Statistical Analysis of Microarray Data to Identify Key Gene Expression Patterns in Primary Hyperoxaluria

Usha Adiga[1], Banubadi Anil Kishore[1], P. Supriya[1], Alfred J. Augustine[2] and Sampara Vasishta[1,*]

[1]*Department of Biochemistry, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India*

[2]*Department of Surgery, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India*

Abstract: This study aims to utilize microarray data deposited by Romero *et al*. and conduct bioinformatic analysis for identifying differentially expressed genes (*DEGs*) associated with a novel method involving gene correction at the Alanine–Glyoxylate Aminotransferase (*AGXT*) locus and direct conversion of fibroblasts from primary hyperoxaluria type 1 (*PH1*) patients into healthy induced hepatocytes (*iHeps*) using Clustered Regularly Interspaced Short Palindromic Repeats - CRISPR-associated protein 9 (*CRISPR-Cas9*) technology. Additionally, the study aims to elucidate the molecular mechanisms underlying hyperoxaluria compared to oxalate crystal formation. Romero *et al.'s* GSE226019 microarray data was retrieved from Gene Expression Omnibus. Statistical analysis was done in R and Bioconductor, utilizing rigorous methods to ensure robust and reproducible results. The limma program compared gene expression levels across groups. Pathway analysis, protein-protein interaction (PPI) network creation, and miRNA-target interaction network analysis were constructed. The top ten *DEGs* included *ANGPTL3*, *SLC38A3*, *KNG1*, *BDH1*, GC, *ADH1C*, *ARG1*, *CYP3A4*, *AMBP*, and *CYP2C9*. Enrichment analysis revealed significant associations with various biological pathways, including Linoleic acid metabolism and Retinol metabolism. Volcano plots and mean difference plots highlighted significant gene expression changes between different sample groups. Protein-protein interaction networks and miRNA-target interaction networks provided insights into molecular interactions and regulatory mechanisms. The top ten differentially expressed genes include *ANGPTL3*, *SLC38A3*, *KNG1*, *BDH1*, GC, *ADH1C*, *ARG1*, *CYP3A4*, *AMBP*, and *CYP2C9*—emerge as key players with strong associations to critical biological pathways like Linoleic acid metabolism and drug metabolism-cytochrome P450. Understanding the regulatory role of specific miRNAs (hsa-miR-4501, hsa-miR-5692c, hsa-miR-6731-3p, hsa-miR-6867-5p, hsa-miR-616-3p, hsa-miR-4468, hsa-miR-3692-3p, hsa-miR-4277, hsa-miR-4763-5p, hsa-miR-4797-5p) in gene expression could provide further insights into disease mechanisms and potential therapeutic avenues. The statistical findings provide a foundation for predictive modeling, hypothesis testing, and exploring personalized therapeutic strategies.

Keywords: Primary Hyperoxaluria, Gene Expression Analysis, Molecular Mechanisms, Differential Gene Expression, Bioinformatics, Statistical Analysis, Data Visualization.

INTRODUCTION

Primary hyperoxalurias (PH) encompass rare metabolic disorders characterized by hepatic oxalate overproduction, with primary hyperoxaluria type 1 (*PH1*) standing out as the most severe form. It arises due to a deficiency in alanine-glyoxylate aminotransferase (AGT), a peroxisomal enzyme encoded by the *AGXT* gene, resulting in impaired conversion of glyoxylate to glycine in the liver. This metabolic anomaly leads to the accumulation of oxalate, a metabolite that cannot be efficiently processed by mammals and consequently precipitates as calcium oxalate crystals, primarily in the kidneys, resulting in renal damage and often progressing to end-stage renal disease (ESRD) and systemic oxalosis [1].

Historically, treatment modalities have focused on preserving renal function and reducing urinary calcium oxalate saturation. However, the advent of substrate reduction therapy, involving periodic administration of therapeutic siRNA targeting the upstream enzyme glycolate oxidase (GO), has provided a novel therapeutic avenue. Despite these advancements, patients often require intensive dialysis as a temporary measure, awaiting combined liver-kidney transplantation (LKTx), the only definitive curative treatment due to the hepatic origin of oxalate overproduction [2-6].

Although Liver-Kidney Transplantation (LKTx) remains the primary treatment, challenges such as donor organ shortage, transplantation-associated complications, and lifelong immunosuppression underscore the need for alternative therapeutic strategies. Gene editing has emerged as a promising avenue, with targeted correction of *AGXT* mutations using *CRISPR-Cas9* technology showing potential. Additionally, hepatocyte transplantation has shown promise in preclinical studies and as a bridging procedure in clinical cases. Efforts to generate induced hepatocytes from patient-derived cells offer an attractive alternative for liver cell replacement therapy,

*Address correspondence to this author at the Department of Biochemistry, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India; Tel: (+91) 8939143993;
E-mail: vasishta94525@gmail.com

E-ISSN: 1929-6029/24

minimizing tumorigenic risks associated with pluripotent cell-based approaches.

Romero *et al*. explored innovative strategies to address *PH1*, combining gene correction at the *AGXT* locus with direct conversion of patient-derived fibroblasts into healthy induced hepatocytes (*iHeps*) [7]. They employed *CRISPR-Cas9* technology for site-specific *AGXT* correction through homology-directed repair (HDR), utilizing two distinct methods: precise point mutation correction and insertion of an enhanced *AGXT* cDNA. Subsequently, they induced the transformation of corrected cells into *iHeps* by overexpressing hepatic transcription factors. The resulting *AGXT*-corrected *iHeps* exhibited hepatic gene expression profiles and demonstrated reduced oxalate accumulation compared to non-edited *PH1*-derived *iHeps*, presenting a promising alternative cellular source for liver cell replacement therapy and a personalized *in vitro* model for studying *PH1* [7].

Furthermore, advancements in hepatocyte differentiation from induced pluripotent stem cells (iPSCs) and direct cell reprogramming of somatic cells into induced hepatocytes offer additional avenues for exploring autologous hepatocyte-based therapies, minimizing the risks associated with allogenic transplantation and pluripotent cell-derived tumorigenicity. These approaches hold immense potential in addressing the unmet clinical needs of *PH1* patients, paving the way for effective and personalized therapeutic interventions.

The aim of the study is to utilize microarray data, deposited by the author Romero *et al.*, for bioinformatic analysis to identify differentially expressed genes (*DEGs*) linked with a novel method combining gene correction at the *AGXT* locus with direct conversion of fibroblasts from *PH1* patients into healthy induced hepatocytes (*iHeps*) using *CRISPR-Cas9* technology.

Additionally, the study intends to investigate gene enrichment, gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway involvement, protein-protein interaction (PPI) networks, and miRNA-target interaction networks to elucidate the molecular mechanisms underlying hyperoxaluria compared to the oxalate crystal formation stage. Statistical analysis played a pivotal role, employing linear models in the limma package, which are tailored for microarray data due to their flexibility and precision in handling high-dimensional datasets. These methods contributed to identifying key hub genes and pathways, with implications in predictive modeling, hypothesis testing, and personalized medicine.

METHODOLOGY

Statistical and Bioinformatics Analysis

This study's core novelty lies in the biostatistical and bioinformatics methodologies applied to analyze publicly accessible microarray data (accession number GSE226019) deposited by Romero *et al*. Statistical analyses were conducted using R 4.0.1 in conjunction with Bioconductor packages to ensure reproducibility and robustness. The limma package was employed for differential gene expression analysis, offering tools for calibration and standardization across experimental batches. A gene linear model was fitted to assess differential expression, with genes filtered based on fold-change thresholds and statistical significance ($p < 0.05$).

Visualization of differentially expressed genes (*DEGs*) was achieved through ggplot2 for volcano plots and pheatmap for clustering significant *DEGs*.

Dataset Selection and Preprocessing

Secondary data analysis was performed on microarray datasets obtained from the GEO database for their relevance to Primary Hyperoxaluria Type 1 (*PH1*) and *CRISPR-Cas9*-mediated gene repair experiments. The dataset included five groups:

Skin fibroblasts of healthy controls.

Skin fibroblasts with *AGXT* knock-in.

Induced hepatocytes (*iHeps*) of control donors.

Liver hepatocytes of control donors.

iHeps of *PH1* patients with *AGXT* mutation.

Raw data from GEO was preprocessed for background correction, normalization, and quality assessment using limma, ensuring consistency across sample sets.

Functional and Pathway Enrichment Analysis

To evaluate the biological functions and pathway enrichments of *DEGs*, the ClusterProfiler package was employed for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. Signaling pathways and biological functions were analyzed, with a statistically significant threshold set at $p < 0.05$. GO analysis provided insights into

cellular components, biological processes, and molecular functions, while KEGG analyses illuminated relevant pathways.

Protein-Protein Interaction (PPI) Network Analysis

A PPI network was constructed using the STRING database and visualized in Cytoscape to predict protein functions and elucidate cellular mechanisms. The cytoHubba module identified hub genes with higher connectivity scores, indicating their critical roles in the PPI network. Validation of hub gene statistical significance was conducted using GEO2R, with a focus on $p < 0.05$ for cross-sample comparisons.

MicroRNA-Target Gene Network

The miRNA-target gene relationships for *DEGs* and hub genes were identified using the miRDB database. This database integrates curated metadata, experimental evidence, and computational predictions. Interactions between miRNAs and target genes were systematically explored to understand their regulatory roles in *PH1*-related pathways.

Experimental Details and Data Sources

Data was derived from microarray experiments conducted by Romero *et al*. Human fibroblasts were isolated, immortalized via hTERT lentiviral transduction, and subjected to *CRISPR-Cas9*-mediated *AGXT* gene editing. Subsequent reprogramming of fibroblasts into *iHeps* involved hepatic transcription factors and culture in hepatic-specific media. Experimental validation included flow cytometry, RNA sequencing, immunoprecipitation, and western blotting. The processed data deposited in the GEO database (accession number GSE226019) served as the foundation for our analysis.

Ethical Considerations

This study utilized publicly available microarray data, eliminating ethical concerns. No direct patient recruitment or personal data usage occurred. Ethical standards adhered to those established by the original data authors.

RESULTS

The top ten differentially expressed genes were *ANGPTL3, SLC38A3, KNG1, BDH1, GC, ADH1C, ARG1, CYP3A4, AMBP, CYP2C9.* The table presents the enrichment results for the top 10 Differentially Expressed Genes (*DEGs*) in the study. These *DEGs* are significantly associated with various biological pathways, as indicated by the low False Discovery Rate (FDR) values, suggesting strong statistical significance. Notably, pathways such as Linoleic acid metabolism, Arginine biosynthesis, Retinol metabolism, and drug metabolism-cytochrome P450 exhibit high fold enrichment values, indicating a substantial overrepresentation of *DEGs* compared to random expectation. This suggests their potential importance in the context of the studied condition. Additionally, the involvement of pathways like Metabolic pathways and Chemical carcinogenesis-DNA adducts highlights the broader biological processes underlying the differential gene expression observed in the study. Overall, these findings provide valuable insights into the molecular mechanisms and pathways associated with the identified *DEGs*, offering a deeper understanding of the underlying biology of the studied condition.

A volcano plot demonstrates statistical significance ($-\log_{10}$ P value) versus magnitude of change ($\log_2$ fold change) and is useful for envisioning differentially expressed genes. GSE226019 were used to screen *DEGs* (upregulated and Generated using limma).

Volcano Plots-Group 1 vs Group 2

In the context of primary hyperoxaluria, comparing gene expression between Group 1 (skin fibroblasts of healthy controls) and Group 2 (skin fibroblasts with *AGXT* targeted knock-in) revealed noteworthy

Table 1: Enrichment Results of Top 10 *DEGs*

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
$5.1E^{-04}$	2	30	169.5	Linoleic acid metabolism
$4.2E^{-02}$	1	22	115.6	Arginine biosynthesis
$3.2E^{-05}$	3	68	112.2	Retinol metabolism
$3.2E^{-05}$	3	69	110.5	Drug metabolism-cytochrome P450
$3.2E^{-05}$	3	75	101.7	Metabolism of xenobiotics by cytochrome P450
2.1^{-03}	2	68	74.8	Chemical carcinogenesis-DNA adducts
1.1^{-02}	4	1538	6.6	Metabolic pathways

alterations in various gene expressions. Specifically, HES5 displayed a significant upregulation with a $\log_2$(fold change) of 6.938 and a corresponding -$\log_{10}$(Pvalue) of 1.354, suggesting its potential relevance to the condition. Conversely, genes like *PRDM16* and *EPHA2* exhibited substantial downregulation, with $\log_2$(fold changes) of -6.376 and -4.381, respectively, implicating their involvement in the pathophysiology. Moreover, FBXO2 and TMEM51 showed notable upregulation with $\log_2$(fold changes) of 4.676 and 9.021, respectively, whereas WNT4 and LOC105376845 displayed significant downregulation with $\log_2$(fold changes) of -5.904 and -3.894, respectively.

Volcano Plots-Group 3 vs Group 4

In comparing Group 3 (induced hepatocytes of control donors) to Group 4 (liver hepatocytes of control donors), significant changes in gene expression were observed. Notably, MXRA8 exhibited the most substantial upregulation, with a $\log_2$(fold change) of 6.156 and a corresponding -$\log_{10}$(Pvalue) of 13.198, suggesting its potential importance in the studied condition. Conversely, genes such as ARHGEF16 and ESPN displayed significant downregulation, with $\log_2$(fold changes) of -6.418 and -9.725, respectively, implicating them in the pathogenesis of the disease. Additionally, genes like GNB1 and ICMT showed notable upregulation with $\log_2$(fold changes) of 1.449 and 2.283, respectively, indicating their potential involvement in disease mechanisms. Conversely, TNFRSF14 and PLEKHG5 exhibited moderate changes in expression levels, with $\log_2$(fold changes) of -2.088 and 2.608, respectively.

Volcano Plots- Group 3 vs Group 5

In comparing Group 3 (induced hepatocytes of control donors) to Group 5 (induced hepatocytes of *PH1* patients with *AGXT* mutation), the analysis of gene expression data revealed significant changes in several genes. Notably, AJAP1 exhibited substantial upregulation with a $\log_2$(fold change) of 7.576, while TNFRSF14 showed moderate downregulation with a $\log_2$(fold change) of -1.106, suggesting their potential roles in underlying biological processes. Additionally, genes like SLC35E2A displayed significant upregulation with a $\log_2$(fold change) of 1.315, indicating potential involvement in cellular mechanisms. Conversely, genes such as *PRDM16* and SLC2A5 showed considerable downregulation with $\log_2$(fold changes) of -5.78 and -3.069, respectively, suggesting potential implications in disease pathogenesis. Furthermore, genes like GPR157 and PRXL2B exhibited notable changes in expression levels with $\log_2$(fold changes) of 1.345 and 1.015, respectively.

Volcano Plots- Group 4 vs Group 5

In comparing Group 4 (liver hepatocytes of control donors) to Group 5 (induced hepatocytes of *PH1* patients with *AGXT* mutation), the gene expression analysis revealed significant changes in the expression levels of several genes. For instance, ESPN displayed substantial upregulation with a $\log_2$(fold change) of 9.307, while HES2 exhibited significant downregulation with a $\log_2$(fold change) of -6.546. Additionally, genes like ATAD3C and TMEM52 showed notable upregulation with $\log_2$(fold changes) of 4.854 and 5.33, respectively, suggesting potential roles in the underlying biological processes. Conversely, SLC2A5 and SLC45A1 displayed considerable downregulation with $\log_2$(fold changes) of -8.839 and -5.152, respectively, indicating potential implications in disease pathogenesis. Furthermore, PLCH2 and ARHGEF16 exhibited significant changes in expression levels with $\log_2$(fold changes) of 3.838 and 6.262, respectively.

The "Log_2 Fold Change" indicates the extent of change in gene expression between two conditions. A positive value denotes upregulation (increased expression) in the second condition, while a negative value signifies downregulation (decreased expression). Conversely, the "-Log_{10}(Pvalue)" reflects the statistical significance of the expression change, with higher values suggesting stronger evidence against random chance. Typically, a threshold of -$\log_{10}$(Pvalue) > 2, corresponding to a p-value less than 0.01, is considered significant. Among the most significant changes observed, ESPN exhibits the highest upregulation, with a $\log_2$ fold change of 9.3 and a -$\log_{10}$(Pvalue) of 74.15, indicating a substantial increase in expression. Understanding the role of ESPN in this context could be crucial for further investigation. Conversely, MXRA8 shows the most significant downregulation, with a $\log_2$ fold change of -6.21 and a -$\log_{10}$(P value) of 14.92, prompting the need to explore why its expression is markedly decreased. Additionally, several genes, including HES2, SLC2A5, VAMP3, and ARHGEF16, exhibit noteworthy changes in expression levels, warranting further investigation into their roles in the observed differences between conditions. It's worth noting that some entries represent pseudogenes or uncharacterized LOCs, whose significance may require

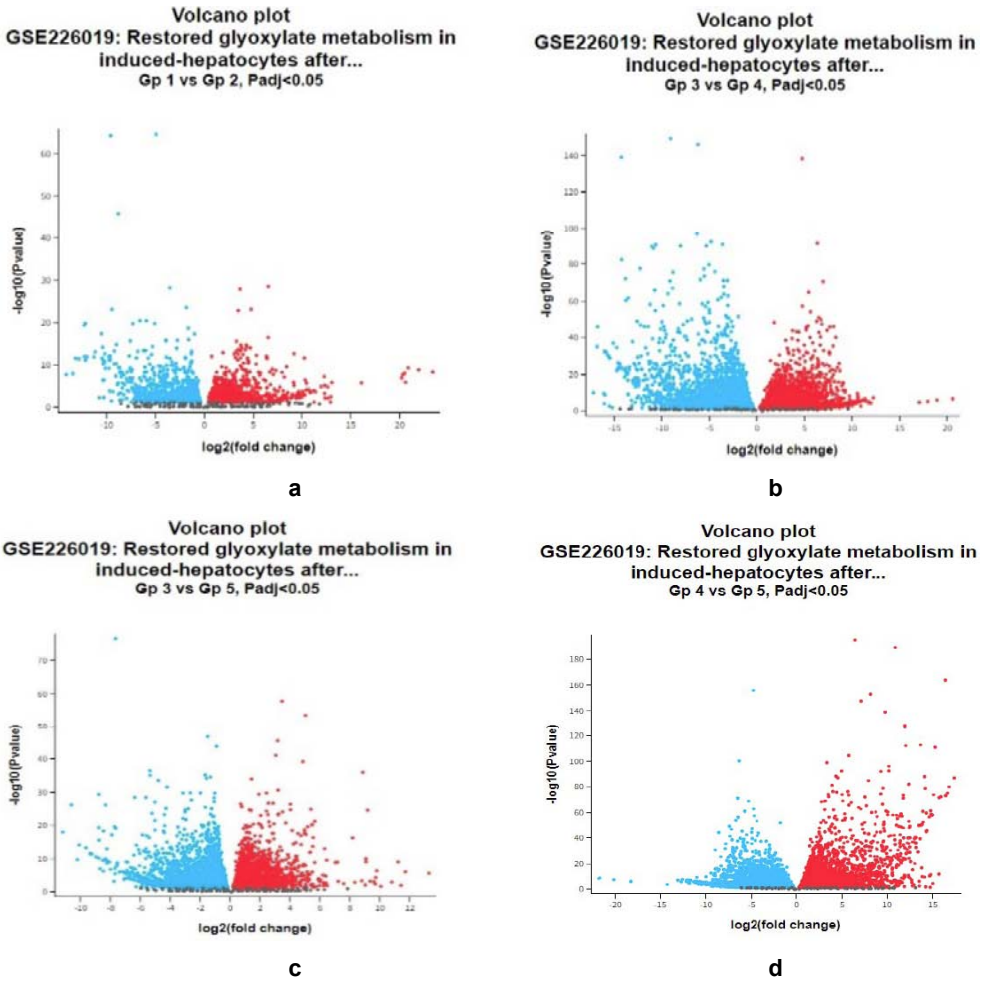

Figure 1: a: Volcano plot- Group 1x2. **b**: Volcano plot- Group 3x4. **c**: Volcano plot- Group 3x5. **d**: Volcano plot- Group 4x5.

additional scrutiny due to their ambiguous functional implications.

Overall, this data suggests significant changes in the expression of various genes. Further analysis and investigation into the most relevant genes would be necessary to understand the biological processes underlying the observed differences between the two conditions. These findings underscore a wide range of gene expression alterations that contribute to the molecular mechanisms underlying primary hyperoxaluria. Further functional analysis of these genes could illuminate novel therapeutic targets and deepen our understanding of the disease processes.

The $\log_2$ fold change in contrast to the average $\log_2$ expression value is shown in a mean difference (MD) plot (Figure **2a** and **2b**) created with limma (plotMD), which is helpful for demonstrating differentially expressed genes.

A mean distinction plot shows the test results for a solitary difference, compares the disagreement or differences between two quantitative measurements, i.e., the expression of genes.

Mean Difference Plots Group 1 versus Group 2

In comparing Group 1 (skin fibroblasts of healthy controls) to Group 2 (skin fibroblasts *AGXT* targeted knock-in), the mean difference plots illustrate the expression changes of various genes across different conditions. Notably, HES5, a member of the Hes family bHLH transcription factors, showed a substantial increase in expression with a $\log_2$(fold change) of 6.938, suggesting a potential role in the regulatory processes underlying the studied condition. Conversely, *PRDM16* exhibited significant downregulation with a $\log_2$(fold change) of -6.376, indicating its potential involvement as a suppressor or modulator of the studied pathways. Additionally, genes like TMEM51 and RUNX3 displayed notable upregulation with $\log_2$(fold changes) of 9.021 and 9.778, respectively, suggesting their potential importance in the biological processes being

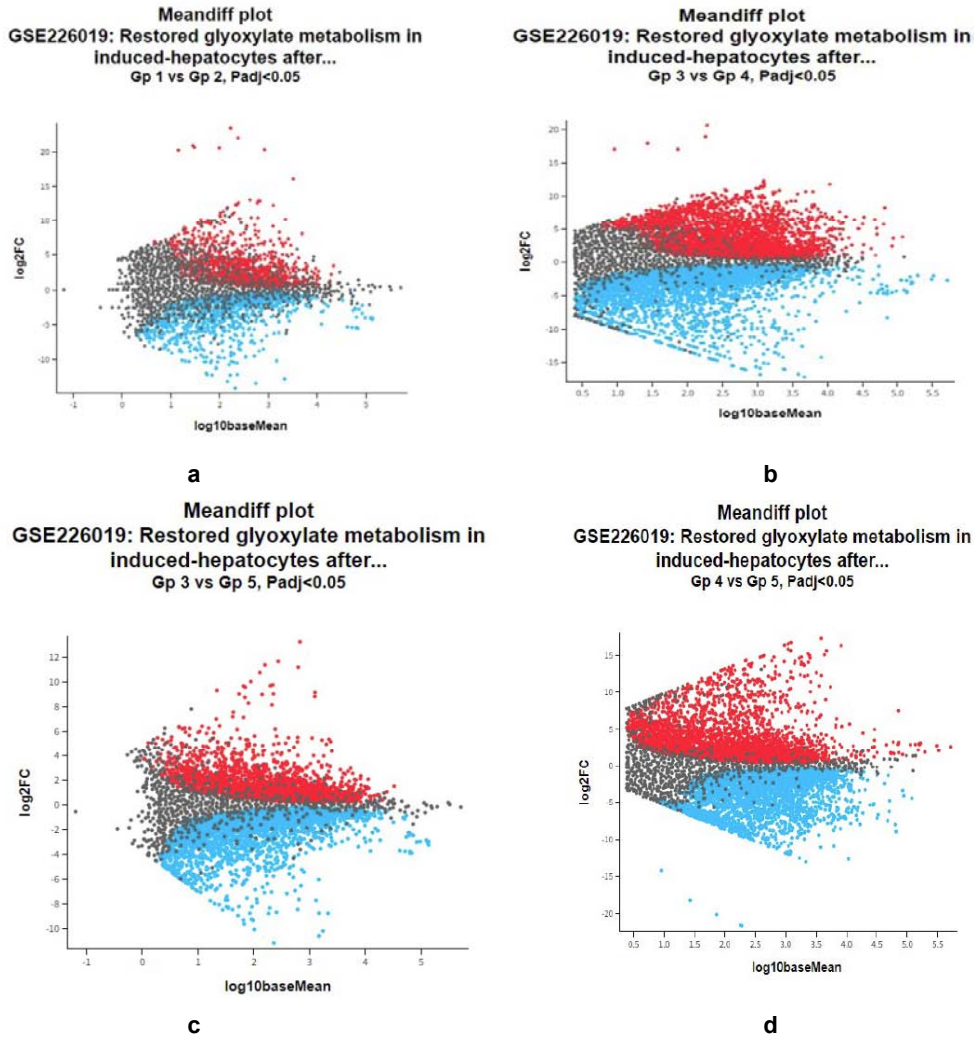

Figure 2: a: Mean difference- Group 1 x 2. **b**: Mean difference- Group 3x4. **c**: Mean difference -Group 3x5. **d**: Mean difference - Group 4x5.

investigated. Conversely, genes like WNT4 exhibited significant downregulation with a $\log_2$(fold change) of -5.904, indicating potential implications in disease pathogenesis. Furthermore, genes such as GPR3 showed substantial upregulation with a $\log_2$(fold change) of 6.974, suggesting their potential roles as mediators or regulators of the studied pathways. These findings highlight the diverse molecular landscape associated with *PH1* and underscore potential candidate genes for further investigation to elucidate their roles in disease progression and pathophysiology.

Mean Difference Plot – Group 3 versus Group 4

The mean difference plot comparing Group 3 (induced hepatocytes from control donors) to Group 4 (liver hepatocytes from control donors) illustrates significant alterations in gene expression levels across these conditions. Notably, genes such as WASH7P, LOC729737, WASH9P, and LOC100132287 exhibit consistent downregulation, indicating a decrease in their expression compared to the baseline. Conversely, genes like HES2 display significant upregulation, suggesting an increase in expression levels. Furthermore, the plot highlights genes with considerable variability in expression levels, as indicated by the wide range of $\log_2$ fold changes observed. These findings offer valuable insights into the molecular mechanisms underlying the observed biological phenomena, particularly in the context of hepatocyte differentiation and function in both induced and native states. Additionally, they underscore the importance of further investigating the roles of these genes in cellular processes, particularly those relevant to hepatocyte function and liver physiology.

Mean Difference -Group 3 versus Group 5

The mean difference plot comparing Group 3 (induced hepatocytes from control donors) to Group 5

(induced hepatocytes from *PH1* patients with *AGXT* mutation) reveals significant changes in gene expression levels. Notably, AJAP1 (adherens junctions associated protein 1) exhibits a substantial increase in expression with a high $\log_2$ fold change of 7.576, indicating a pronounced upregulation in *PH1*. Conversely, *PRDM16* (PR/SET domain 16) shows a considerable decrease in expression with a $\log_2$ fold change of -5.78, suggesting downregulation in *PH1*. Additionally, genes like GPR157 (G protein-coupled receptor 157) and PIK3CD (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta) also display noteworthy changes in expression levels, albeit to a lesser extent. Furthermore, some genes exhibit relatively small $\log_2$ fold changes, indicating minor alterations in expression levels in the context of *PH1*. These findings provide valuable insights into the molecular differences between induced hepatocytes from control donors and those from *PH1* patients, shedding light on potential mechanisms underlying the pathology of the disease and identifying candidate genes for further investigation.

Mean Difference -Group 4 vs Group 5

The mean difference plot comparing Group 4 (liver hepatocytes from control donors) to Group 5 (induced hepatocytes from *PH1* patients with *AGXT* mutation) reveals a diverse landscape of transcriptional changes across various genes and non-coding RNAs. Significant alterations in expression levels are observed, indicating potential roles in various biological processes. Interestingly, the dataset encompasses pseudogenes, non-coding RNAs, and protein-coding genes, highlighting the complexity of the transcriptome. Genes such as ARHGEF16 and TMEM52 exhibit substantial fold changes, suggesting their involvement in cellular signaling and membrane transport. Moreover, non-coding RNAs like LINC01128 and PRKCZ-AS1 display distinct expression patterns, hinting at their regulatory functions. Furthermore, genes with negative fold changes, including LRRC47 and SLC2A5, may play roles in metabolic processes and cellular homeostasis. Notably, outliers like ESPN and HES2 demonstrate extreme expression changes, warranting further investigation into their potential implications in health and disease. Overall, this dataset offers valuable insights into the dynamic regulation of gene expression, identifying potential targets for future research in molecular biology and biomedical sciences.

The R boxplot method, illustrated in Figure **3**, was utilized to visualize the distribution of values within the selected samples. These samples were grouped by

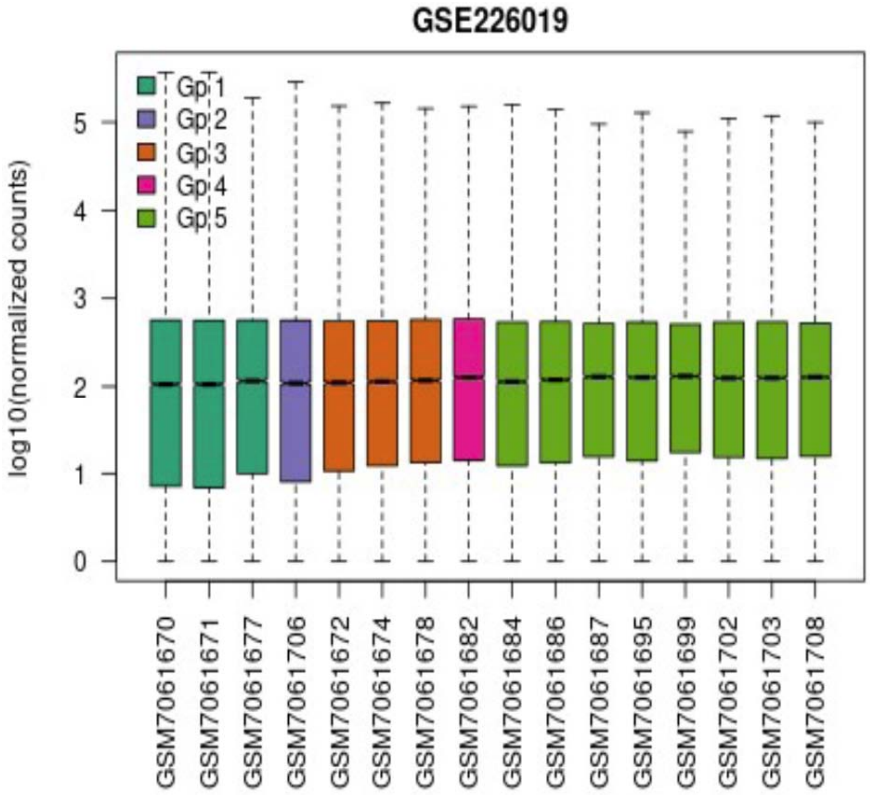

Figure 3: Box plot.

color to facilitate comparison. Assessing the distribution aids in determining the suitability of the chosen samples for further analysis of differential expression. Typically, data centered around the median suggests compatibility for cross-comparison and normalization.

The plot shows data after log transform and normalization, if they were performed.

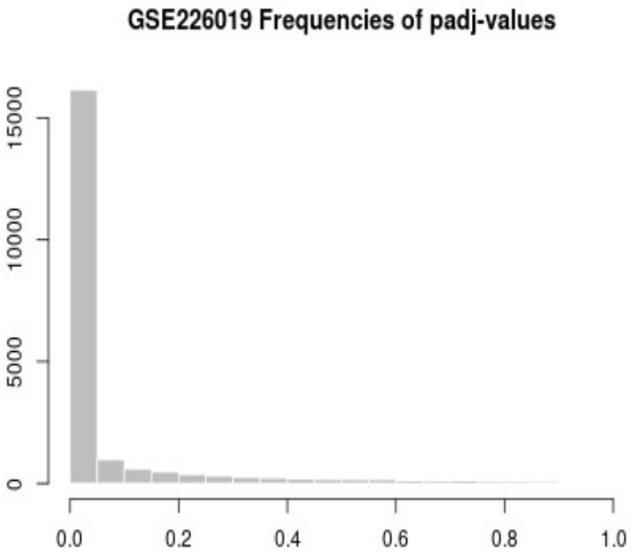

Figure 4: Histogram.

The histogram depicted in Figure **4**, generated using the hist function, serves to visualize the distribution of P-values obtained from the analysis results. These P-values were calculated using all selected contrasts, mirroring those presented in the top table of differentially expressed genes.

Following the fitting of a linear model, a mean variance plot, generated using R limma (plotSA, vooma), serves to confirm the relationship between the mean and variance of the expression data. This plot visually represents the level of variability within the data and helps assess the presence of a significant mean-variance trend. Additionally, it aids in determining whether utilizing the precision weights option to address this trend is advisable. Precision weights are particularly beneficial when a clear mean-variance trend is observed, as they improve the accuracy of test results. Each point on the plot corresponds to a gene, with the red line representing the mean-variance trend approximation. This approximation may already factor in the precision weight option during the analysis of differential gene expression. Conversely, the blue line represents an approximation of constant variance.

Produced utilizing UMAP, or Uniform Manifold Approximation and Projection (UMAP) (Figure **6**), this method is beneficial for dimensionality reduction, simplifying the visualization of the relationships between samples. By reducing the data to two dimensions, it offers insight into the proximity of samples to each other. The plot also illustrates the number of closest neighbors employed in the computation.

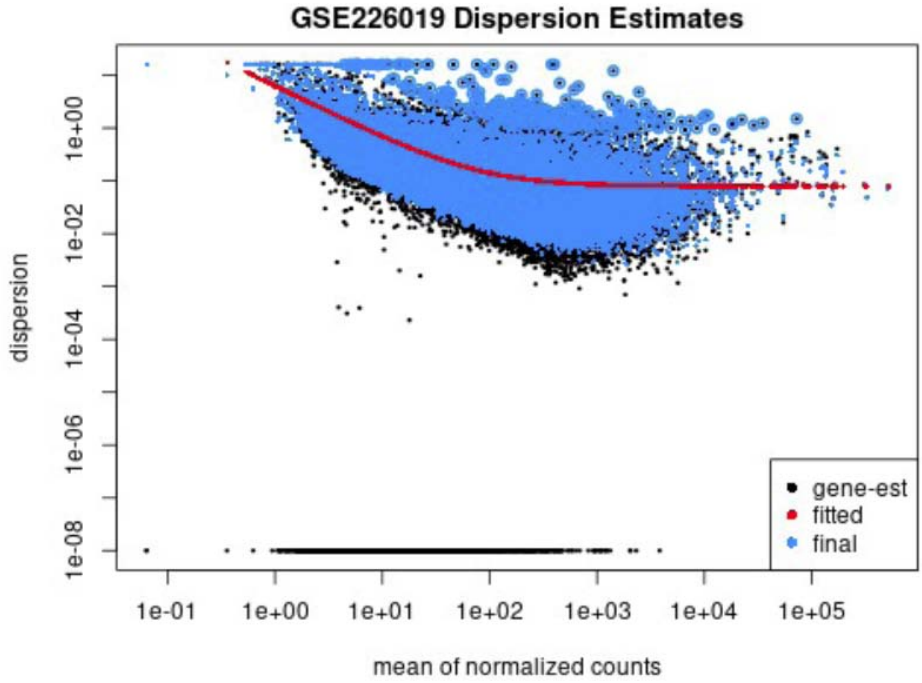

Figure 5: Mean variance plot.

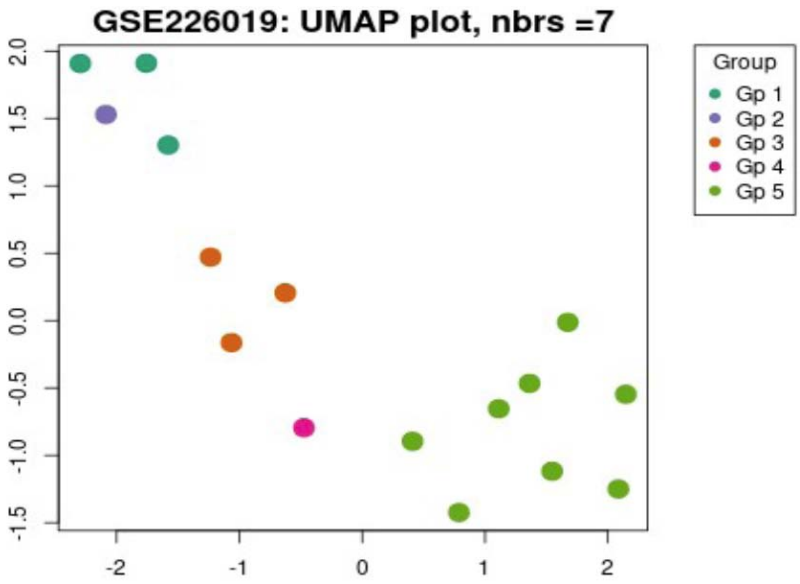

Figure 6: UMAP plot.

Using the limma algorithm, a Venn diagram (Figure **7**) is generated to explore and identify significant gene overlap among different contrasts, facilitating the comparison and download of pertinent data.

Venn Diagram Group 1 and 2

The Venn diagram comparing Group 1 (skin fibroblasts of healthy controls) and Group 2 (skin fibroblasts *AGXT* targeted knock-in) illustrates a complex interplay of gene expression changes across a diverse set of genes and non-coding RNAs. Several genes exhibit significant alterations in expression levels, suggesting potential involvement in various biological processes. Notably, the dataset encompasses pseudogenes, non-coding RNAs, and protein-coding genes, emphasizing the intricate nature of the transcriptome. Genes like ARHGEF16 and TMEM52 demonstrate substantial fold changes, implicating their roles in cellular signaling and membrane transport. Moreover, distinct expression patterns observed in non-coding RNAs such as LINC01128 and PRKCZ-AS1 hint at their regulatory functions. Additionally, genes showing negative fold changes, including LRRC47 and SLC2A5, may contribute to metabolic processes and cellular homeostasis. Remarkably, outliers like ESPN and HES2 exhibit extreme expression changes, meriting further exploration into their potential implications for health and disease. Overall, this comprehensive dataset offers valuable insights into the dynamic regulation of gene expression, presenting potential avenues for future research in molecular biology and biomedical sciences.

Venn – Group 3 and 4

The Venn diagram comparing Group 3 (induced hepatocytes of control donors) and Group 4 (liver hepatocytes of control donors) reveals distinct patterns of gene expression changes, shedding light on potential differences in underlying biological processes. Among the genes showing significant alterations, TNFRSF9 stands out with the highest fold change ($log_2FC = 6.098$), suggesting its potential importance in the investigated biological phenomena. Similarly, genes such as PADI3 ($log_2FC = 7.038$), SLC2A1-DT ($log_2FC = 5.529$), and RPE65 ($log_2FC = 5.805$) exhibit substantial upregulation, indicating their potential roles in cellular functions. In contrast, genes like LRRC8B ($log_2FC = -3.031$), LINC02788 ($log_2FC = -7.688$), and SNORA66 ($log_2FC = -3.834$) demonstrate notable downregulation, suggesting their involvement in different regulatory pathways. Moreover, genes like CDC14A, DDX20, and VPS45 display moderate yet significant changes in expression levels, hinting at their potential contributions to the studied conditions. These findings highlight the complex interplay of gene expression dynamics between Group 3 and Group 4, providing valuable insights for further investigation into the underlying mechanisms of these biological processes.

Venn -Group 3 and 5

The Venn diagram comparing Group 3 (induced hepatocytes of control donors) and Group 5 (induced hepatocytes of *PH1* patients with *AGXT* mutation) reveals distinct patterns of gene expression changes,

suggesting differences in underlying biological processes between the two groups. Among the genes showing significant alterations, AJAP1 stands out with a considerable upregulation, indicated by a $\log_2$ fold change of 7.576, suggesting its potential importance in the biological processes under investigation. Similarly, genes such as TENT5B ($\log_2$FC = 3.172), PADI1 ($\log_2$FC = 2.348), and *EPHA2*-AS1 ($\log_2$FC = 2.471) also exhibit notable upregulation, indicating their potential roles in cellular functions. Conversely, genes like CLSPN ($\log_2$FC = -1.424), CLCNKB ($\log_2$FC = -3.345), and TINAGL1 ($\log_2$FC = -5.153) demonstrate significant downregulation, suggesting their involvement in different regulatory pathways. These findings highlight the complex interplay of gene expression dynamics between Group 3 and Group 5, providing valuable insights for further investigation into the underlying mechanisms of these biological processes.

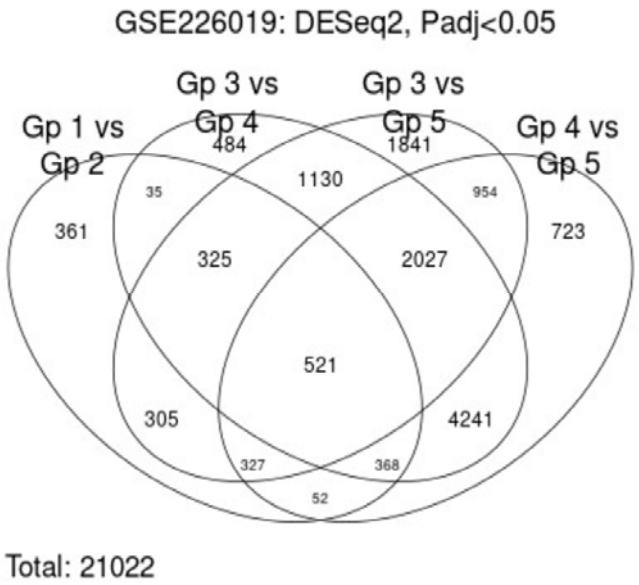

Figure 7: Venn diagram.

Venn Group 4 and 5

The Venn diagram comparing Group 4 (liver hepatocytes of control donors) and Group 5 (induced hepatocytes of *PH1* patients with *AGXT* mutation) highlights distinctive patterns of gene expression changes between the two groups. Among the genes exhibiting significant alterations, several stand out with notable upregulation, including NPPA ($\log_2$FC = 2.638), ZPLD2P ($\log_2$FC = 4.878), and ERRFI1 ($\log_2$FC = 2.179), suggesting their potential roles in the biological processes under investigation. Conversely, genes such as SLC45A1 ($\log_2$FC = -5.152), TAL1 ($\log_2$FC = -5.116), and NGF ($\log_2$FC = -5.26) demonstrate substantial downregulation, indicating their potential involvement as suppressors or inhibitors in the studied pathways or conditions. These findings provide valuable insights into the differential gene expression profiles between Group 4 and Group 5, offering avenues for further research to elucidate the functional significance of these genes in the context of the investigated biological processes.

The string analysis results indicate strong associations between various genes based on their interactions and functions, such as the interplay between *AMBP* and GC in inhibiting calcium oxalate crystallization, the involvement of *ANGPTL3* and GC in lipid and glucose metabolism regulation, and the significant metabolic interactions between *CYP2C9* and *CYP3A4* in the metabolism of endogenous substrates.

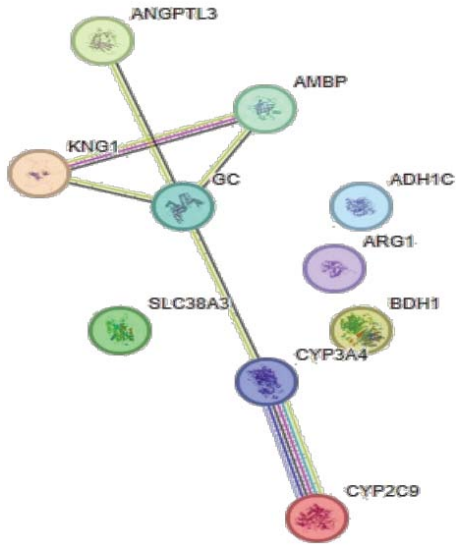

Figure 8: Protein-protein interactions of top 10 genes.

The network statistics indicate that there are 10 nodes (genes or proteins) and 6 edges (interactions) present. The average node degree is 1.2, which means that on average, each node is connected to 1.2 other nodes. The average local clustering coefficient is 0.417, suggesting a relatively high degree of clustering or interconnectedness among neighboring nodes. The expected number of edges is 1, but the actual number of edges is significantly higher, indicating that the network has more interactions between nodes than expected by chance.

The network analysis reveals significant enrichment in various biological processes related to metabolism. Among these processes, the Monoterpenoid metabolic process and Terpenoid metabolic process stand out, indicating the involvement of specific genes/proteins in

the synthesis and modification of monoterpenoids and terpenoids, respectively. Additionally, the Steroid metabolic process, Alcohol metabolic process, Organic hydroxy compound metabolic process, and Lipid metabolic process are also enriched, suggesting the activity of genes/proteins associated with the biosynthesis, breakdown, and modification of steroids, alcohols, organic hydroxy compounds, and lipids. These findings underscore the importance of these metabolic pathways in cellular functions, including energy production, signaling, and structural integrity. The network's composition of genes/proteins involved in these processes provides valuable insights into the molecular mechanisms underlying various physiological and pathological conditions associated with metabolism, such as hormone regulation, lipid disorders, and metabolic diseases. Further investigation into the specific roles of these genes/proteins can contribute to a better understanding of metabolic pathways and their implications for human health and disease.

The molecular function analysis highlights the presence of specific activities within the network. Caffeine oxidase activity, represented by 2 out of 4 interactions, suggests the involvement of enzymes responsible for metabolizing caffeine. This activity indicates the potential for caffeine metabolism within the system, which may have implications for caffeine clearance or its effects on downstream cellular processes. Oxidoreductase activity, observed in 5 out of 731 interactions, denotes a broader category of enzymes involved in oxidation-reduction reactions, implying their role in electron transfer processes. This activity is fundamental to various cellular functions, including energy production and metabolism. Furthermore, Heme binding, found in 3 out of 140 interactions, suggests the presence of proteins capable of binding to heme molecules. Heme binding proteins often play crucial roles in transporting, storing, or catalyzing reactions involving heme-containing molecules, such as hemoglobin and cytochromes. These molecular functions provide insights into the biochemical processes occurring within the network and highlight the diversity of enzymatic activities involved in cellular metabolism and signaling.

The KEGG pathway analysis highlights several significant pathways enriched within the network, providing insights into the molecular mechanisms at play. Firstly, "Drug metabolism - cytochrome P450" (hsa00982) signifies the involvement of cytochrome P450 enzymes in metabolizing various drugs, emphasizing the network's role in drug detoxification and elimination. Additionally, "Retinol metabolism" (hsa00830) suggests active processes related to the metabolism of vitamin A and its derivatives, crucial for vision and cellular function. The presence of "Metabolism of xenobiotics by cytochrome P450" (hsa00980) underscores the network's participation in the biotransformation of foreign compounds, contributing to detoxification mechanisms. "Chemical carcinogenesis" (hsa05204) indicates potential involvement in pathways related to cancer initiation and progression, while "Linoleic acid metabolism" (hsa00591) suggests activities linked to omega-6 fatty acid metabolism, important for cellular signaling. Lastly, the enrichment of "Metabolic pathways" (hsa01100) underscores the diverse metabolic activities occurring within the network, encompassing various biosynthetic and degradation processes essential for cellular homeostasis. Collectively, these pathway enrichments provide a comprehensive understanding of the network's biological functions and potential implications in health and disease.

Table 2: miRNAs for Top 10 *DEGs*

Genes	MiRNA
ANGPTL3	hsa-miR-4501
SLC38A3	hsa-miR-5692c
KNG1	hsa-miR-6731-3p
BDH1	hsa-miR-6867-5p
GC	hsa-miR-616-3p
ADH1C	hsa-miR-4468
ARG1	hsa-miR-3692-3p
CYP3A4	hsa-miR-4277
AMBP	hsa-miR-4763-5p
CYP2C9	hsa-miR-4797-5p

DISCUSSION

Role of miRNAs in Gene Regulation and Kidney Function

MiRNAs, capable of targeting over 60% of human genes, are pivotal in regulating diverse biological processes, including gene expression, development, and homeostasis [17,18]. In the context of kidney biology, miRNAs are integral to the development, structure, and function of renal tissues. They influence critical processes such as fluid and electrolyte balance, acid-base homeostasis, and blood pressure regulation.

Furthermore, their involvement in pathogenic mechanisms underscores their potential as therapeutic targets.

Notably, the stability of miRNAs in serum and urine, irrespective of storage conditions, enhances their appeal as reliable biomarkers for diagnosing and monitoring kidney injuries [19]. This highlights the translational potential of miRNA-target networks in advancing our understanding of diseases like Primary Hyperoxaluria Type 1 (*PH1*).

Contribution of Statistical Methodologies

The robust statistical framework employed in this study significantly contributed to the accuracy and reliability of the findings. The use of limma for differential expression analysis ensured precise normalization and calibration across experimental batches, minimizing potential biases in the dataset. The selection criteria based on fold-change and p-value thresholds ($p < 0.05$) allowed for the identification of biologically significant *DEGs*, which formed the foundation for subsequent functional and pathway analyses.

Visualization tools such as volcano plots and heatmaps facilitated intuitive interpretations of *DEGs*, providing insights into gene expression patterns across experimental groups. The incorporation of GO and KEGG enrichment analyses using ClusterProfiler further underscored the relevance of identified genes in pathways critical to *PH1* pathophysiology.

Protein-Protein Interaction (PPI) networks, constructed via STRING and visualized in Cytoscape, revealed key hub genes with central roles in cellular processes. The validation of these hub genes using GEO2R enhanced the statistical rigor, ensuring that the findings were not artifacts of a specific dataset but were reproducible across similar experimental settings.

Applicability in Related Research

The statistical methodologies utilized in this study are widely applicable in other research areas involving high-throughput omics datasets. For example:

Differential Gene Expression Analysis: The approach can be adapted to study other genetic diseases, cancers, or metabolic disorders by analyzing transcriptomic data.

Pathway Enrichment and PPI Networks: GO and KEGG analyses, coupled with PPI construction, offer a comprehensive view of molecular mechanisms in diverse conditions, from neurodegenerative diseases to cardiovascular pathologies.

MiRNA-Target Networks: By integrating databases like miRDB, researchers can identify regulatory networks in various diseases, extending the utility of the methodology to precision medicine applications.

By demonstrating how biostatistical tools and frameworks can uncover clinically relevant biomarkers and therapeutic targets, this study provides a model for leveraging bioinformatics in translational research. The reproducibility and adaptability of these methods ensure their applicability in addressing similar questions across a wide range of biological disciplines.

CONCLUSION

The top ten differentially expressed genes identified—*ANGPTL3*, *SLC38A3*, *KNG1*, *BDH1*, GC, *ADH1C*, *ARG1*, *CYP3A4*, *AMBP*, and *CYP2C9*—exhibit strong associations with various biological pathways. Pathways like Linoleic acid metabolism and drug metabolism-cytochrome P450 demonstrate significant overrepresentation of these genes. This provides insights into the molecular mechanisms underlying the studied condition. Notably, genes like ESPN show the highest upregulation, while MXRA8 demonstrates the most significant downregulation, suggesting their potential roles in disease pathogenesis. Furthermore, network analysis highlights the involvement of these genes in critical metabolic processes, offering potential targets for further investigation. Understanding the regulatory role of specific miRNAs (hsa-miR-4501,hsa-miR-5692c,hsa-miR-6731-3p,hsa-miR-6867-5p,hsa-miR-616-3p,hsa-miR-4468,hsa-miR-3692-3p,hsa-miR-4277,hsa-miR-4763-5p,hsa-miR-4797-5p) in gene expression could provide further insights into disease mechanisms and potential therapeutic avenues. Overall, this study enhances our understanding of primary hyperoxaluria's molecular landscape and identifies potential targets for future research and therapeutic interventions.

OUTCOMES OF THE STUDY

The research found 10 pivotal differentially expressed genes (*DEGs*) *ANGPTL3, SLC38A3, KNG1, BDH1, GC, ADH1C, ARG1, CYP3A4, AMBP,* and *CYP2C9* that are strongly linked to essential biological activities, including linoleic acid metabolism and drug metabolism via cytochrome P450 pathways. Protein-

protein interaction networks and miRNA-target interaction networks were established, uncovering complex molecular and regulatory pathways. Particular miRNAs, such as hsa-miR-4501 and hsa-miR-5692c, were identified as regulators of these *DEGs*, providing significant insights into disease processes and prospective treatment targets.

RATIONALE OF THE STUDY

The research seeks to explore a new treatment strategy for Primary Hyperoxaluria Type 1 (*PH1*) by examining the effects of gene repair at the *AGXT* locus and the direct transformation of fibroblasts from *PH1* patients into induced hepatocytes (*iHeps*) via *CRISPR-Cas9* technology. The research aims to elucidate the molecular processes of hyperoxaluria and its advancement to oxalate crystal formation by the analysis of gene expression data. This bioinformatics-based method enables the identification of crucial regulatory genes and pathways, connecting genetic repair procedures with their biological consequences.

LIMITATIONS OF THE STUDY

The work offers useful insights on differentially expressed genes and related pathways; nevertheless, it is constrained by its dependence on microarray data, which may not possess the depth and resolution of RNA-sequencing data. Furthermore, the results stem from bioinformatic predictions and need experimental confirmation to verify the functions of the detected *DEGs* and miRNAs. The study concentrates on a particular dataset, perhaps constraining the applicability of the findings to wider populations or alternative datasets.

REFERENCES

[1] Li Y, Zheng R, Xu G, Huang Y, Li Y, Li D, *et al.* Generation and characterization of a novel rat model of primary hyperoxaluria type 1 with a nonsense mutation in alanine-glyoxylate aminotransferase gene. Am J Physiol Renal Physiol 2021; 320(3): F475-F484. https://doi.org/10.1152/ajprenal.00514.2020

[2] Groothoff JW, Metry E, Deesker L, Garrelfs S, Acquaviva C, Almardini R, *et al.* Clinical practice recommendations for primary hyperoxaluria: an expert consensus statement from ERKNet and OxalEurope. Nat Rev Nephrol 2023; 19(3): 194-211. https://doi.org/10.1038/s41581-022-00661-1

[3] Mandrile G, Beck B, Acquaviva C, Rumsby G, Deesker L, Garrelfs S, *et al.* Genetic assessment in primary hyperoxaluria: why it matters. Pediatr Nephrol 2023; 38(3): 625-634. https://doi.org/10.1007/s00467-022-05613-2

[4] Wannous H. Primary hyperoxaluria type 1 in children: clinical and laboratory manifestations and outcome. Pediatr Nephrol 2023; 38(8): 2643-2648. https://doi.org/10.1007/s00467-023-05917-x

[5] Gang X, Liu F, Mao J. Lumasiran for primary hyperoxaluria type 1: What we have learned? Front Pediatr 2023; 10: 1052625. https://doi.org/10.3389/fped.2022.1052625

[6] Soliman NA, Mabrouk S. Primary hyperoxaluria type 1 in developing countries: novel challenges in a new therapeutic era. Clin Kidney J 2022; 15(Suppl 1): i33-i36. https://doi.org/10.1093/ckj/sfab203

[7] Nieto-Romero V, García-Torralba A, Molinos-Vicente A, Moya FJ, Rodríguez-Perales S, García-Escudero R, *et al.* Restored glyoxylate metabolism after *AGXT* gene correction and direct reprogramming of primary hyperoxaluria type 1 fibroblasts. iScience 2024; 27(4): 109530. https://doi.org/10.1016/j.isci.2024.109530

[8] Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, *et al.* The gene ontology knowledgebase in 2023. Genetics 2023; 224(1): iyad031.

[9] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res 2019; 47(D1): D590-D595. https://doi.org/10.1093/nar/gky962

[10] Khatri A, Singh VK, Prasad R, Kumar A, Singh VK, Joshi D. In Silico Functional Network Analysis for the Identification of Novel Target Associated with SCN1A Gene. Biomed Biotechnol Res J 2023; 7(2): 163-169. https://doi.org/10.4103/bbrj.bbrj_46_23

[11] Pande A. Co-regulatory network of transcription factor and microRNA: A key player of gene regulation. Biomed Biotechnol Res J 2021; 5(4): 374-379. https://doi.org/10.4103/bbrj.bbrj_182_21

[12] Ashwini K, Gollapalli P, Shetty SS, Raghotham A, Shetty P, Shetty J, *et al.* Gene enrichment analysis and protein-protein interaction network topology delineates S-phase kinase-associated protein 1 and catenin beta-1 as potential signature genes linked to glioblastoma prognosis. Biomed Biotechnol Res J 2023; 7(1): 37-47. https://doi.org/10.4103/bbrj.bbrj_344_22

[13] Yadav R. Gene expression analysis to network construction for the identification of hub genes involved in neurodevelopment. Biomed Biotechnol Res J 2021; 5(4): 425-434. https://doi.org/10.4103/bbrj.bbrj_250_21

[14] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019; 47: D607-D613. https://doi.org/10.1093/nar/gky1131

[15] Atoum MF, Alowaisy D, Deeb AA. Analysis of MicroRNA Processing Machinery Gene DROSHA, DICER1, and XPO5 Variants Association with Atherosclerosis: A Case-control Study. Biomed Biotechnol Res J 2024; 8(4): 447-454. https://doi.org/10.4103/bbrj.bbrj_256_24

[16] Pu M, Chen J, Tao Z, Miao L, Qi X, Wang Y, *et al.* Regulatory network of miRNA on its target: coordination between transcriptional and post-transcriptional regulation of gene expression. Cell Mol Life Sci 2019; 76(3): 441-451. https://doi.org/10.1007/s00018-018-2940-7

[17] Panizo S, Martínez-Arias L, Alonso-Montes C, Cannata P, Martín-Carro B, Fernández-Martín JL, *et al.* Fibrosis in chronic kidney disease: pathogenesis and consequences. Int J Mol Sci 2021; 22. https://doi.org/10.3390/ijms22010408

[18] Carbonell T, Gomes AV. MicroRNAs in the regulation of cellular redox status and its implications in myocardial ischemia-reperfusion injury. Redox Biol 2020; 36: 101607. https://doi.org/10.1016/j.redox.2020.101607

[19] Peters LJF, Floege J, Biessen EAL, Jankowski J, van der Vorst EPC. MicroRNAs in chronic kidney disease: four candidates for clinical application. Int J Mol Sci 2020; 21. https://doi.org/10.3390/ijms21186547

Received on 28-10-2024 Accepted on 26-11-2024 Published on 27-12-2024

https://doi.org/10.6000/1929-6029.2024.13.38