in the treatment of human diseases. *Frontiers in pharmacology*, 10, 1172.
- CHUNG, S., HANLON, C. D. & ANDREW, D. J. 2014. Building and specializing epithelial tubular organs: the *Drosophila* salivary gland as a model system for revealing how epithelial organs are specified, form and specialize. *Wiley Interdisciplinary Reviews: Developmental Biology*, 3, 281-300.
- COLE, L. A. 2016. *Biology of life: biochemistry, physiology and philosophy*, Academic Press.
- CONESA, A. & GÖTZ, S. 2008. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, 2008.
- COORDINATORS, N. R. 2015. Database resources of the national center for biotechnology information. *Nucleic acids research*, 43, D6-D17.

- CUI, R., CHAE, Y. & AN, Y.-J. 2017. Dimension-dependent toxicity of silver nanomaterials on the cladocerans *Daphnia magna* and *Daphnia galeata*. *Chemosphere*, 185, 205-212.
- DE FELICE, F., MARCHETTI, C., MARAMPON, F., CASCIALLI, G., MUZZI, L. & TOMBOLINI, V. 2019. Radiation effects on male fertility. *Andrology*, 7, 2-7.
- DE SOUSA, N., RODRÍGUEZ-ESTEBAN, G., ROJO-LAGUNA, J. I., SALÓ, E. & ADELL, T. 2018. Hippo signaling controls cell cycle and restricts cell plasticity in planarians. *PLoS biology*, 16, e2002399.
- DERENZINI, M., MONTANARO, L. & TRERE, D. 2017. Ribosome biogenesis and cancer. *Acta histochemica*, 119, 190-197.
- DINGES, N., MORIN, V., KREIM, N., SOUTHALL, T. D. & ROIGNANT, J.-Y. 2017. Comprehensive characterization of the complex *lola* locus reveals a novel role in the octopaminergic pathway via tyramine Beta-Hydroxylase regulation. *Cell reports*, 21, 2911-2925.
- DONATI, S., SANDER, T. & LINK, H. 2018. Crosstalk between transcription and metabolism: how much enzyme is enough for a cell? *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 10, e1396.
- DOS SANTOS, G. C., RENOVATO-MARTINS, M. & DE BRITO, N. M. 2021. The remodel of the “central dogma”: a metabolomics interaction perspective. *Metabolomics*, 17, 1-15.
- EMMS, D. M. & KELLY, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20, 1-14.
- ESTRELA, J. M., ORTEGA, A. & OBRADOR, E. 2006. Glutathione in cancer biology and therapy. *Critical reviews in clinical laboratory sciences*, 43, 143-181.
- EULALIO, A., HUNTZINGER, E., NISHIHARA, T., REHWINKEL, J., FAUSER, M. & IZAUARRALDE, E. 2009. Deadenylation is a widespread effect of miRNA regulation. *Rna*, 15, 21-32.
- FABREGAT, A., SIDIROPOULOS, K., VITERI, G., FORNER, O., MARIN-GARCIA, P., ARNAU, V., D'EUSTACHIO, P., STEIN, L. & HERMJAKOB, H. 2017. Reactome pathway analysis: a high-performance in-memory approach. *BMC bioinformatics*, 18, 1-9.
- FERRIER, D. E. K. 2016. Evolution of homeobox gene clusters in animals: the giga-cluster and primary vs. secondary clustering. *Frontiers in Ecology and Evolution*, 4, 36.
- FUMAGALLI, M., LECCA, D., ABBRACCHIO, M. P. & CERUTI, S. 2017. Pathophysiological role of purines and pyrimidines in neurodevelopment: unveiling new pharmacological approaches to congenital brain diseases. *Frontiers in pharmacology*, 8, 941.
- G. GRITSENKO, P., ILINA, O. & FRIEDL, P. 2012. Interstitial guidance of cancer invasion. *The Journal of pathology*, 226, 185-199.
- GAO, X., PUJOS-GUILLOT, E. & SÉBÉDIO, J.-L. 2010. Development of a quantitative metabolomic approach to study clinical human fecal water metabolome based on trimethylsilylation derivatization and GC/MS analysis. *Analytical chemistry*, 82, 6447-6456.
- GARNIER-LAPLACE, J., DELLA-VEDOVA, C., ANDERSSON, P., COPPLESTONE, D., CAILES, C., BERESFORD, N. A., HOWARD, B. J., HOWE, P. & WHITEHOUSE, P. 2010. A multi-criteria weight of evidence approach for deriving ecological benchmarks for radioactive substances. *Journal of Radiological Protection*, 30, 215.
- GENE ONTOLOGY, C. 2015. Gene ontology consortium: going forward. *Nucleic acids research*, 43, D1049-D1056.
- GOLLA, S., GOLLA, J. P., KRAUSZ, K. W., MANNA, S. K., SIMILLION, C., BEYOĞLU, D., IDLE, J. R. & GONZALEZ, F. J. 2017. Metabolomic analysis of mice exposed to gamma radiation reveals a systemic understanding of total-body exposure. *Radiation research*, 187, 612-629.
- GOMES, T., SONG, Y., BREDE, D. A., XIE, L., GUTZKOW, K. B., SALBU, B. & TOLLEFSEN, K. E. 2018. Gamma radiation induces dose-dependent oxidative stress and transcriptional alterations in the freshwater crustacean *Daphnia magna*. *Science of the total environment*, 628, 206-216.

- GROSSMANN, S., BAUER, S., ROBINSON, P. N. & VINGRON, M. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. *Bioinformatics*, 23, 3024-3031.
- HAFER, K., RIVINA, L. & SCHIESTL, R. H. 2010. Cell cycle dependence of ionizing radiation-induced DNA deletions and antioxidant radioprotection in *Saccharomyces cerevisiae*. *Radiation research*, 173, 802-808.
- HASIKOVA, L., KOZLIK, P., KALIKOVA, K., STIBURKOVA, B. & ZAVADA, J. 2020. OP0206 ALLANTOIN-A BIOMARKER OF OXIDATIVE STRESS-IS HIGHER IN PATIENTS WITH GOUT THAN IN HEALTHY VOLUNTEERS, AND CORRESPONDS WITH SEVERITY OF DISEASE. BMJ Publishing Group Ltd.
- HAYDON, P. G. 2012. Purinergic signaling. *Basic neurochemistry*. Elsevier.
- HERNÁNDEZ-DE-DIEGO, R., TARAZONA, S., MARTÍNEZ-MIRA, C., BALZANO-NOGUEIRA, L., FURIÓ-TARÍ, P., PAPPAS JR, G. J. & CONESA, A. 2018. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic acids research*, 46, W503-W509.
- HINZ, B. 2015. The extracellular matrix and transforming growth factor- β 1: Tale of a strained relationship. *Matrix biology*, 47, 54-65.
- HIRUTA, C. & TOCHINAI, S. 2012. How does the alteration of meiosis evolve to parthenogenesis?—Case study in a water flea, *Daphnia pulex*. *Meiosis-Molecular Mechanisms and Cytogenetic Diversity; Swan, A., Ed*, 109-122.
- HIRUTA, C. & TOCHINAI, S. 2014. Formation and structure of the ephippium (resting egg case) in relation to molting and egg laying in the water flea *Daphnia pulex* De Geer (Cladocera: Daphniidae). *Journal of morphology*, 275, 760-767.
- HONG, Y. P. & YANG, Y. J. 2017. Low-Dose Exposure to Bisphenol A in Early Life. In *Bisphenol A Exposure and Health Risks 2017*. InTech.
- HUANG, F., NI, M., CHALISHAZAR, M. D., HUFFMAN, K. E., KIM, J., CAI, L., SHI, X., CAI, F., ZACHARIAS, L. G. & IRELAND, A. S. 2018. Inosine monophosphate dehydrogenase dependence in a subset of small cell lung cancers. *Cell metabolism*, 28, 369-382.
- IMMARIGEON, C., BERNAT-FABRE, S., AUGÉ, B., FAUCHER, C., GOBERT, V., HAENLIN, M., WALTZER, L., PAYET, A., CRIBBS, D. L. & BOURBON, H.-M. G. 2019. Drosophila Mediator subunit Med1 is required for GATA-dependent developmental processes: divergent binding interfaces for conserved coactivator functions. *Molecular and cellular biology*, 39, e00477-18.
- IRANI, S., LOBO, J. M., GRAY, G. R. & TODD, C. D. 2018. Allantoin accumulation in response to increased growth irradiance in *Arabidopsis thaliana*. *Biologia plantarum*, 62, 181-187.
- ISSERLIN, R., MERICCO, D., VOISIN, V. & BADER, G. D. 2014. Enrichment Map—a Cytoscape app to visualize and explore OMICs pathway enrichment results. *F1000Research*, 3.
- JOLLY, M. K., BOARETO, M., HUANG, B., JIA, D., LU, M., BEN-JACOB, E., ONUCHIC, J. N. & LEVINE, H. 2015. Implications of the hybrid epithelial/mesenchymal phenotype in metastasis. *Frontiers in oncology*, 5, 155.
- JORGE, T. F. & ANTÓNIO, C. 2018. Plant metabolomics in a changing world: Metabolite responses to abiotic stress combinations. *Plant Abiotic Stress Responses Clim. Chang.*
- KANEHISA, M., FURUMICHI, M., SATO, Y., ISHIGURO-WATANABE, M. & TANABE, M. 2021a. KEGG: integrating viruses and cellular organisms. *Nucleic acids research*, 49, D545-D551.
- KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28, 27-30.
- KANEHISA, M., SATO, Y., FURUMICHI, M., MORISHIMA, K. & TANABE, M. 2019. New approach for understanding genome variations in KEGG. *Nucleic acids research*, 47, D590-D595.
- KANEHISA, M., SATO, Y. & KAWASHIMA, M. 2021b. KEGG mapping tools for uncovering hidden features in biological data. *Protein Science*.

- KELLERT, S. H. & SKLAR, L. 1997. In the wake of chaos: Unpredictable order in dynamical systems. *Philosophy of Science*, 64.
- KHANG, T. F. & LAU, C. Y. 2015. Getting the most out of RNA-seq data analysis. *PeerJ*, 3, e1360.
- KIM, S. K., CHOI, K. H. & KIM, Y. C. 2003. Effect of acute betaine administration on hepatic metabolism of S-amino acids in rats and mice. *Biochemical pharmacology*, 65, 1565-1574.
- KREAMER, B. L., SIEGEL, F. L. & GOURLEY, G. R. 2001. A novel inhibitor of β -glucuronidase: L-aspartic acid. *Pediatric research*, 50, 460-466.
- KUCERA, M., ISSERLIN, R., ARKHANGORODSKY, A. & BADER, G. D. 2016. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Research*, 5.
- LACKEY, K. H., POPE, P. M. & JOHNSON, M. D. 2003. Expression of 1L-myoinositol-1-phosphate synthase in organelles. *Plant Physiology*, 132, 2240-2247.
- LANGFELDER, P. & HORVATH, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9, 1-13.
- LAWLER, K. J. 2010. Transcriptional and post-transcriptional regulation of gene expression: computational analysis of microarray studies in fungal species.
- LI, L., STOECKERT, C. J. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13, 2178-2189.
- LI, W., WANG, F., MENUT, L. & GAO, F.-B. 2004. BTB/POZ-zinc finger protein abruptly suppresses dendritic branching in a neuronal subtype-specific and dosage-dependent manner. *Neuron*, 43, 823-834.
- LIU, F., LI, K., LI, J., HU, D., ZHAO, J., HE, Y., ZOU, Y., FENG, Y. & HUA, H. 2015. Apterous A modulates wing size, bristle formation and patterning in *Nilaparvata lugens*. *Scientific reports*, 5, 1-12.
- LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15, 1-21.
- LUNDGREN, S. E., CALLAHAN, C. A., THOR, S. & THOMAS, J. B. 1995. Control of neuronal pathway selection by the *Drosophila* LIM homeodomain gene *apterous*. *Development*, 121, 1769-1773.
- MAERE, S., HEYMANS, K. & KUIPER, M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21, 3448-3449.
- MAGLOTT, D., OSTELL, J., PRUITT, K. D. & TATUSOVA, T. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 33, D54-D58.
- MANDAL, L., BANERJEE, U. & HARTENSTEIN, V. 2004. Evidence for a fruit fly hemangioblast and similarities between lymph-gland hematopoiesis in fruit fly and mammal aorta-gonadal-mesonephros mesoderm. *Nature genetics*, 36, 1019-1023.
- MANJANG, K., TRIPATHI, S., YLI-HARJA, O., DEHMER, M. & EMMERT-STREIB, F. 2020. Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance. *Scientific reports*, 10, 1-16.
- MARBACH, D., COSTELLO, J. C., KÜFFNER, R., VEGA, N. M., PRILL, R. J., CAMACHO, D. M., ALLISON, K. R., KELLIS, M., COLLINS, J. J. & STOLOVITZKY, G. 2012. Wisdom of crowds for robust gene network inference. *Nature methods*, 9, 796-804.
- MARFIL, V., BLAZQUEZ, M., SERRANO, F., CASTELL, J. V. & BORT, R. 2015. Growth-promoting and tumorigenic activity of c-Myc is suppressed by Hhex. *Oncogene*, 34, 3011-3022.
- MASON, M. J., FAN, G., PLATH, K., ZHOU, Q. & HORVATH, S. 2009. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC genomics*, 10, 1-25.
- MAZANDU, G. K. & MULDER, N. J. 2014. Information content-based gene ontology functional similarity measures: which one to use for a given biological data type? *PLoS one*, 9, e113859.
- MERICO, D., ISSERLIN, R., STUEKER, O., EMILI, A. & BADER, G. D. 2010. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS one*, 5, e13984.

- MOU, T., DENG, W., GU, F., PAWITAN, Y. & VU, T. N. 2020. Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. *Frontiers in genetics*, 10, 1331.
- NATIONS, U. 2000. World Investment Report. *New York: United Nations*.
- NAVAS, L. E. & CARNERO, A. 2021. NAD⁺ metabolism, stemness, the immune response, and cancer. *Signal Transduction and Targeted Therapy*, 6, 1-20.
- NEPH, S., KUEHN, M. S., REYNOLDS, A. P., HAUGEN, E., THURMAN, R. E., JOHNSON, A. K., RYNES, E., MAURANO, M. T., VIERSTRA, J. & THOMAS, S. 2012a. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28, 1919-1920.
- NEPH, S., STERGACHIS, A. B., REYNOLDS, A., SANDSTROM, R., BORENSTEIN, E. & STAMATOYANNOPOULOS, J. A. 2012b. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150, 1274-1286.
- NEWELL, N. R., NEW, F. N., DALTON, J. E., MCINTYRE, L. M. & ARBEITMAN, M. N. 2016. Neurons that underlie *Drosophila melanogaster* reproductive behaviors: detection of a large male-bias in gene expression in fruitless-expressing neurons. *G3: Genes, Genomes, Genetics*, 6, 2455-2465.
- OECD 2008. OECD guideline for testing of chemicals 211. *Daphnia magna* reproduction test. Organisation for Economic Cooperation and Development (OECD) Paris, France.
- OLIVE, P. L. 1998. The role of DNA single- and double-strand breaks in cell killing by ionizing radiation. *Radiation research*, 150, S42-S51.
- ONOZATO, Y., KAIDA, A., HARADA, H. & MIURA, M. 2017. Radiosensitivity of quiescent and proliferating cells grown as multicellular tumor spheroids. *Cancer science*, 108, 704-712.
- PELLETIER, J., RIAÑO - CANALIAS, F., ALMACELLAS, E., MAUVEZIN, C., SAMINO, S., FEU, S., MENOYO, S., DOMOSTEGUI, A., GARCIA - CAJIDE, M. & SALAZAR, R. 2020. Nucleotide depletion reveals the impaired ribosome biogenesis checkpoint as a barrier against DNA damage. *The EMBO journal*, 39, e103838.
- PETERSEN, A. J., KATZENBERGER, R. J. & WASSARMAN, D. A. 2013. The innate immune response transcription factor relish is necessary for neurodegeneration in a *Drosophila* model of ataxia-telangiectasia. *Genetics*, 194, 133-142.
- PITTFNER, F. & HOFFMANN-OSTENHOF, O. 1976. Studies on the Biosynthesis of Cyclitols, XXXV [1]. On the Mechanism of Action of myo-Inositol-1-phosphate Synthase from Rat Testicles.
- PRAKASH, A. & MONTEIRO, A. 2018. apterous A specifies dorsal wing patterns and sexual traits in butterflies. *Proceedings of the Royal Society B: Biological Sciences*, 285, 20172685.
- PRAKASH, V., CARSON, B. B., FEENSTRA, J. M., DASS, R. A., SEKYROVA, P., HOSHINO, A., PETERSEN, J., GUO, Y., PARKS, M. M. & KURYLO, C. M. 2019. Ribosome biogenesis during cell cycle arrest fuels EMT in development and disease. *Nature communications*, 10, 1-16.
- PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33, D501-D504.
- PUJARI, G., BERNI, A., PALITTI, F. & CHATTERJEE, A. 2009. Influence of glutathione levels on radiation-induced chromosomal DNA damage and repair in human peripheral lymphocytes. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 675, 23-28.
- QIAN, L. & BODMER, R. 2009. Partial loss of GATA factor Pannier impairs adult heart function in *Drosophila*. *Human molecular genetics*, 18, 3153-3163.
- RITCHIE, M. E., PHIPSON, B., WU, D. I., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43, e47-e47.
- SAELEN, W., CANNODT, R. & SAEYS, Y. 2018. A comprehensive evaluation of module detection methods for gene expression data. *Nature communications*, 9, 1-12.

- SANZARI, J. K., WAN, X. S., KRIGSFELD, G. S., WROE, A. J., GRIDLEY, D. S. & KENNEDY, A. R. 2013. The effects of gamma and proton radiation exposure on hematopoietic cell counts in the ferret model. *Gravitational and space research: publication of the American Society for Gravitational and Space Research*, 1, 79.
- SCHINAMAN, J. M., GIESEY, R. L., MIZUTANI, C. M., LUKACSOVICH, T. & SOUSA-NEVES, R. 2014. The KRÜPPEL-like transcription factor DATILÓGRAFO is required in specific cholinergic neurons for sexual receptivity in *Drosophila* females. *PLoS biology*, 12, e1001964.
- SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13, 2498-2504.
- SHE, M., TANG, M., JIANG, T. & ZENG, Q. 2021. The roles of the LIM domain proteins in *Drosophila* cardiac and hematopoietic morphogenesis. *Frontiers in Cardiovascular Medicine*, 8.
- SHEN, L. 2014. GeneOverlap: An R package to test and visualize gene overlaps. *R Package*.
- SHIMBORI, C., GAULDIE, J. & KOLB, M. 2013. Extracellular matrix microenvironment contributes actively to pulmonary fibrosis. *Current opinion in pulmonary medicine*, 19, 446-452.
- SHUVALOV, O., PETUKHOV, A., DAKS, A., FEDOROVA, O., VASILEVA, E. & BARLEV, N. A. 2017. One-carbon metabolism and nucleotide biosynthesis as attractive targets for anticancer therapy. *Oncotarget*, 8, 23955.
- SIDDIQUI, A. & CEPPI, P. 2020. A non-proliferative role of pyrimidine metabolism in cancer. *Molecular metabolism*, 35, 100962.
- SONG, Y., XIE, L., LEE, Y., BREDE, D. A., LYNE, F., KASSAYE, Y., THAULOW, J., CALDWELL, G., SALBU, B. & TOLLEFSEN, K. E. 2020. Integrative assessment of low-dose gamma radiation effects on *Daphnia magna* reproduction: Toxicity pathway assembly and AOP development. *Science of the Total Environment*, 705, 135912.
- SOTO - HEREDERO, G., GOMEZ DE LAS HERAS, M. M., GABANDÉ - RODRÍGUEZ, E., OLLER, J. & MITTELBRUNN, M. 2020. Glycolysis – a key player in the inflammatory response. *The FEBS journal*, 287, 3350-3369.
- STREFFER, C. 2004. Bystander effects, adaptive response and genomic instability induced by prenatal irradiation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 568, 79-87.
- SUN, Y. V. & HU, Y.-J. 2016. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Advances in genetics*, 93, 147-190.
- TAYLOR, M. A., SOSSEY-ALAOUI, K., THOMPSON, C. L., DANIELPOUR, D. & SCHIEMANN, W. P. 2013. TGF- β upregulates miR-181a expression to promote breast cancer metastasis. *The Journal of clinical investigation*, 123, 150-163.
- TODOROVIC, V. & RIFKIN, D. B. 2012. LTBP3, more than just an escort service. *Journal of cellular biochemistry*, 113, 410-418.
- TREFFKORN, S. & MAYER, G. 2019. Expression of NK genes that are not part of the NK cluster in the onychophoran *Euperipatoides rowelli* (Peripatopsidae). *BMC developmental biology*, 19, 1-22.
- UBELLACKER, J. M., TASDOGAN, A., RAMESH, V., SHEN, B., MITCHELL, E. C., MARTIN-SANDOVAL, M. S., GU, Z., MCCORMICK, M. L., DURHAM, A. B. & SPITZ, D. R. 2020. Lymph protects metastasizing melanoma cells from ferroptosis. *Nature*, 585, 113-118.
- ULRICH, K. & JAKOB, U. 2019. The role of thiols in antioxidant systems. *Free Radical Biology and Medicine*, 140, 14-27.
- VAN DAM, S., VOSA, U., VAN DER GRAAF, A., FRANKE, L. & DE MAGALHAES, J. P. 2018. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19, 575-592.

- VANDENBERG, L. N., COLBORN, T., HAYES, T. B., HEINDEL, J. J., JACOBS JR, D. R., LEE, D.-H., SHIODA, T., SOTO, A. M., VOM SAAL, F. S. & WELSHONS, W. V. 2012. Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose responses. *Endocrine reviews*, 33, 378-455.
- WANG, S.-T., CHEN, H.-W., SHEEN, L.-Y. & LII, C.-K. 1997. Methionine and cysteine affect glutathione level, glutathione-related enzyme activities and the expression of glutathione S-transferase isozymes in rat hepatocytes. *The Journal of nutrition*, 127, 2135-2141.
- WANG, W., FRIDMAN, A., BLACKLEDGE, W., CONNELLY, S., WILSON, I. A., PILZ, R. B. & BOSS, G. R. 2009. The phosphatidylinositol 3-kinase/akt cassette regulates purine nucleotide synthesis. *Journal of Biological Chemistry*, 284, 3521-3528.
- WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R. & FEDERHEN, S. 2007. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36, D13-D21.
- XU, L., DONG, Z., FANG, L., LUO, Y., WEI, Z., GUO, H., ZHANG, G., GU, Y. Q., COLEMAN-DERR, D. & XIA, Q. 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic acids research*, 47, W52-W58.
- YANG, C., KO, B., HENSLEY, C. T., JIANG, L., WASTI, A. T., KIM, J., SUDDERTH, J., CALVARUSO, M. A., LUMATA, L. & MITSCHKE, M. 2014. Glutamine oxidation maintains the TCA cycle and cell survival during impaired mitochondrial pyruvate transport. *Molecular cell*, 56, 414-424.
- YE, L. F., CHAUDHARY, K. R., ZANDKARIMI, F., HARKEN, A. D., KINSLOW, C. J., UPADHYAYULA, P. S., DOVAS, A., HIGGINS, D. M., TAN, H. & ZHANG, Y. 2020. Radiation-induced lipid peroxidation triggers ferroptosis and synergizes with ferroptosis inducers. *ACS chemical biology*, 15, 469-484.
- YU, G. & HE, Q.-Y. 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12, 477-479.
- YU, G., WANG, L.-G., HAN, Y. & HE, Q.-Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16, 284-287.
- ZHANG, B. & HORVATH, S. 2005. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4.
- ZHANG, L., ZHOU, F. & TEN DIJKE, P. 2013. Signaling interplay between transforming growth factor- β receptor and PI3K/AKT pathways in cancer. *Trends in biochemical sciences*, 38, 612-620.
- ZHU, F.-Y., CHEN, M.-X., YE, N.-H., QIAO, W.-M., GAO, B., LAW, W.-K., TIAN, Y., ZHANG, D., ZHANG, D. & LIU, T.-Y. 2018. Comparative performance of the BGISEQ-500 and Illumina HiSeq4000 sequencing platforms for transcriptome analysis in plants. *Plant Methods*, 14, 1-14.