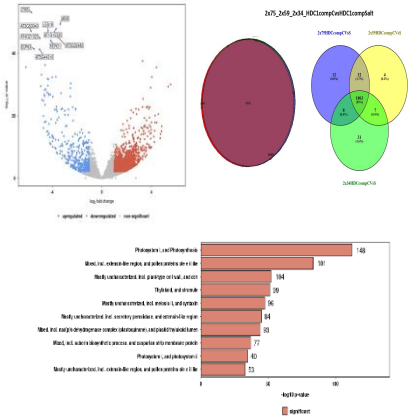
Evaluating the impact of read length on RNAseq-based differential expression using Galaxy and Searchlight.



MSc Biotechnology 2022-23

BIOL5228P-Dry Investigative Project

Student ID: 2805500k

2805500k@student.gla.ac.uk

07.08.2023

Supervisor

Dr. Pawel Herzyk

Table of contents

Abstract:

Background: RNA-seq has emerged as a significant tool for studying the transcriptome, providing enhanced resolution, and reduced technical variability. Continuous research in this field strives to improve gene expression analysis tools, enhancing their effectiveness and value for scientific research and applications. Read lengths play an important role in RNA-seq analysis as they directly impact the accuracy, sensitivity, and resolution of the result. The leading Illumina technology allows generating short single-end or paired-end reads of different lengths at a different cost to the consumer. *Aim:* Our main objective is to investigate whether the performance of RNAseq achieved with the golden standard paired-end 2x75bp reads could be matched by using shorter reads of 2x59bp and 2x34bp and to what extent the paired-end performance could be matched by single-end one. *Methods:* The RNA-seq analysis involved samples with three read lengths of four biological conditions on Galaxy. Specific pipelines were utilized to process and analyze the data, accounting for the read length variations. *Results:* The study shows that 2 × 59 paired-end reads exhibit a stronger correlation with the golden standard 2 × 75 paired-end reads in this RNA-seq analysis, compared to 1 × 75 single-end reads. The correlations are stronger at the transcripts, genes, and functional levels, favoring 2 × 59 paired-end reads over 1 × 75 single-end reads. This indicates that 2 × 59 paired-end reads provide more precise and dependable results for cost-effective differential expression analysis. *Conclusion:* Shorter paired-end reads offer cost-effective gene-level expression analysis, robust expression estimates, and isoform-level differentiation results.

Abbreviations:

RNA-Seq – RNA sequencing

HDC1-Histone Deacetylation Complex 1

HDC1compC- Histone Deacetylation Complex 1 Control

HDC1compS- Histone Deacetylation Complex 1 Salt

KO- Knockout

KOC- Knockout Control

KOS- Knockout Salt

DEG – differentially expressed genes.

DGE – differential gene expression

PCA - Principal component analysis

SE- Single-ended

PE- Paired-ended

1. Introduction:

Over the past few decades, transcriptome profiling has emerged as a highly popular method for investigating differences between normal and diseased conditions at the molecular level (Hong *et al*., 2020). RNA sequencing has brought about a revolutionary transformation in this field by enabling the quantification of gene expression levels, the detection of new genes and introns, and the exploration and analysis of different variations of mRNA and gene fusions (Casamassimi *et al.*, 2017). One notable application is the identification of differentially expressed genes by comparing changes in gene expression across different biological conditions. Consequently, this analysis of differentially expressed genes (DEG) aids in the identification of pathways implicated in different diseases and enables the development of personalized treatments to enhance effectiveness (Udhaya Kumar et al., 2020). Advancements in sequencing technologies have resulted in improved accuracy, higher throughput, and decreased costs. The cost of sequencing has significantly reduced since the completion of the first human genome in 2003 (McCombie WR and McPherson JD., 2019). Although sequencing technologies have made significant advancements in read length, it is the short read Illumina technology which is the most effective in DEG analysis. Yet different Illumina sequencers generate different read lengths at different cost, the longer the read, the higher the cost, and it is not overwhelmingly clear whether, for some sequencers, the performance of longer reads could be replicated with shorter, most cost-effective reads. DEG identification typically necessitates familiarity with various command-line tools and scripting. However, to streamline the process and enhance user-friendliness, a web-based tool named Galaxy has been developed for this analysis.

1.1 Objectives:

Glasgow Polyomics (GP) research facility recently upgraded its aging Illumina NextSeq 500 sequencer to the state-of-the-art NextSeq 2000. Consequently, the paired-end 2x75bp sequencing, which had been used for years as the golden standard of RNAseq, was no longer available as for NextSeq 2000 sequencer Illumina introduced “50 cycles”, “100 cycles” and “200 cycles” sequencing kits only. The latter, which allowed PE 2x100bp reads, has been used routinely in GP as its performance should not compromise the 2x75bp golden standard.

In this study, we are asking whether the performance of the RNAseq based on the golden standard 2x75bp could be matched by NextSeq 2000 sequencer run with “100 cycles” or even “50 cycles” sequencing kits, which would be far more cost-effective than the current “200 cycles” kit. Both the above kits have 38-cycle redundancy and can provide 138 and 88 cycles, respectively, resulting in paired-end 2x59bp and 2x34bp reads, given that 20 cycles are reserved for sample barcoding. Additionally, we are also investigating to what extent the performance achieved with paired-end reads could be matched with the corresponding single-end reads.

In order to achieve the above aims, we will identify differentially expressed genes (DEGs) and gene ontologies of datasets generated with different read lengths, including 75, 59, and 34 bps for both single-end and paired-end reads. By doing so, we aimed to gain a comprehensive understanding of the impact of read length variations on DEG identification and gene ontology analysis in the context of our research. The analysis was performed on the raw RNA-sequenced data to gain insights into the gene expression patterns and functional annotations of the studied biological samples.

In the study, both single-end and paired-end datasets were obtained and trimmed to the required read length. After undergoing quality checks to ensure data integrity, the reads were mapped to the reference genome. After the alignment process was finished, the counted mapped reads were used to measure the expression levels of genes. Ultimately, differential expression analysis was executed to pinpoint genes exhibiting significant changes in expression between the various conditions or samples under investigation. Each dataset's results were compared to the established 2x75 golden standard to ascertain whether shorter paired-end reads (2x59) exhibited a stronger correlation with it.

2. Materials:

2.1 RNA seq Dataset:

This study utilized biological samples prepared in Prof. Anna Amtmann's laboratory in the Plant Science Group in the School of Molecular Biosciences at the University of Glasgow. The sequencing was performed in the Glasgow Polyomics research facility. The samples came from a two-factorial experiment where responses to salt stress (Salt versus Control) were investigated in *Arabidopsis thaliana* seedlings in two genotypes, one involving the Histone Deacetylation Complex 1 (HDC1), which regulates the sensitivity of Arabidopsis seedlings to salt stress(equivalent to wild-type), and the other where HDC1 was knocked out(KO). Consequently, there were four experimental conditions, namely HDC1-Control (HDC1compC), HDC1-Salt (HDCcompSalt), KO-Control (KOC) and KO-Salt (KOSalt). Each of the conditions had three biological replicates, which gave a total of 12 samples, resulting in the initial set of 24 fastqc files (two files for read 1 and read 2 separately per sample)

The datasets used in this study are accessible at the NCBI Gene Expression Omnibus (GEO) with the accession number GSE205893. Currently, the data is private, but it is scheduled to be released on June 10, 2024. The reads were mapped to the reference genome of Arabidopsis thaliana (TAIR10.1).

3. Methods:

In this study, our main goal was to identify differentially expressed genes (DEGs) and gene ontologies from datasets with different read lengths, including 75, 59, and 34 for both single-end and paired-end reads. Here, we utilized the original 2x75bp data-sets and, subsequently, the trimmed data sets with different read lengths, including 2x59, 2x34, 1x75, 1x59, and 1x34. The single-end data sets were created by considering only the first read of the original paired reads. Throughout the analysis, we considered the 2x75 read length as the golden standard, against which we compared and evaluated the results from other read lengths. By doing so, we aimed to gain a comprehensive understanding of the impact of read length variations on DEG identification and gene ontology analysis in the context of our research. The analysis was performed on the raw RNA-sequenced data to gain insights into the gene expression patterns and functional annotations of the studied biological samples.

3.1 Preprocessing datasets :

The raw reads in fastq format were obtained and uploaded to the Galaxy server, an open web tool widely used for different omics analyses. Galaxy offers a comprehensive RNA-seq analysis pipeline specifically designed for identifying differentially expressed genes (The Galaxy Community., 2022). To ensure the accuracy of the sequenced data and prevent potential

errors, FASTQC, a quality check was conducted on all datasets (Adrews., 2010). The raw reads underwent trimming to attain the required length for the experiment using Trimmomatic, and the tool also eliminated low-quality reads and adapter contents (Bolger, Lohse, and Usadel, 2014). After the reads were trimmed, they were subjected to quality assessment using FASTQC to verify that the trimming process successfully eliminated any potential errors.

3.2 Expression Analysis:

Following that, those reads were mapped to the reference genome using HISAT2. HISAT2 is an efficient aligner for mapping sequencing reads to a reference genome, particularly for repetitive sequences. Its indexing algorithm for repeat sequences addresses this challenge by projecting alignments of redundant reads to a single location within the repetitive region. HISAT2 stores information about other potential alignment locations, allowing it to recover them if needed. This approach balances speed and accuracy when dealing with repetitive regions in the genome. It efficiently aligns reads to the reference genome while allowing for the exploration and evaluation of multiple alignment possibilities for repetitive reads. (Kim *et al*., 2019). It is followed by Feature Counts, a program designed to summarize the number of reads in RNA sequencing experiments. It employs efficient techniques, making it faster than current methods and with low memory requirements. It supports both single and paired-end reads and provides a range of sequencing options (Liao, Smyth, and Shi, 2013). Once the gene-specific read counts were acquired, the analysis for differential expression was carried out using DESeq2. (v1.30.1) (Love, Huber, and Anders, 2014)). DESeq2 performed a comparison of read counts between the control and knockout groups of samples for various read lengths to identify the differentially expressed (DE) genes. DESeq2 normalizes gene count numbers to correct for variations in sequencing depth, and it employs the Benjamini and Hochberg method to address multiple testing issues, and the adjusted p-value was set to (adj p-value < 0.05). Then, the DE genes are categorized into two groups, upregulated and downregulated genes, based on the criteria of log2FoldChange > 1 and log2FoldChange < -1, respectively. DESeq2 generates lists of DE genes using these criteria, which are then used to generate Venn diagrams using the eulerr package (v7.0.0, Larsson., 2022)) and the web tool Venny (v2.1, Oliveros., 2007-2015). These Venn diagrams enable a comparison of the numbers of DE genes among the different read lengths.

Searchlight is an open-source RNA-seq visualization and interpretation pipeline that functions similarly to big data tools. It provides extensive statistical and visual analysis capabilities at various levels, including pathways, most expressed genes, and gene ontologies. It can accommodate a wide range of differential experimental designs and generates reports. Moreover, it allows users to customize plots using R-scripts and a Shiny app in subsequent analyses (Cole et al., 2021). We compared the significant genes against various gene ontology and pathway analysis databases to identify enriched gene sets. The analysis was based on the hypergeometric test, considering p-values and log2 fold changes.

The overall chart showing the methodology details is presented in Figure 1.

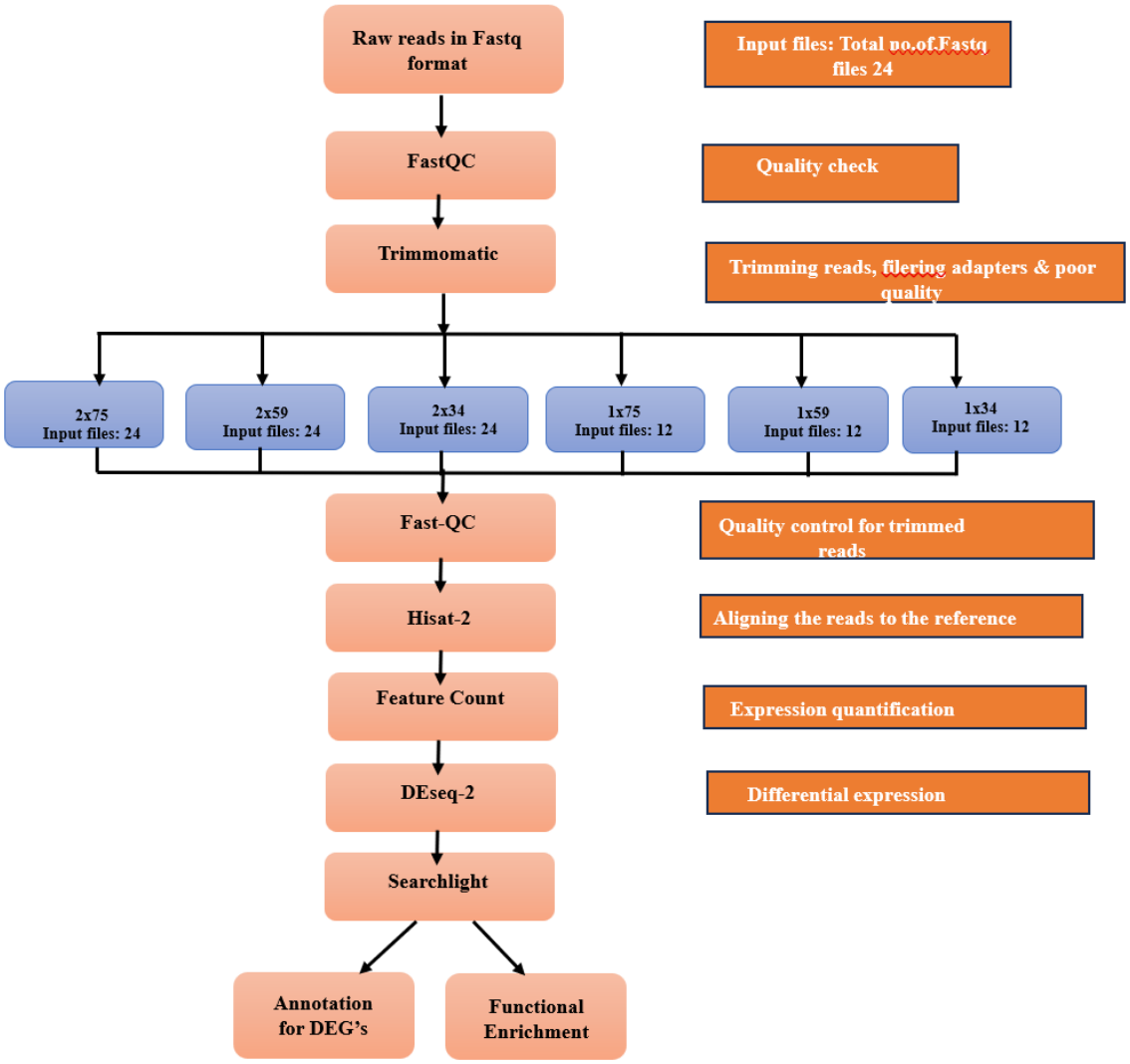


Figure: 1. A flowchart illustrating the workflow of raw reads. *The initial step involves quality-checking the raw reads in FastQ format using Fastqc. Subsequently, the reads are trimmed to specific lengths (75, 59, and 34 bp) while eliminating low-quality segments and adapter sequences using Trimmomatic. After the trimming process, another quality check is performed using Fastqc and the reads are aligned to the reference genome using Hisat2, and the read counts per gene are subsequently calculated using Feature count. Next, differentially expressed genes are estimated using Deseq2. Finally, the obtained estimates are annotated, plots are generated, and enrichment analysis is conducted to complete the downstream analysis.*

4. Results:

4.1. Comparing PCA Plots of Biological Conditions Using Different Read Lengths

In this section, we present the results of comparing the principal component analysis (PCA) plots generated from RNA-seq data obtained with different read lengths. The read lengths investigated includes paired-end reads of (2x75, 2x59, and 2x34) and single-ended reads of (1x75, 1x59, and 1x34). The reference condition with a 2x75 read length is considered the golden standard for comparison.

4.1.1 PCA plot comparison for pipelines with different paired-end read lengths:

In the PCA analysis presented in Figure 2, we observed consistent and tight clustering of biological replicates in each condition across all read lengths. This result indicates that the choice of read length in the paired-end sequencing did not significantly impact the grouping of samples. 85% of between-sample variation can be attributed to PC1 and relates to the difference between Salt and Control in both genotypes, with a significantly higher variation occurring in the KO genotype. The second principal component an explained additional 5% of variance which could be broadly attributed to the difference between genotypes.

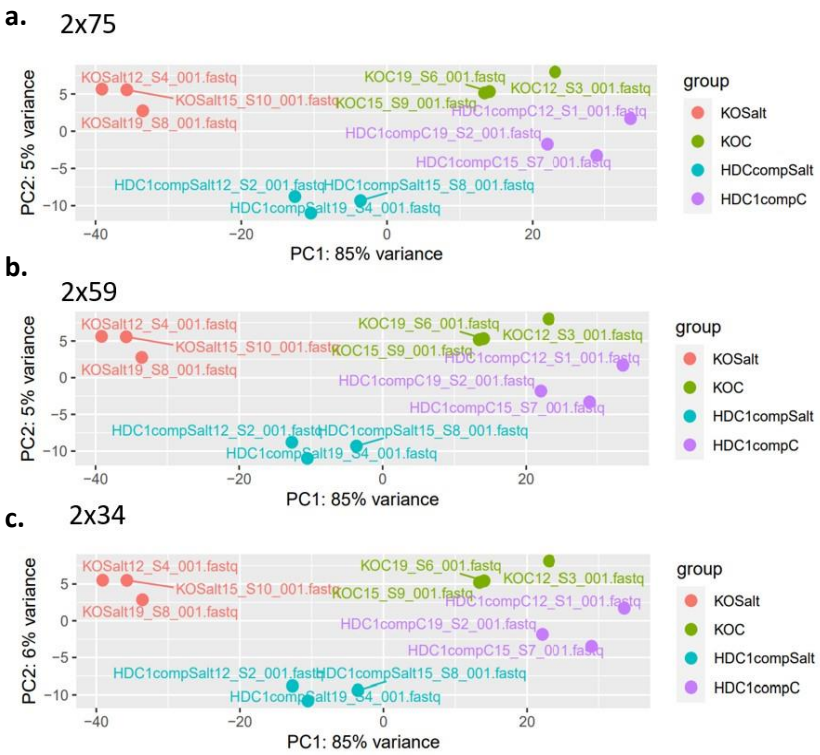


Figure 2: PCA plot- comparison for different paired-end read lengths. *a. indicates the read length 2x75, b. indicates the comparison of HDC1compControl vs HDC1compSalt for the read length 2x59, and c. indicates the read length 2x34.*

4.1.2 PCA plot comparison for pipelines with different single-end read lengths:

The PCA plots for 1x59 and 1x34 seem identical. Although the PCA plot for 1x75 seems different, it becomes obvious that after flipping over the PC1 axis followed by flipping over

the PC2 axis, the 1x75 plot will match the other two. Therefore, we can conclude that the PCA results for single-end reads closely match the results obtained from paired-end reads.

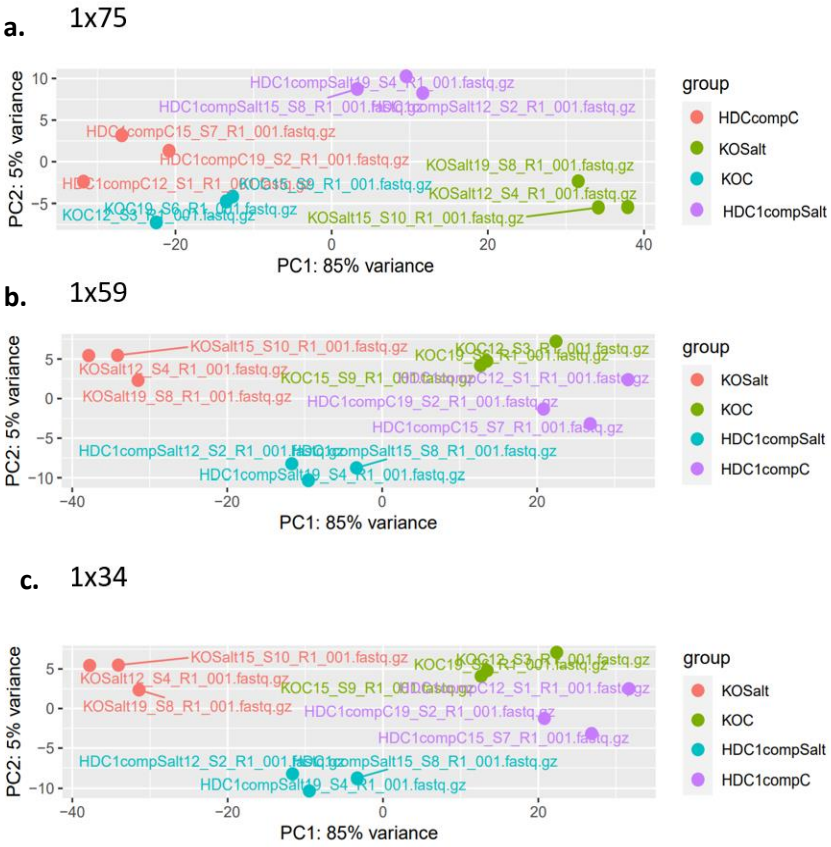


Figure 3: PCA plot comparison for different single-end read lengths. *a. indicates the read length 2x75,b. indicates read length 2x59, and c. indicates the read length 2x34.*

4.2. Difference in Read length affects DEG:

The differentially expressed genes between Salt and Control in two different genotypes were identified with DESeq2 software, as explained in the Methods section. Specifically, we compared read lengths of 2x75, 2x59, 2x34, 1x75, 1x59, and 1x34 in two separate comparisons: HDC1compCvsHDC1compSalt and KOCvsKOSalt.

The results revealed the following number of DEG for each read length in the HDC1compCvsHDC1compSalt comparison: 1855 (2x75), 1846 (2x59), 1849 (2x34), 1343 (1x75), 1785 (1x59), and 1782 (1x34) genes. Similarly, for the KOCvsKOSalt comparison, the number of DEGs for each read length was as follows: 4728 (2x75), 4736 (2x59), 4728 (2x34), 5619 (1x75), 4536 (1x59), and 4512 (1x34) genes.

We plotted the results in a bar graph in Figure 4 a and b, which clearly depicted the variation in the number of DEGs based on different read lengths for both comparisons. Notably, the KOCvsKOSalt comparison resulted in a higher number of DEGs compared to HDC1compCvsHDC1compSalt, suggesting a more pronounced gene expression difference in

the knockout genotype. The top 10 DE genes for each dataset are listed in Table A1- A12, respectively.

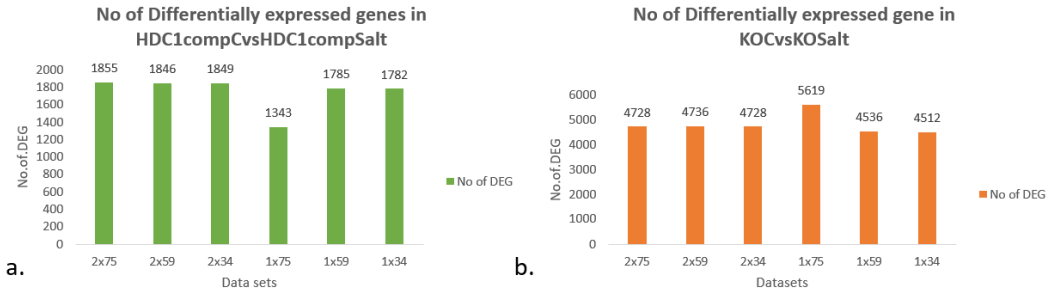


Figure 4: Total number of differentially expressed genes in the datasets. *A bar graph representing the total number of DEG in the wild-type and the knocked-out conditions. a. It represents the number of DEG's for datasets 2x75, 2x59, 2x34, 1x75, 1x59, and 1x34 for HDC1compControl vs HDC1compSalt. b. It represents the number of DEG for datasets 2x75, 2x59, 2x34, 1x75, 1x59, and 1x34 for KOControl vs KOSalt.*

4.3. Analysing the correlation of DEG with the golden standard 2x75 of HDC1compCvsHDC1compSalt and Knocked out Control and Salt:

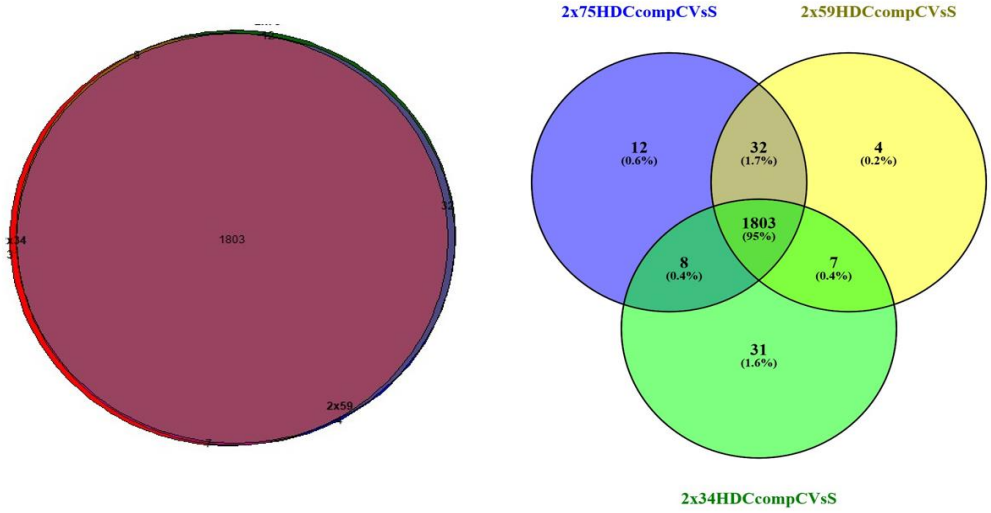
When comparing the gene expression of 2x59 and 2x34 to the golden standard 2x75 in HDC1compCvsSalt, we represented each comparison as a Venn diagram. The Venn diagrams visually displayed the overlap of differentially expressed genes (DEGs) between the read lengths.

In both HDC1compCvsSalt and KOCvsKOSalt comparisons, the Venn diagrams for 2x59 and 2x34 exhibited a significant overlap with the 2x75 dataset, indicating a strong correlation between these short paired-end read lengths and the golden standard. This suggests that 2x59 and 2x34 reliably captured similar gene expression patterns as the 2x75 dataset.

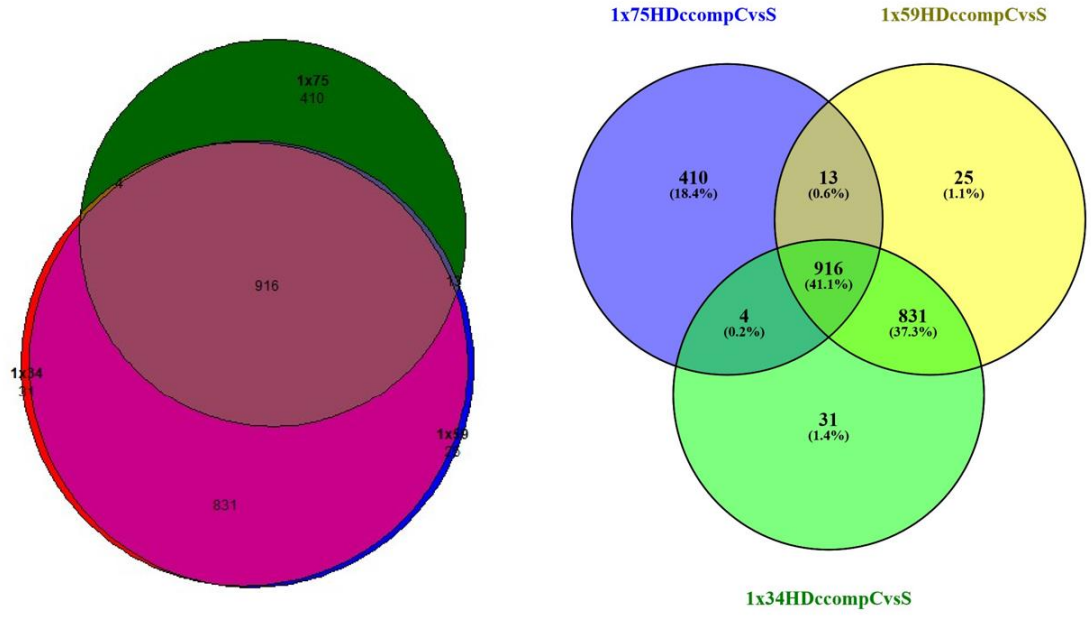
On the other hand, in the single-end read dataset (1x75), we observed notable deviations from the 2x75 golden standard in both HDC1compCvsSalt and KOCvsKOSalt comparisons. The Venn diagrams indicated fewer shared DEGs with the 2x75 dataset, implying reduced agreement between 1x75 and 2x75 expression profiles.

However, among the long single-end read datasets, we noticed that 1x59 and 1x34 exhibited lesser deviations from the 2x75 dataset. These read lengths demonstrated a stronger correlation with the golden standard, capturing more similar DEGs.

a. 2x75_2x59_2x34_HDC1compCvsHDC1compSalt



b. 1x75 _1x59_1x34 _HDC1compCvsHDC1compSalt



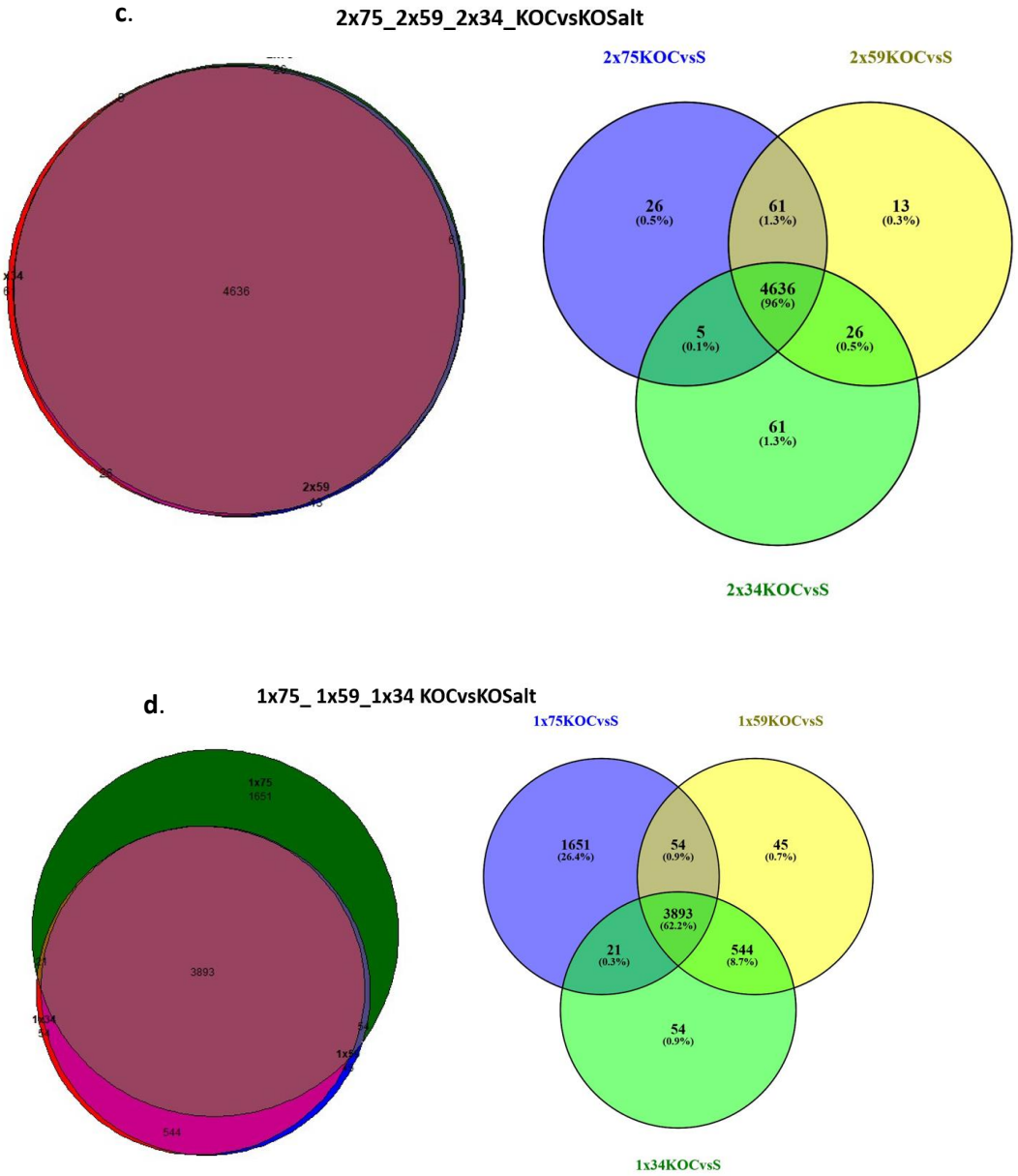
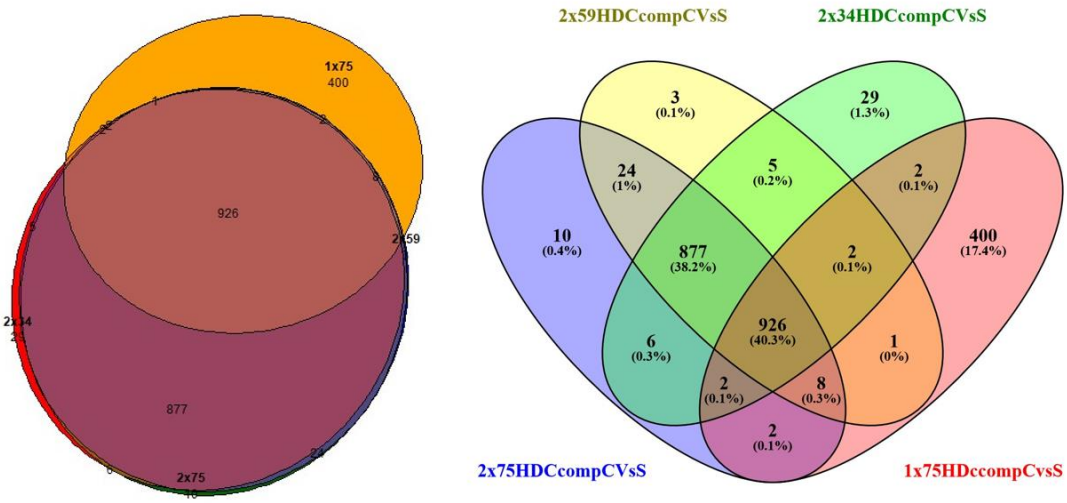


Figure 5: Overlap of differentially expressed genes between different comparisons generated using various read lengths. *Differentially expressed genes from various datasets were represented in Venn diagrams a. paired-end comparisons of HDC1compCvsHDC1compSalt of 2x75, 2x59, and 2x34 b. single-end comparisons of HDC1compCvsHDC1compSalt of 1x75, 1x59, and 1x34 c. paired end comparisons of KOCvsKOSalt of 2x75, 2x59, and 2x34 d. single end comparisons of KOCvsKOSalt of 1x75, 1x59, and 1x34. In the paired-end comparison of wild-type and KO conditions, both 59 and 34 show higher overlapping regions with 79 datasets, and in In the single-end comparison of wild-type and KO conditions, both 59 and 34 show higher deviations from 75 datasets.*

To investigate how well single-end data can match the results obtained with paired-end data, we compared short paired-end datasets (2x59, 2x34) and the long single-end dataset (1x75) in both wild-type and knocked-out conditions with the golden standard 2x75.

From the Venn diagrams presented in Figure 6, we observed that in both wild-type and knocked-out conditions, 2x59 exhibited the highest level of correlation with the golden standard 2x75. The largest overlap of differentially expressed genes (DEGs) was observed between the samples labeled as "2x59" and "2x75," indicating a robust agreement between their expression profiles. Next in correlation came 2x34, which also showed a significant overlap with the golden standard, though slightly less than 2x59.On the other hand, the long single-end dataset (1x75) exhibited deviations from the golden standard in both conditions. The shared DEGs between 1x75 and 2x75 were fewer, highlighting the discrepancies between their gene expression patterns.

a. 2x75_2x59_2x34_1x75_HDC1compCvsHDC1compSalt



b. 2x75_2x59_2x34_KOCvsKOSalt

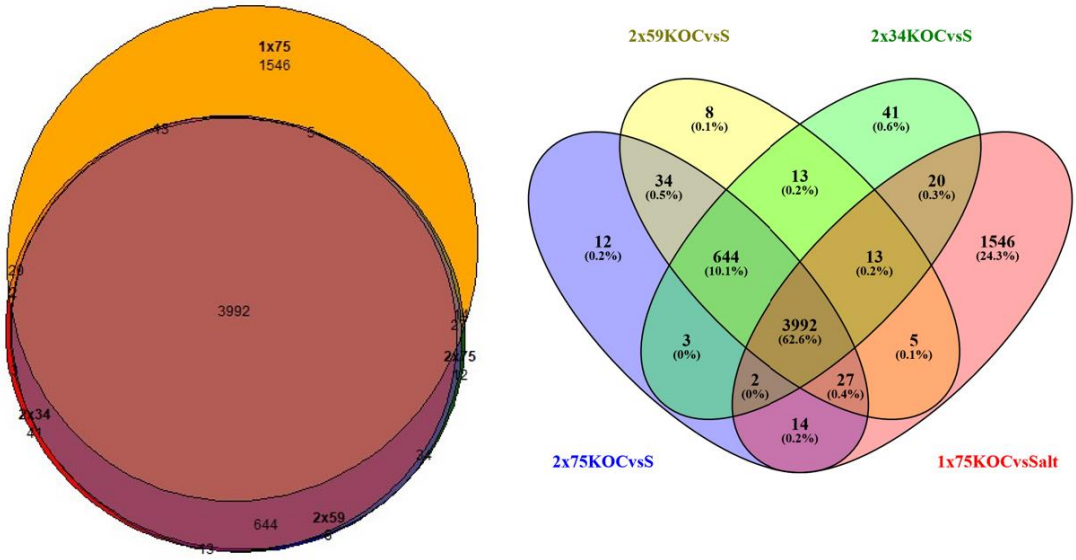


Figure 6: *Overlap of differentially expressed genes between different comparisons generated using paired-end and single-end reads, namely 2x59, 2x34, and 1x75. a. represents the wild-type genotype's comparison HDC1compCvsHDC1compSalt and b. represents the knockout genotype's comparison KOCvsKOSalt. The Venn diagrams illustrate the comparisons. In both scenarios, the overlap between 2x59 and 2x34 was substantial, with 2x59 being the clear leader. On the other hand, 1x75 had only a small, shared region and exhibited more variations compared to the other cases.*

4.4. Analysing the most expressed genes of DEG with the golden standard 2x75 of HDC1compC, HDC1compSalt, KOC, and KOSalt:

We compiled a list of the top ten most differentially expressed genes in the wild-type and KO samples for all datasets, including 2x59, 2x34, 1x75, 1x59, and 1x34. These gene lists were compared with the golden standard dataset (2x75), and the genes were ranked based on their mean expression in each sample group.

Our observations revealed that the dataset with 2x59 reads displayed a strong correlation with the golden standard across almost all sample groups. The order of genes listed based on mean expression closely matched that of the golden standard. However, we noted some discrepancies in the gene lists. Genes shown in red colour indicate mismatches with the golden standard, while other genes in bold had their order changed in the list. Despite these variations, 2x34 and 1x75 datasets followed 2x59 in sharing identity with the golden standard, albeit with differing gene orders.

Table1. ***Most expressed gene list for various read lengths in different biological conditions***

a. Top ten expressed gene list in HDC1compC

Golden std	HDC1compC				
2x75	2x59	2x34	1x75	1x59	1x34
RBCS1A	RBCS1A	RBCS1A	RBCS1A	RBCS1A	RBCS1A
ICL	ICL	ICL	ICL	ICL	ICL
CAB1	CAB1	CAB1	CAB1	CAB1	PYK10
PYK10	PYK10	PYK10	PYK10	PYK10	**CAB1**
AT5G60390	AT5G60390	AT5G60390	AT5G60390	AT5G60390	**AAC1**
AAC1	AAC1	AAC1	AAC1	AAC1	**AT5G60390**
ATMS1	ATMS1	ATMS1	ATMS1	ATMS1	ATMS1
AT3G16420	AT3G16420	AT3G16420	AT3G16420	AT3G16420	AT3G16420
CAB2	CAB2	GAPC1	**HSC70-1**	GAPC1	HSC70-1
HSC70-1	GAPC1	HOG1	HOG1	HSC70-1	HOG1

b. Top ten expressed gene list in HDC1compSalt

Golden Std	HDC1compsalt				
2X75	2x59	2x34	1x75	1x59	1x34
ICL	ICL	ICL	ICL	ICL	ICL
RBCS1A	RBCS1A	R3CS1A	RBCS1A	RBCS1A	RBCS1A
PYK10	PYK10	PYK10	PYK10	PYK10	PYK10
AT5G60390	AT5G60390	AT5G60390	AT5G60390	AT5G60390	AT5G60390
CAB1	CAB1	**AAC1**	**AAC1**	**AAC1**	**AAC1**
AAC1	AAC1	**CAB1**	**CAB1**	**HSC70-1**	**HSC70-1**
HSC70-1	HSC70-1	**ATMS1**	HSC70-1	**ATMS1**	**ATMS1**
ATMS1	ATMS1	**HSC70-1**	ATMS1	**MLS**	**MLS**
PCK1	PCK1	PCK1	PCK1	PCK1	PCK1
MLS	MLS	MLS	MLS	**CAB1**	HOG1

c. Top ten expressed gene list in KOC

Golden Std	KOC				
2X75	2x59	2x34	1x75	1x59	1x34
RBCS1A	RBCS1A	RBCS1A	RBCS1A	RBCS1A	ICL
ICL	ICL	ICL	ICL	ICL	RBCS1A
CAB1	CAB1	**PYK10**	**PYK10**	**PYK10**	PYK10
PYK10	PYK10	**CAB1**	**CAB1**	**CAB1**	AAC1
AT5G60390	AT5G60390	AT5G60390	AT5G60390	AAC1	CAB1
AAC1	AAC1	AAC1	AAC1	AT5G60390	AT5G60390
ATMS1	ATMS1	ATMS1	ATMS1	ATMS1	ATMS1
AT3G16420	**HSC70-1**	HSC70-1	HSC70-1	HSC70-1	HSC70-1
HSC70-1	**AT3G16420**	AT3G16420	AT3G16420	AT3G16420	AT3G16420
AT4G24150	**AT4G24150**	AT4G24150	HOG1	HOG1	HOG1

d. Top ten expressed gene list in KOC

Golden Std	KOSalt				
2X75	2x59	2x34	1x75	1x59	1x34
ICL	ICL	ICL	ICL	ICL	ICL
PYK10	PYK10	PYK10	PYK10	PYK10	PYK10
AT5G60390	AT5G60390	AT5G60390	AT5G60390	**AT3G02480**	**AT3G02480**
HSC70-1	**AT3G02480**	**AT3G02480**	HSC70-1	HSC70-1	HSC70-1
AT3G15670	**HSC70-1**	**HSC70-1**	**AT3G02480**	**AT5G60390**	**AT5G60390**
AT3G02480	**AT3G15670**	**AT3G15670**	**AT3G15670**	**AT3G15670**	**AT3G15670**
RBCS1A	RBCS1A	R3CS1A	RBCS1A	RBCS1A	RBCS1A
MLS	MLS	MLS	MLS	MLS	MLS
AT3G41768	AAC1	AAC1	AAC1	AAC1	AAC1
AAC1	PCK1	PCK1	PCK1	PCK1	PCK1

Overall, our analysis confirms that 2x59 read length exhibits a robust correlation with the golden standard dataset (2x75). It consistently captures highly expressed genes in various sample groups, making it a reliable choice for accurately representing gene expression patterns.

The slight variations observed in the other datasets highlight the importance of read length selection in RNA-seq experiments to ensure accurate and consistent gene expression analysis.

4.5 Comparison of Gene ontologies of DEG's with the golden standard 2x75 of HDC1compC, HDC1compSalt, Knocked out Control, and Salt:

To investigate the impact of differentially expressed (DE) genes between the golden standard (2x75) and datasets with varying read lengths on pathways and biological functions, we conducted gene ontology (GO) analysis. The significant DE genes were subjected to analysis using the tool Search Light2 to assess if 2x59 correlates with the golden standard. We compared the significant genes against various gene ontology and pathway analysis databases to identify enriched gene sets. The analysis was based on the hypergeometric test, considering p-values and log2 fold changes. Figure 7 displays the top 10 most enriched gene sets for each Salt versus Control comparison obtained with different read lengths. We further compared these enriched gene sets with the golden standard 2x75 for both wild-type and knockout genotypes. In the enrichment analysis results, the functions that are highlighted in yellow correspond to gene sets that match the golden standard, meaning they agree with the expected biological functions. On the other hand, the functions highlighted in blue are mismatched with the golden standard, indicating a discrepancy or lack of agreement between the enriched gene sets and the expected biological functions. We observed that 2x59 and 2x34 showed a strong correlation with the golden standard in all biological conditions. These read lengths consistently represented similar gene sets, indicating their reliability in capturing important biological pathways and functions. However, 1x75 exhibited some deviations in the analysis, suggesting discrepancies in the representation of enriched gene sets compared to the golden standard.

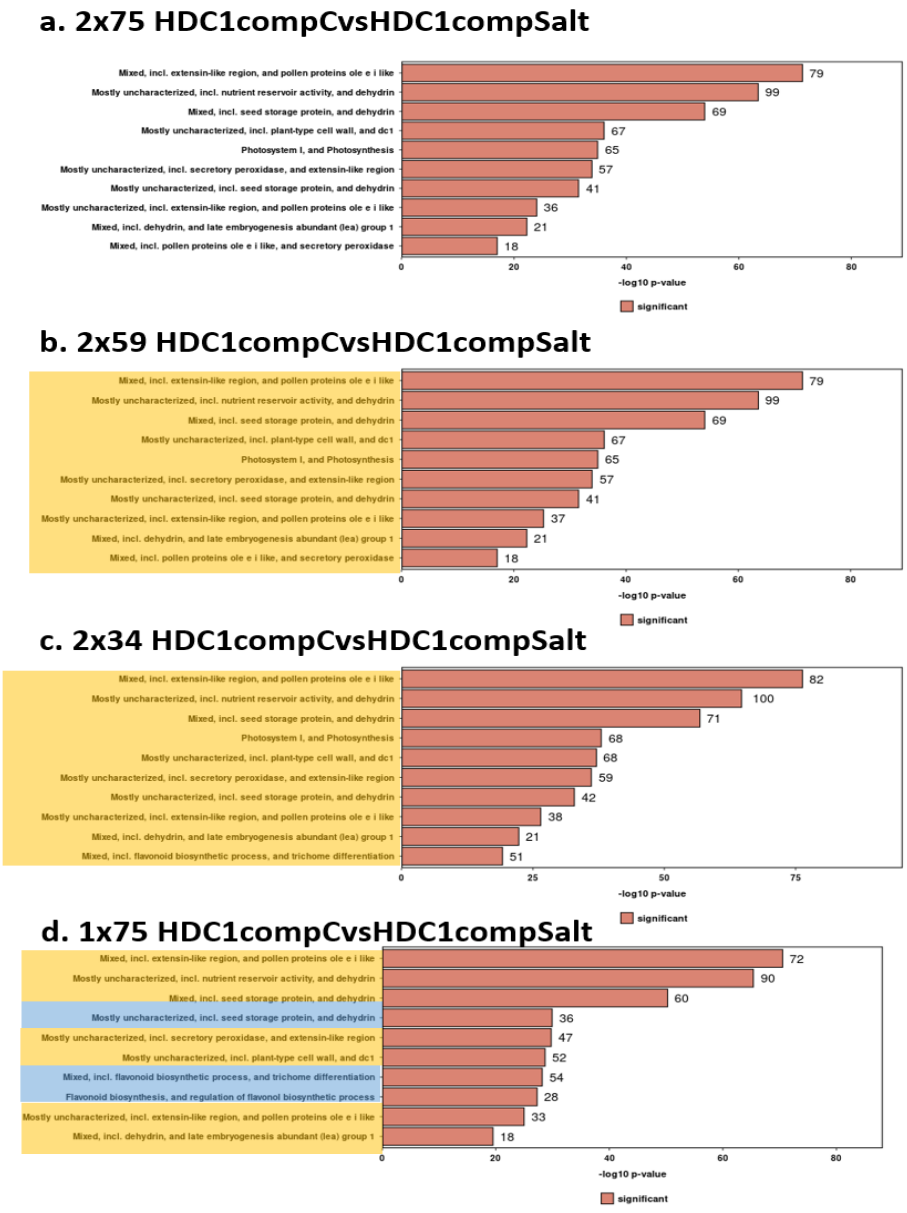


Figure 7: The top 10 significant enrichment gene sets identified in Salt versus Control comparison in the wild-type genotype (p-value). *The most significant gene enrichment analysis was conducted based on p-value >0.05 using string_11.5 for all HDC1compControlvsHDC1compSalt datasets. a. 2x75, the golden standard b. 2x59 c. 2x34 and d. 1x75. Each gene set is characterized by functional terms (on the left) and corresponding -log(p-values) and the size of the gene-set (on the right). When they were compared with the golden standard, 2x59 and 2x34 showed maximum identity, and 1x75 showed some variation. Those functions highlighted in yellow represent those that matched with the golden standard, and the blue highlighted are mismatched. And the same comparison for the Knockout condition is included in the appendix(A2).*

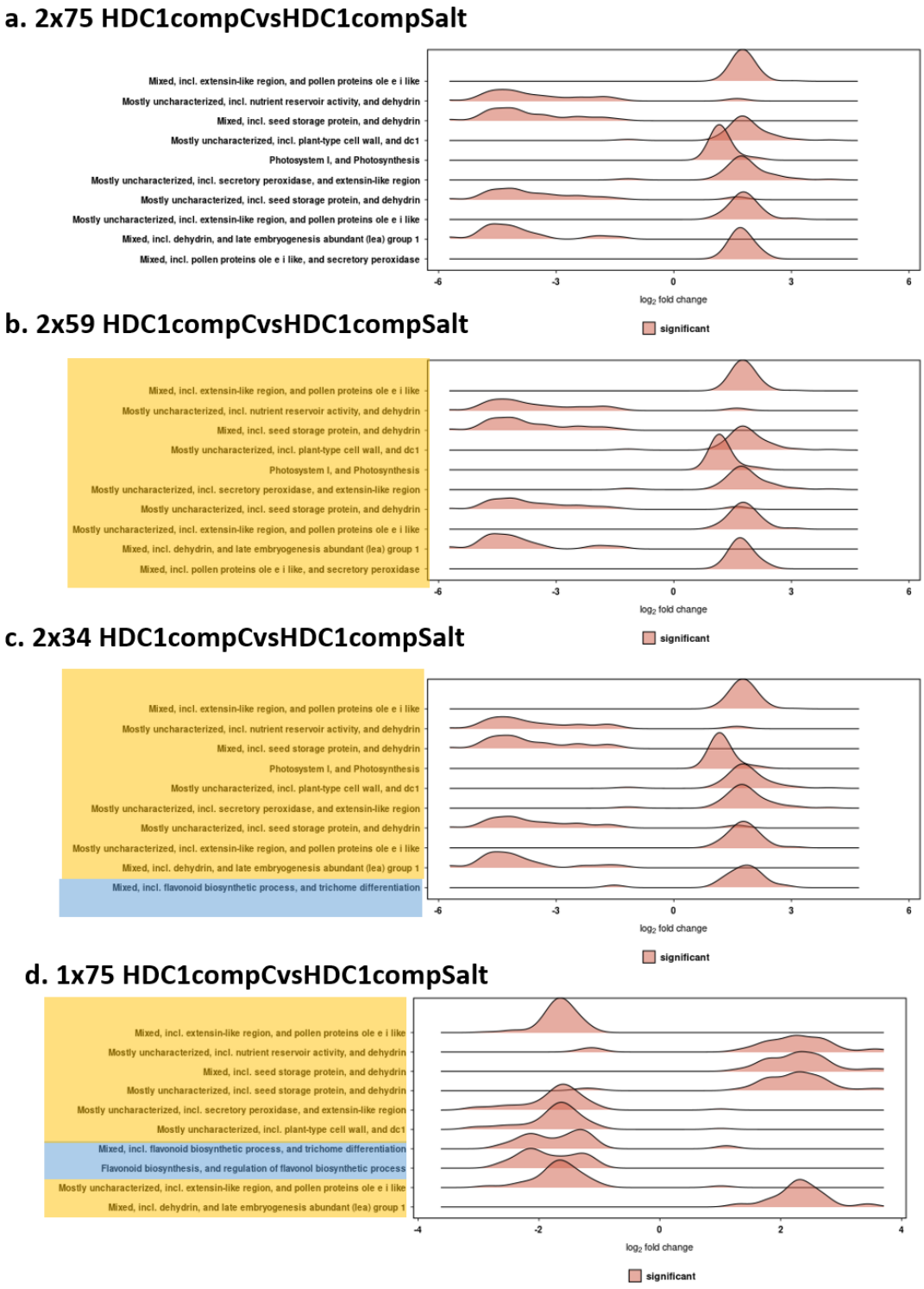


Figure 8: The top 10 significant enrichment gene sets identified in Salt versus Control comparison in the wild-type genotype (log fold change). *The most significant gene enrichment analysis was performed based on log fold change > 0.0 using string_11.5 database for all HDC1compControlvsHDC1compSalt datasets. a. 2x75, the golden standard b. 2x59 c. 2x34, and d. 1x75. When they were compared with the golden standard, 2x59 shows maximum identity, and 2x34 and 1x75 show some variation. And the same comparison for the Knockout condition is included in the appendix (A1).*

Overall, our gene ontology analysis supports the strong correlation of 2x59 and 2x34 read lengths with the golden standard, emphasizing their suitability for reliable gene expression analysis in capturing biologically relevant pathways and functions. Conversely, 1x75 showed deviations, underscoring the importance of selecting appropriate read lengths for accurate gene ontology analysis in RNA-seq experiments.

5. Discussion:

5.1 Correlation analysis of 2x59 and 2x34 with the golden standard:

In this study, we aimed to evaluate the performance of paired-end and single-ended reads of different read lengths namely, 75, 59, and 34 bp in various analyses, with a specific focus on comparing the results obtained from 2x59 reads to the gold standard of 2x75 reads. Our findings reveal that 2x59 read length demonstrates a remarkably close correlation to 2x75 across multiple analyses, including differential expression of genes (DEG), identification of the most expressed genes, and gene ontology. Based on the result in sections 4.3, 4.4, and 4.5, it becomes evident that 2x59 is the most accurate read length, followed by 2x34, for capturing gene expression patterns compared to the golden standard 2x75 in both wild-type and knocked-out comparisons.

Furthermore, 2x34 read length also shows notable correlations in these analyses. Shorter the read, the more chance of multiple mapping; therefore, one would expect that RNAseq with reads 34bp long would lead to inferior results and reduces the power to effectively detect genetic variants (Li, Ruan, and Durbin, 2008). These limitations could be overcome by increasing the sequencing depth, utilizing longer insert sizes, improving library preparation, and utilizing advanced and effective algorithms (McKernan et al., 2009)(Li *et al.*, 2008) .

In this study, we formulated our hypothesis based on past research, which explored a similar topic, and suggested that short paired-end reads could yield accurate results and offer cost-benefits when compared to longer read lengths in various analyses (Freedman *et al*.,2020). Drawing from their findings, we hypothesized that 2x59 read length would exhibit a close correlation with the gold standard of 2x75 read length in multiple analyses, including differential expression of genes, identification of the most expressed genes, and gene ontology. Additionally, we anticipated that 2x34 read length might also demonstrate notable correlations in these analyses, although potentially not as strong as 2x59.

Regarding differential expression analysis, we observed that the expression patterns derived from 2x59 reads are highly concordant with those obtained from 2x75 reads. The overlap in significantly differentially expressed genes between the two read lengths suggests that 2x59 adequately captures the major signals of differential expression, similar to the gold standard, 2x75. Furthermore, in our exploration of the most expressed genes, we found that 2x59 read length exhibits a strong agreement with 2x75. The shared top-expressed genes (Table 1) indicate that 2x59 effectively captures the highly expressed transcripts in a manner comparable to 2x75. Additionally, the gene ontology analyses (Figures 7 & 8) yielded similar results for both 2x59 and 2x75 read lengths. The enriched biological processes, molecular functions, and cellular components identified from 2x59 were highly consistent with those detected using 2x75. This robust correlation indicates that 2x59 is reliable in capturing the relevant biological functions, close to the gold standard.

Based on our comprehensive analyses, we suggest that 2x75 read length can be potentially replaced with 2x59 in gene expression studies without compromising accuracy and reliability as shared paired-end expressions are more reliable (Ringeling et al., 2022) and provide information on various versions of mRNA transcripts (Katz et al., 2010). Furthermore, our results also indicate that 2x34 read length exhibits substantial correlations in various analyses, making it another viable option, though not as close as 2x59. Given that shorter paired-end 2x59 reads display comparable performance to 2x75 while retaining a high level of accuracy, it becomes an attractive and cost-effective alternative for gene expression studies (Freedman *et al*.,2020). The reduced sequencing cost associated with 2x59 read length can significantly benefit researchers, especially when dealing with large-scale experiments or projects with budget constraints. Glasgow Polyomics charges £2.42, £3.15 and £4.06 for 1 million of PE reads 2x34, 2x59 and 2x100, respectively. Given that GP uses a NextSeq2000 sequencer running 2x100 reads routinely, the move to 2x59 or 2x34 would result in a 22% or 40% cost reduction, respectively. The saved resources could, in principle, be used to increase the number of replicates, which would lead to higher statistical power of DEG detection.

5.2 Correlation analysis of 1x75, 1x59, and 1x34 with the golden standard

Further, we aimed to compare the performance of different read lengths, specifically 1x75, 1x59, and 1x34, in various gene expression analyses. Our primary focus was on how these read lengths compared to the gold standard, 2x75, in terms of reliability and cost-effectiveness. Upon analyzing the results, we observed that 1x59 and 1x34 read lengths displayed some similarities with the gold standard in one or two analyses. However, these similarities were not consistent across all analyses, and intensive trimming like 1x34 can significantly influence the composition of expression estimates, leading us to question their reliability (Williams *et al*., 2016). As a result, we chose to disregard their similarities to the gold standard and focus on the more robust findings. In contrast, 1x75 read length showed noticeable deviations from the gold standard in most of the analyses. These deviations suggested that 1x75 may not provide accurate and reliable results comparable to the gold standard, making it a less suitable option for gene expression analysis. Single-end reads (1x75) showed more discrepancies, emphasizing the advantages of shorter read lengths (1x59 and 1x34) in achieving more consistent and accurate gene expression analysis and further emphasizing the importance of utilizing short paired-end read lengths for more reliable and consistent gene expression analysis.

5.3 Expression analysis of the wild-type and knockout datasets:

In our study, we noticed that DEG for HDC1comp in control and salt condition were less compared to that of knock out of HDC1comp in both conditions. The knockout of HDC1comp in control and salt conditions likely induces a cascade of gene expression changes, affecting multiple pathways and regulatory networks. These changes contribute to a higher number of DEGs detected in the RNA-seq data. Understanding the underlying molecular mechanisms and the specific genes affected by HDC1comp KO is essential to gain insights into its biological function and the cellular response to salt stress.

In conclusion, our study provides compelling evidence in support of shorter paired-end 2x59 reads as the better option for gene expression analysis. We emphasize the importance of considering both reliability and cost-effectiveness when choosing the appropriate read length for specific research objectives.

5.4 Limitations and Future Works:

1. This work was only done on one data set. Although we mitigated this shortcoming by analysing two genotypes with distinctly different DEG numbers, the data still come from one experiment.

2. This study was done on data from one model organism Arabidopsis thaliana which has medium size eukaryotic genome. The results, therefore, are limited to this organism or perhaps to other organisms with similar genome sizes. It is likely that for large genomes, e.g., human or mouse, some of the results may not hold. In particular, we would expect that the difference between the golden standard and 2x34 could possibly be deeper as for large genomes, there is more chance of multiple mapping of 34bp long read.

3. Consequently, in an ideal scenario, one would run this study for a number of different datasets and for a number of model organisms of different genome sizes, but that would go beyond the timeline of this project.

6. Acknowledgments

I would like to express my sincere gratitude to my Supervisor, Dr. Pawel Herzyk, for this mentorship and expertise. His insightful guidance, continuous support, and dedication have played an important role in shaping the direction of this project and enhancing its quality.

I extend my deepest appreciation to Dr. John Cole and Dr. David Mcguinness for their invaluable support throughout the project. Lastly, I would like to express my heartfelt gratitude to my family members, well-wishers and friends for their support and prayers throughout this endeavor. Their encouragement and trust in my abilities have been my driving force.

The successful completion of this project owes much to the combined efforts of everyone mentioned above. I deeply appreciate their valuable contributions.

7. References:

1. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

2. Bolger, A.M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114–2120. doi:https://doi.org/10.1093/bioinformatics/btu170.

3. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int J Mol Sci. 2017 Jul 29;18(8):1652. doi: 10.3390/ijms18081652. PMID: 28758927; PMCID: PMC5578042.

4. Cole, J., Faydaci, B.A., McGuinness, D., Shaw, R., Maciewicz, R.A., Robertson, N. and Goodyear, C.S. (2021). Searchlight: automated bulk RNA-seq exploration and visualisation using dynamically generated R scripts. 22(1). doi:https://doi.org/10.1186/s12859-021-04321-2.

5. Freedman, A.H., Gaspar, J.M. & Sackton, T.B. Short paired-end reads trump long single-end reads for expression analysis. *BMC Bioinformatics* **21**, 149 (2020). https://doi.org/10.1186/s12859-020-3484-z

6. Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D. and Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology*, 13(1). doi:https://doi.org/10.1186/s13045-020-01005-x.

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

8. Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), pp.907–915. doi:https://doi.org/10.1038/s41587-019-0201-4.

9. Larsson J (2022). *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses*. R package version 7.0.0, https://CRAN.R-project.org/package=eulerr.

10. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008 Nov;18(11):1851-8. doi: 10.1101/gr.078212.108. Epub 2008 Aug 19. PMID: 18714091; PMCID: PMC2577856.

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

12. Love, M.I., Huber, W. and Anders, S. (2014). Love MI, Huber W, Anders S.. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol 15: 550.

13. McCombie WR, McPherson JD. Future Promises and Concerns of Ubiquitous Next-Generation Sequencing. Cold Spring Harb Perspect Med. 2019 Sep 3;9(9):a025783. doi: 10.1101/cshperspect.a025783. PMID: 30478095; PMCID: PMC6719590.

14. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 2009 Sep;19(9):1527-41. doi: 10.1101/gr.091868.109. Epub 2009 Jun 22. PMID: 19546169; PMCID: PMC2752135.

15. Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html

16. R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

17. RStudio Team. 2021. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

18. Ringeling, F.R., Chakraborty, S., Vissers, C., Reiman, D., Patel, A.M., Lee, K.-H., Hong, A., Park, C.-W., Reska, T., Gagneur, J., Chang, H., Spletter, M.L., Yoon, K.-J., Ming, G., Song, H. and Canzar, S. (2022). Partitioning RNAs by length improves transcriptome reconstruction from short-read RNA-seq data. *Nature Biotechnology*, [online] 40(5), pp.741–750. doi:https://doi.org/10.1038/s41587-021-01136-7.

19. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, doi:10.1093/nar/gkac247

20. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics. 2016 Feb 25;17:103. doi: 10.1186/s12859-016-0956-2. PMID: 26911985; PMCID: PMC4766705.

8. Appendix

Table A1. The Top ten most DEG in HDC1compCvsHDC1compSalt in 2x75.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT5G52300	13243	-5.68007	1.70E-49	LTI65
AT2G36270	2377.337	-3.35667	5.13E-46	ABI5
AT1G15330	7131.398	-3.69142	8.81E-45	AT1G15330
AT3G53040	3159.103	-5.72172	3.33E-44	AT3G53040
AT2G35300	14844.43	-4.3436	3.33E-44	LEA18
AT3G02480	65851.71	-3.54129	2.25E-42	AT3G02480
AT4G21020	10452.59	-4.99766	4.27E-42	AT4G21020
AT3G50980	2256.366	-4.12529	2.32E-41	XERO1
AT2G36640	7298.376	-5.46917	9.68E-39	ECP63
AT5G44310	8422.031	-5.04058	1.11E-38	AT5G44310

Table A2. Top ten most differentially expressed genes in KOCvsKOSalt in 2x75.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT4G22880	5851.702	5.011225	8.50E-99	LDOX
AT5G42800	6485.039	6.03591	5.41E-95	DFR
AT5G44440	1247.546	6.044941	5.76E-77	AT5G44440
AT3G01500	4713.453	6.659926	2.16E-69	CA1
AT1G02930	9767.326	4.616743	6.58E-64	GSTF6
AT5G54060	1714.324	4.803509	1.38E-62	UF3GT
AT4G14090	1699.443	4.354248	4.75E-62	AT4G14090
AT3G12960	268.5487	-4.73415	4.16E-56	SMP1
AT4G08870	1057.67	6.272248	4.16E-56	ARGAH2
AT1G48470	2991.876	-3.46088	2.48E-55	GLN1

Table A3. Top ten most differentially expressed genes in HDC1compCvsHDC1compSalt in 2x59.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT5G52300	13280.2	-5.6854	1.56E-49	LTI65
AT2G36270	2370.477	-3.3543	6.37E-46	ABI5
AT1G15330	7128.355	-3.691	1.08E-44	AT1G15330
AT3G02480	74163.46	-3.58606	3.79E-44	AT3G02480
AT3G53040	3157.701	-5.71979	3.79E-44	AT3G53040
AT2G35300	14828.71	-4.34133	3.79E-44	LEA18
AT4G21020	10454.4	-4.99422	6.58E-42	AT4G21020
AT3G50980	2268.624	-4.13824	3.09E-41	XERO1
AT2G36640	7291.311	-5.46121	2.06E-38	ECP63
AT5G44310	8434.069	-5.03254	3.52E-38	AT5G44310

Table A4. Top ten most differentially expressed genes in KOCvsKOSalt in 2x59.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT4G22880	5723.614	5.009788	7.70E-101	LDOX
AT5G42800	6481.544	6.045495	2.74E-96	DFR
AT5G44440	1247.413	6.043458	8.64E-77	AT5G44440
AT3G01500	4718.153	6.663448	2.75E-69	CA1
AT1G02930	9380.463	4.632491	2.14E-65	GSTF6
AT4G14090	1737.325	4.351073	5.09E-62	AT4G14090
AT5G54060	1692.533	4.784687	1.34E-61	UF3GT
AT4G08870	1057.797	6.270682	5.17E-56	ARGAH2
AT1G48470	2995.948	-3.45728	3.10E-54	GLN1
AT3G50980	2268.624	-4.53539	1.19E-53	XERO1

Table A5. Top ten most differentially expressed genes in HDC1compCvsHDC1compSalt in 2x34.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT5G52300	13172.9306	-5.68622	2.77E-49	LTI65
AT2G36270	2370.34665	-3.35022	5.41E-45	ABI5
AT1G15330	7112.62675	-3.68846	1.28E-44	AT1G15330
AT3G53040	3150.43739	-5.71821	1.76E-44	AT3G53040
AT2G35300	14742.7673	-4.34845	4.04E-44	LEA18
AT3G02480	74459.3831	-3.58711	1.64E-43	AT3G02480
AT4G21020	10459.8071	-4.99694	2.59E-42	AT4G21020
AT3G50980	2261.73898	-4.14794	1.23E-41	XERO1
AT2G36640	7324.49048	-5.46386	2.89E-38	ECP63
AT5G44310	8749.12512	-5.02346	5.38E-37	AT5G44310

Table A6. Top ten most differentially expressed genes in KOCvsKOSalt in 2x34.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT4G22880	5513.62579	5.00873	1.19E-101	LDOX
AT5G42800	6444.01708	6.047483	2.10E-95	DFR
AT5G44440	1217.10262	6.045459	2.74E-75	AT5G44440
AT1G02930	8824.58883	4.611476	1.47E-68	GSTF6
AT3G01500	4705.86773	6.650145	1.47E-68	CA1
AT5G54060	1598.65331	4.819757	3.14E-63	UF3GT
AT4G14090	1769.06852	4.355458	2.54E-61	AT4G14090
AT1G48470	3020.14491	-3.4776	1.50E-55	GLN1
AT4G08870	1031.79646	6.321437	1.50E-55	ARGAH2
AT3G50980	2261.73898	-4.53978	4.44E-54	XERO1

Table A7. Top ten most differentially expressed genes in HDC1compCvsHDC1compSalt in 1x75.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT4G25100	16920.19	-5.03785	2.94E-84	FSD1
AT5G08450	6064.232	3.374484	3.82E-71	AT5G08450
AT2G28190	4288.681	2.520412	9.88E-29	CSD2
AT5G52300	6873.504	4.148683	4.95E-26	LTI65
AT1G12520	1111.391	2.138852	2.24E-22	CCS
AT5G38940	536.559	-4.36592	2.54E-21	AT5G38940
AT1G15330	3729.921	2.528379	1.49E-20	AT1G15330
AT4G22880	3023.588	-2.18245	1.89E-19	LDOX
AT4G36700	923.4555	3.556974	3.12E-19	AT4G36700
AT5G42800	3367.565	-2.61824	1.04E-18	DFR

Table A8. Top ten most differentially expressed genes in KOCvsKOSalt in 1x75.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT3G01500	2485.398	-9.15522	4.29E-128	CA1
AT3G02480	37127.96	5.851912	7.89E-119	AT3G02480
AT1G48470	1562.798	4.629201	2.71E-102	GLN1
AT5G42800	3367.565	-6.15155	1.01E-101	DFR
AT3G50980	1173.103	6.509274	1.16E-99	XERO1
AT5G08450	6064.232	-3.93515	2.30E-97	AT5G08450
AT4G22880	3023.588	-4.86198	5.77E-97	LDOX
AT5G44440	647.6369	-7.18016	2.93E-91	AT5G44440
AT2G36270	1234.281	4.603472	3.49E-88	ABI5
AT1G15330	3729.921	4.968986	1.78E-83	AT1G15330

Table A9. Top ten most differentially expressed genes in HDC1compCvsHDC1compSalt in 1x59.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT5G52300	6868.895	-5.64938	9.33E-49	LTI65
AT1G15330	3727.822	-3.69878	8.29E-45	AT1G15330
AT2G36270	1227.927	-3.32619	8.29E-45	ABI5
AT2G35300	7683.623	-4.33474	1.52E-44	LEA18
AT3G02480	38943.05	-3.59163	2.54E-44	AT3G02480
AT3G53040	1621.447	-5.65136	1.30E-43	AT3G53040
AT4G21020	5386.909	-4.95264	5.41E-41	AT4G21020
AT2G36640	3724.84	-5.43882	1.76E-39	ECP63
AT5G44310	4315.298	-5.01075	1.50E-38	AT5G44310
AT3G50980	1173.98	-4.0523	1.56E-37	XERO1

Table A10. Top ten most differentially expressed genes in KOCvsKOSalt in 1x59.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT4G22880	2958.147	4.99566	2.87E-103	LDOX
AT5G42800	3362.467	6.013686	1.94E-98	DFR
AT1G02930	4910.562	4.630316	2.10E-67	GSTF6
AT3G01500	2482.526	6.601278	9.18E-66	CA1
AT4G14090	888.369	4.342363	1.21E-61	AT4G14090
AT5G44440	647.3851	5.896296	1.50E-60	AT5G44440
AT5G54060	885.9694	4.777822	3.55E-60	UF3GT
AT1G48470	1563.178	-3.48499	1.42E-59	GLN1
AT3G29590	688.5362	4.898676	1.32E-52	AT5MAT
AT3G50980	1173.98	-4.47634	2.58E-52	XERO1

Table A11. Top ten most differentially expressed genes in HDC1compCvsHDC1compSalt in 1x59.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT5G52300	6762.70353	-5.6531	4.16E-48	LTI65
AT1G15330	3703.97805	-3.6979	4.38E-45	AT1G15330
AT2G35300	7606.56701	-4.3505	4.38E-45	LEA18
AT3G02480	38404.8531	-3.59141	3.95E-44	AT3G02480
AT2G36270	1218.68655	-3.31905	9.41E-44	ABI5
AT3G53040	1609.19303	-5.61499	1.59E-42	AT3G53040
AT4G21020	5386.81796	-4.95932	1.13E-41	AT4G21020
AT2G36640	3721.64279	-5.44768	9.25E-40	ECP63
AT3G01500	2407.05795	4.098962	1.89E-38	CA1
AT3G50980	1154.83874	-4.08478	2.18E-38	XERO1

Table A12. Top ten most differentially expressed genes in KOCvsKOSalt in 1x34.

Gene_ID	Base mean	log2(FC)	P-adj	Gene_name
AT4G22880	2840.6673	5.011307	3.59E-108	LDOX
AT5G42800	3239.7944	6.050011	2.17E-100	DFR
AT3G01500	2407.05795	6.727882	3.63E-70	CA1
AT1G02930	4634.97744	4.623108	2.13E-68	GSTF6
AT5G54060	835.259576	4.795139	3.87E-60	UF3GT
AT4G14090	903.158208	4.326424	3.87E-60	AT4G14090
AT1G48470	1540.70938	-3.4884	7.49E-60	GLN1
AT5G44440	629.657036	5.858092	3.69E-59	AT5G44440
AT3G50980	1154.83874	-4.46586	1.09E-52	XERO1
AT3G29590	687.189508	4.894928	1.50E-52	AT5MAT

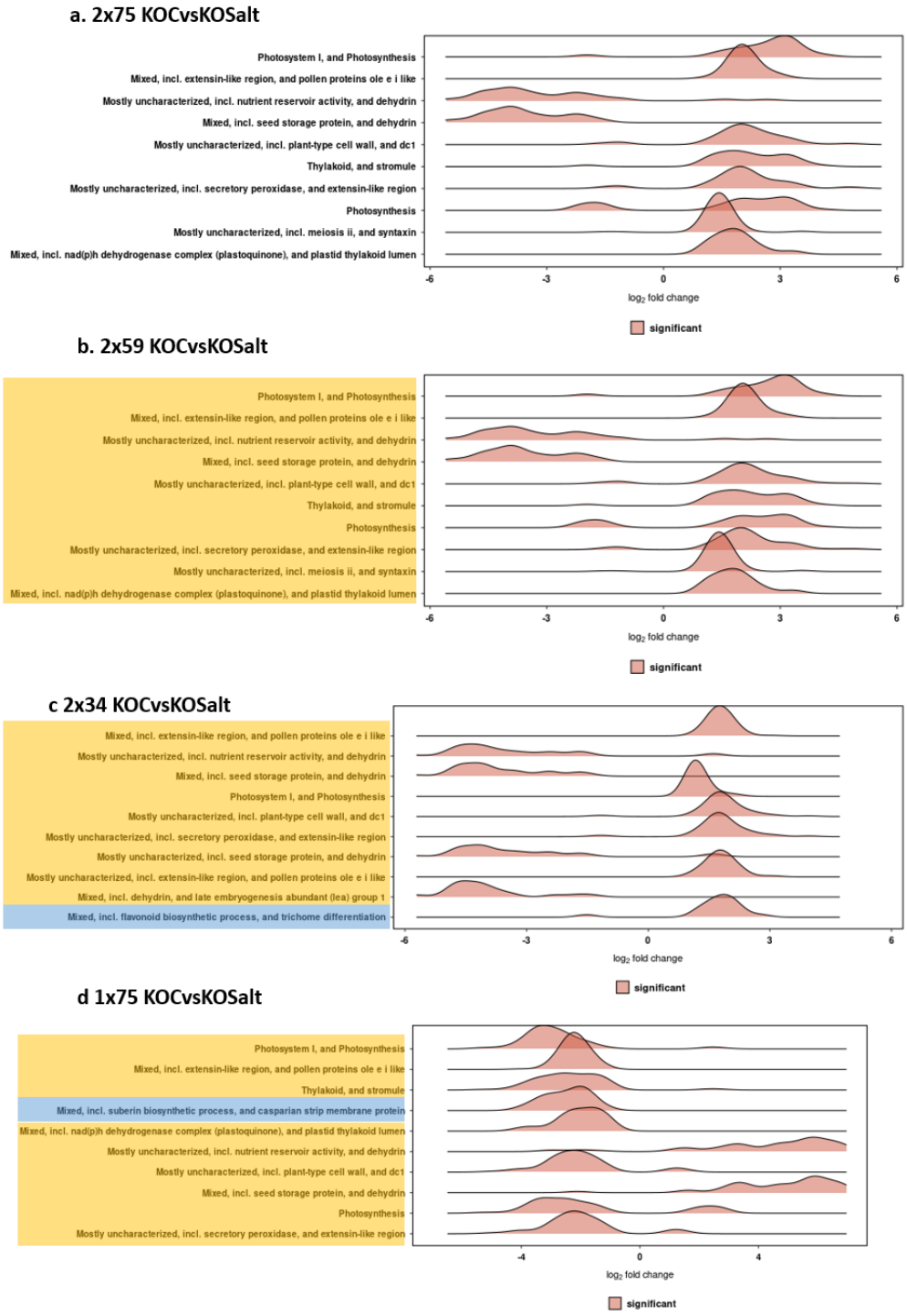


Figure A1: ***The top 10 significant gene enrichment analysis on Knockout comparison (log fold change).*** *The most significant gene enrichment analysis was performed based on log fold change >0.05 using string_11.5 for all HDC1compControlvsHDC1compSalt datasets. a. 2x75 ,the golden standard b. 2x59 c. 2x34 and d. 1x75. When they were compared with the golden standard 2x59 and 2x34 shows maximum identity and 1x75 shows some variation. Those functions highlighted in yellow represents those that matched with the golden standard and blue highlighted are mismatched .*

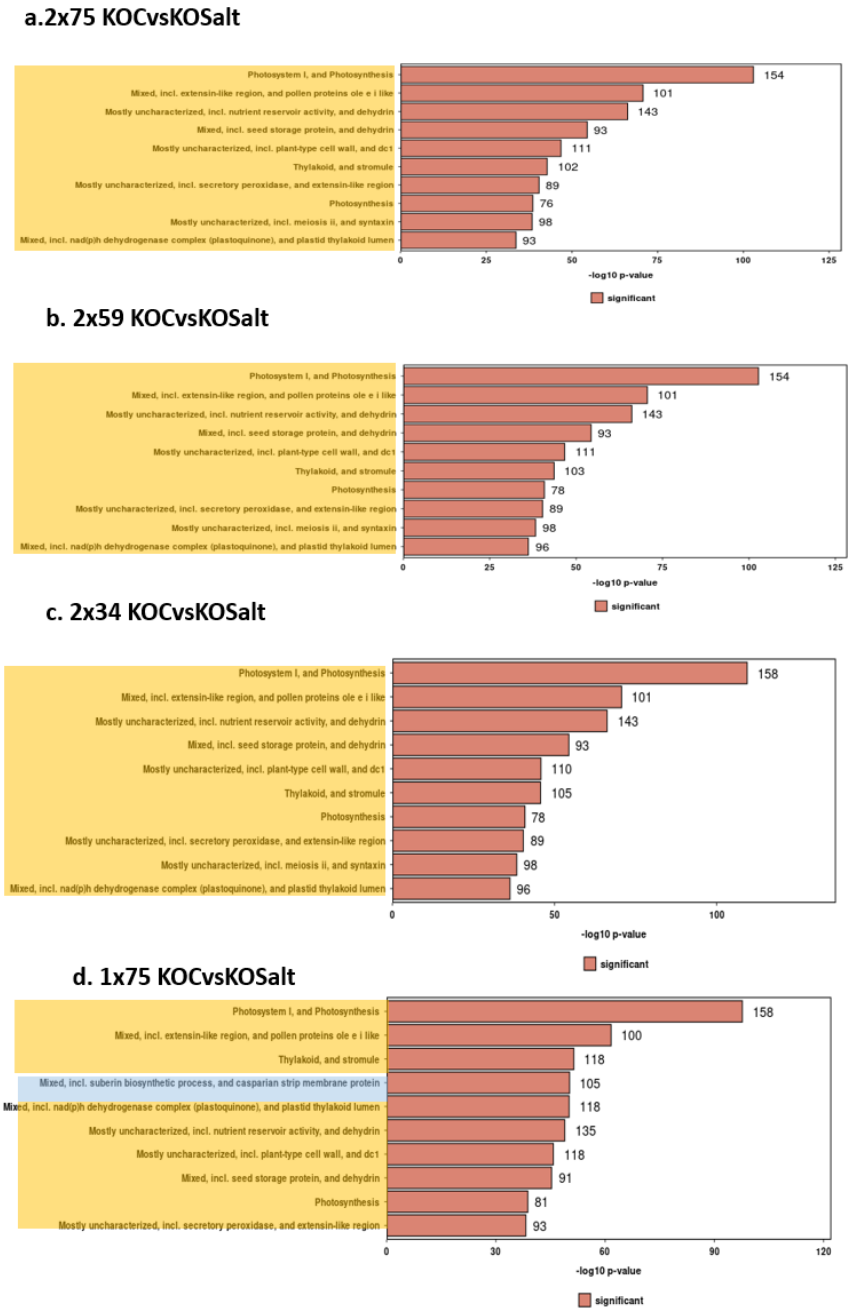


Figure A2: ***The top 10 significant gene enrichment analysis on Knockout comparison (pvalue).*** *The most significant gene enrichment analysis was performed based on pvalue >0.05 using string_11.5 for all HDC1compControlvsHDC1compSalt datasets. a. 2x75 ,the golden standard b. 2x59 c. 2x34 and d. 1x75. Each gene-set is characterised by functional terms (on the left) and corresponding -log(p-values) and the size of the gene-set (on the right). When they were compared with the golden standard 2x59 and 2x34 shows maximum identity and 1x75 shows some variation. Those functions highlighted in yellow represents those that matched with the golden standard and blue highlighted are mismatched .*