IDENTIFICATION OF SIGNIFICANT CONSISTENT GENES AND THEIR RELATIONSHIP WITH HXR9 IC50 VALUE IN PROSTATE CANCER CELL LINES (KG1, HEL92.1.7, KU812F, A375-M, K-562, and MONO-MAC-6)

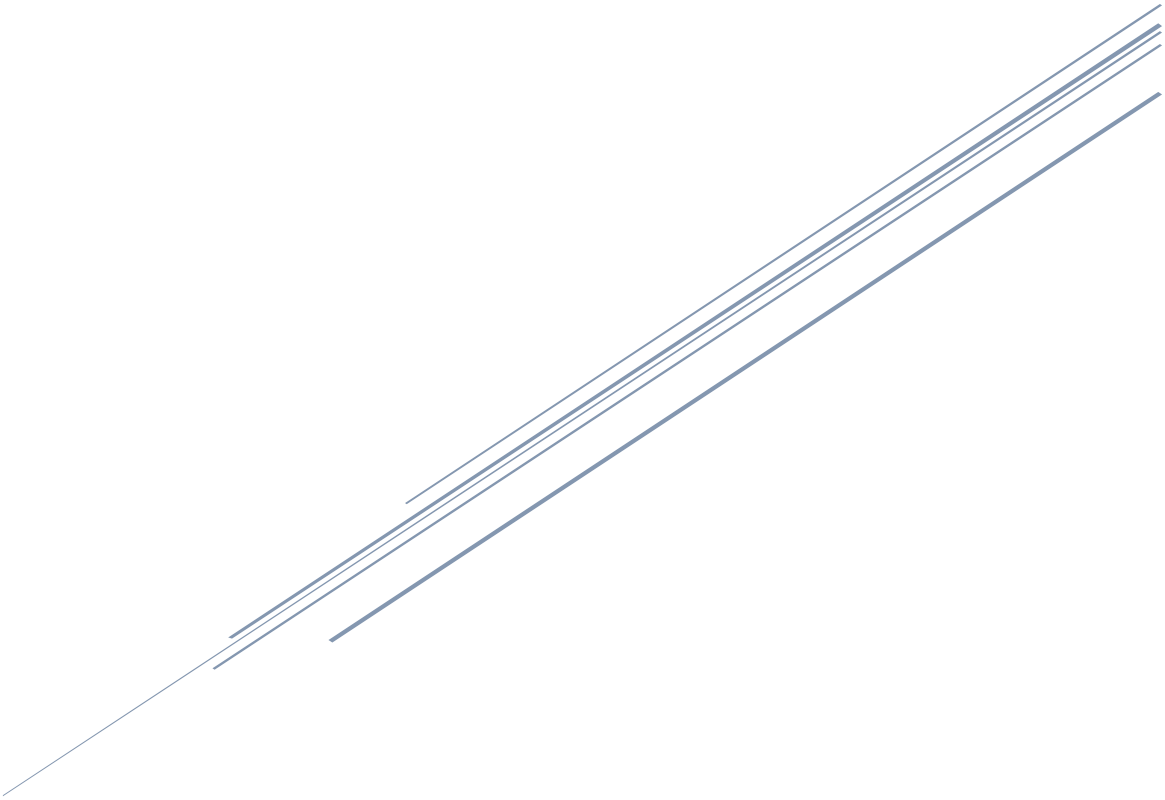
MSc Bioinformatics – 21524331

Rahul Ladhani

School of Biomedical Sciences

University of West London

A dissertation submitted for the degree of Master of Science in Bioinformatics



UNIVERSITY OF
WEST LONDON

Acknowledgement

I would like to take this opportunity to express my sincere gratitude to the individuals whose invaluable contributions have been pivotal to the completion of my project's final report.

First and foremost, I extend my deepest appreciation to Dr. Obed Brew, our esteemed course leader. Dr. Brew's guidance and mentorship have been unwavering sources of support and insight throughout this academic journey. His contributions have left an indelible mark on my research and personal growth.

I am profoundly thankful to Suhirthakumar Puvanendran for generously sharing the gene expression analysis codes from his PhD research project at UWL, titled "Automatic Recognition of Pre-eclampsia - An Application of Artificial Intelligence for Early Pregnancy Risk Detection." This collaboration has significantly enriched the analytical depth of my study, and I am grateful for his expertise and assistance.

I would also like to extend my appreciation to Beatriz Manso for providing the pre-processed and quality-controlled data that formed the bedrock of my research. Her meticulous work and dedication to ensuring data accuracy have been instrumental in the success of my project.

Lastly, I want to convey my heartfelt thanks to my family, friends, and all those who have supported and encouraged me both academically and personally. Your unwavering belief in my abilities has been a driving force behind my research efforts. I am acutely aware that this accomplishment is as much yours as it is mine.

With sincere gratitude,

Rahul Rajenbhai Ladhani

Table of Content

List of Figure

List of Tables

Abstract

Because prostate cancer is a malignancy that is common around the world, it is imperative to fully comprehend its genetic and molecular causes. Particularly concerning prostate cancer, HOX genes, a crucial set of genes, have been linked to several developmental and cancer-related events. The focus is on the synthetic peptide HXR9, which disrupts their connection with the PBX gene family and exhibits anti-cancer activities. This contact is crucial in the development of cancer.

In addition, this work uses the strength of RankProduct statistics to identify genes with constant expression patterns in several prostate cancer cell lines. The study sheds light on the relationship between gene expression and therapeutic response by investigating the link between gene expression and HXR9 IC50 values. A useful test-bed for determining the value of RankProduct statistics is the KG-1 cell line, a well-known model for myeloid leukaemia.

This comprehensive strategy vastly improves our comprehension of the genetic processes underlying prostate cancer. It clarifies the function of the HOX-PBX connection in cancer and highlights HXR9's potential as a cancer preventative. The research offers a holistic picture of prostate cancer research by delivering crucial insights into prospective therapy targets.

1. Introduction

1.1 Prostate cancer

The prostate is a gland that generally has walnut-sized dimensions and enlarges with age. It is located below the bladder and surrounds the urethra, the tube that transfers urine from the body. The prostate gland’s primary function is to produce and secrete a fluid that nourishes and protects sperm during ejaculation(*Prostate cancer UK,*2023).

When prostate cells start to multiply uncontrollably, prostate cancer may develop. Some prostate cancers progress too slowly to pose a threat to health or shorten life expectancy. During the course of their condition, a sizable percentage of prostate cancer patients may not need therapy. Nevertheless, certain prostate tumours have a stronger potential to spread and grow quickly. It has to be treated to prevent it from spreading since it is more likely to cause consequences. (*Prostate cancer UK ,*2023).

Prostate cancer ranked second among all male malignancies worldwide in 2018 (after lung cancer), with around 1,280,000 new cases and more than 345,000 fatalities (nearly 3.7% of all cancer-related deaths in men). Prostate cancer incidence and death

rate are correlated with ageing worldwide, with an average age of 66 at diagnosis. However, it is predicted that there would be around 2,293,000 new cases up until 2040, and a slight change in mortality (an increase of 1.05%) will be seen (Rawla, 2019).

Early stages of prostate cancer frequently have no symptoms and a slow progression, making therapy in some cases limited or non-existent. Although prostatic enlargement can potentially cause these symptoms, the most common complaints are difficulty urinating, increased frequency, and nocturia. As the axis skeleton is the most frequent location of bone metastatic illness, more advanced stages of the disease may manifest with urine incontinence and back discomfort (Rawla, 2019).

Prostate-specific antigen (PSA) blood levels that are high (PSA > 4 ng/mL) are frequently used to identify instances of prostate cancer. Prostate tissue normally produces the glycoprotein known as PSA. However, a tissue biopsy is still required to establish the existence of cancer because it has been noted that men without cancer can also have elevated PSA values (Rawla, 2019).

The onset and spread of prostate cancer are significantly influenced by diet and exercise. Dietary variables can be primarily blamed for the variations in prostate cancer incidence rates seen among diverse foreign and ethnic groupings (Rawla, 2019).

The study of the genes linked to the illness and the mutations in acquired prostate cancer are the main areas of research. A thorough analysis of prostate cancer epidemiology and consideration of risk factors is essential to better comprehend how genetic anomalies and environmental variables combine to create these alterations and foster tumour growth. Increasing our knowledge of the causes and risk factors linked to prostate cancer will help us recognise at-risk individuals and make it easier to create efficient screening and preventative methods (Rawla, 2019).

1.2 HOX gene

Homeobox genes, which include Hox and HOX genes, are a group of highly conserved genes that share the homeobox DNA sequence. HOMEOBOX(HOX) genes are a family of genes that are essential for embryonic development, particularly for the generation of body parts along the anterior-posterior axis (Quinonez and Innis, 2014).

Studies on Drosophila melanogaster, or the fruit fly, led to the discovery of homeobox genes. The homeotic complex in Drosophila is a group of eight homeobox genes that are clustered together in one area of the genome. Clustered homeobox genes, also found in vertebrate genomes, are known as Hox genes in nonhuman vertebrates and HOX genes in humans. At least 39 Hox/HOX genes are present. They appear in four different clusters in mice and human beings. These clusters, which know by the names HOXA, HOXB, HOXC, and HOXD, are situated on chromosomes 7p14, 17q21, 12q13, and 2q31, respectively. (Daftary and Taylor, 2006).

The homeodomain, a DNA-binding domain, is encoded by the homeobox sequence. The proteins made by these genes, known as homeodomain transcription factors, are able to bind to particular DNA sequences and control the expression of other genes due to the homeodomain (Deguchi and Kehrl).

HOX gene mutations can cause serious developmental problems and malformations. HOXA13 gene mutations have been linked to Hand-foot-genital syndrome (HFGS).

Limb malformations, urogenital anomalies, and in some cases, hearing loss are the hallmarks of this condition. It is thought that HOXA13 mutations have a role in the emergence of HFGS and the characteristics that go along with it. Synpolydactyly is a disorder characterised by the fusing of digits (syndactyly) and the presence of additional digits (polydactyly), which can be brought on by mutations in the HOXD13 gene. Bosley-Salih-Alorainy syndrome (BSAS), which includes intellectual incapacity, hearing loss, and facial deformities, is associated with HOXA1 gene mutations. Autosomal recessive pure hair and nail ectodermal dysplasia (ECTD9) is a condition characterized by abnormal hair development and brittle nails, and it is linked to mutations in the HOXC13 gene. On the other hand, prostate cancer risk can be increased by certain mutations in the HOXB13 gene, particularly in individuals with a family history of the disease(Paço, de Bessa Garcia and Freitas, 2020).

The Pre-B-cell Leukaemia Homeobox (PBX) gene is a member of the family of Homeobox genes. Four PBX genes are transcribed in the human genome, and they are homologs of the Drosophila extradenticle gene (Exd). Although they do not form chromosomal clusters, they encode transcription factors that include homeodomains similar to the HOX genes. PBX proteins also have additional highly conserved areas, one of which is necessary for binding to several closely related transcription factors, MEIS and PREP. When a HOX/PBX DNA binding consensus is present, PBX proteins have a high tendency for building complexes with HOX1-11 proteins (Morgan *et al.,* 2017).

Cancer development and progression have been linked to the interaction between PBX and HOX proteins. Numerous studies have shown that abnormal expression or dysregulation of the PBX and HOX proteins can affect vital physiological functions such as cell proliferation, differentiation, and apoptosis, which can lead to oncogenesis. These interactions may activate carcinogenic pathways or interfere with tumour suppressor activities. Additionally, chromosomal translocations involving the PBX and HOX genes have been discovered in several cancer types, indicating their participation in oncogenic transformation (Morgan *et al.,* 2017). According to studies, the development and aggressiveness of prostate cancer are linked to the deregulation of the PBX and HOX proteins, especially HOXB13 and PBX1 (Shah and Sukumar, 2010). A current topic of study is to identify new therapeutic targets and approaches for cancer therapy by elucidating the specific processes behind the PBX-HOX interaction in cancer.

1.3 HXR9 Peptide:

With the use of a polyarginine sequence, scientists have created the synthetic peptide HXR9, which has 18 amino acids and effectively enters cells by endocytosis (Morgan *et al.,* 2014). By preventing the interaction of the oncoproteins c-Myc and Max, it prevents the creation of the c-Myc-Max complex and has anti-cancer properties (Morgan *et al)*.

HXR9:

WYPWMKKHHRRRRRRRRRR

Figure 1 HXR9 amino acid composition

Prostate, breast, ovarian, lung, and melanoma cancer cells have all shown anti-cancer effects when HXR9 was present. The interaction between HOX transcription factors and their cofactor PBX is disrupted by HXR9.HOX genes play critical roles in embryonic development, cell differentiation, and organogenesis, and their dysregulation is implicated in various cancer types(Shen *et al.,* 2019). HXR9 interferes with the transcriptional activity of HOX genes by preventing the HOX-PBX connection, which can cause cancer cells to undergo apoptosis (programmed cell death, as in below figure) (Shen *et al.,* 2019).

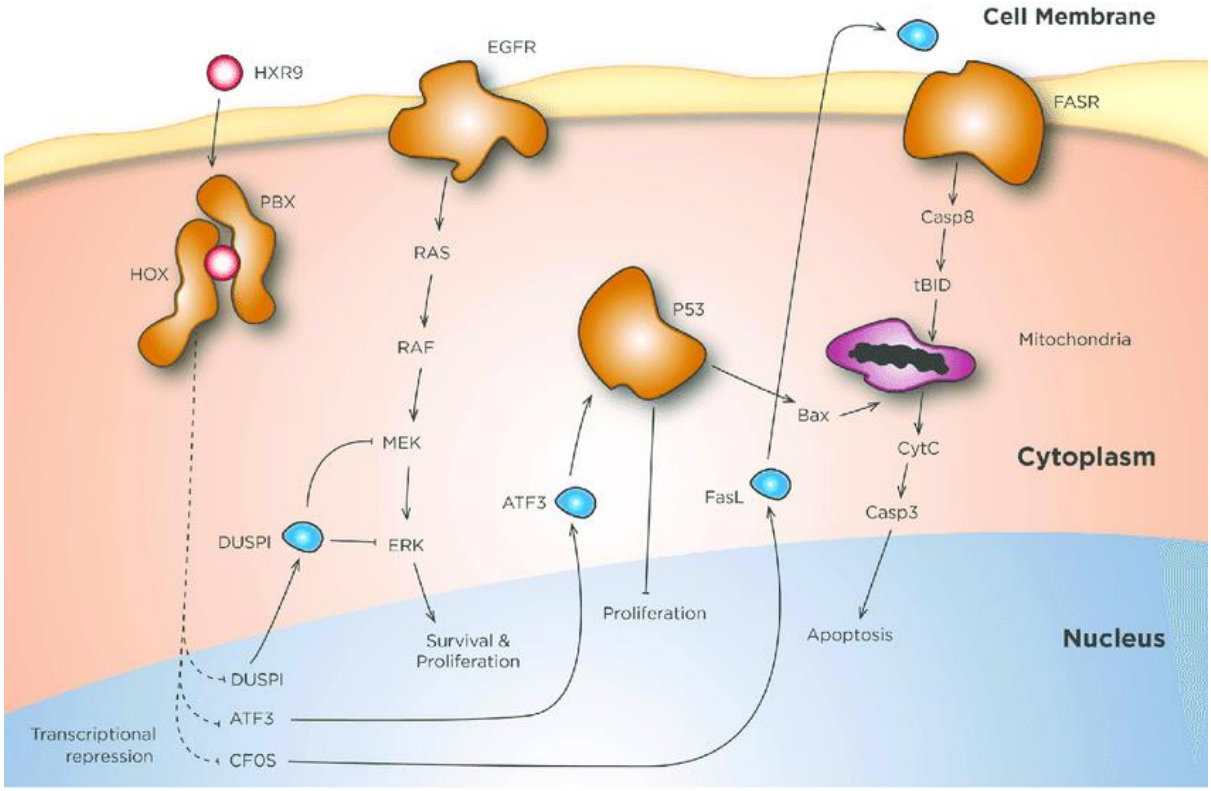


Figure 2 Show HXR9's action mechanisms ©ResearchGate

1.4 KG1 is the cell line that is to be examined in this study.

A well-known human myeloid leukaemia cell line that has been extensively utilised in studies is the KG-1 cell line. It was first created in 1979 using peripheral blood from a patient who had acute myelogenous leukaemia (AML). The myeloblastic appearance of the KG-1 cell line and its capacity to develop into mature granulocytes in response to various stimuli, such as retinoic acid or phorbol esters, are characteristics of this cell line (Drexler *et al.,* 1995).

2. Aim:

The aim of this work is to use Rank Products statistics to pinpoint the genes whose expression is consistently high or low across several prostate cancer cell lines.

The goal is to utilise Rank Products statistics to find inconsistent genes by finding genes that exhibit varying expression levels across several prostate cancer cell lines.

By using Rank Products statistics to analyse the correlation between the expression levels of these genes and the associated HXR9 IC50 values in each cell line, it is intended to better understand the link between consistent genes and HXR9 IC50 values.

The most popular and useful indicator of a drug's effectiveness is its half-maximal inhibitory concentration or IC50. In pharmacological research, it provides a measure of the efficacy of an antagonist medication by indicating the amount needed to block a biological process by 50%.

A statistical technique for locating genes with differential expression is called Rank Products. To determine significance, a rank product (RP) is computed and compared between conditions to compare gene ranks. Using RP values and significance criteria, genes with consistently high or low expression levels are found. Rank Products is resilient, takes into account a wide range of expression distributions, and works with tiny sample numbers. It has been extensively utilised to find recurrent expression patterns in studies on cancer and other topics (Hong *et al.,* 2006).

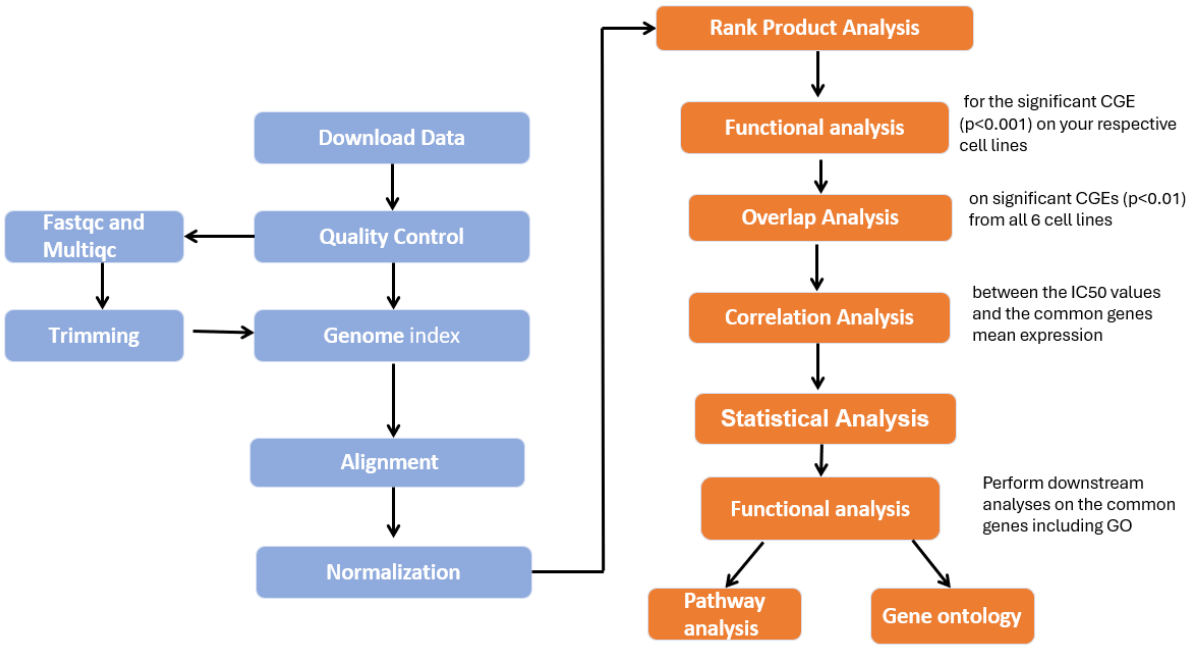


Figure 3 Workflow of the project

3. Methods and Materials

Prior to using the non-parametric statistical approach, Rank Product, to examine the expression levels of both up-and-down-regulated genes in the samples, pre-processing was first performed on the samples. An overlap analysis was then performed to determine distinct signature genes for each cell line. Functional analysis was carried out utilising the gene sets relevant to the discovered gene signatures of each cell line in order to acquire a deeper understanding of the HOX/HXR9 affinity.

3.1 Materials:

This study's raw data was obtained from various kinds of databases, including NCBI-GEO(Gene Expression Omnibus) and ArrayExpress. These databases are well-known repositories for gene expression and microarray data, offering useful details for study and analysis. Indeed, the European Nucleotide Archive, Cellosaurus, UniProt, PubMed, and these are extremely beneficial databases with a wide range of uses in scientific fields. A wide variety of packages and libraries from the Ubuntu, RStudio, and Python environments were used in the study. Snakemake(version 7.26.0) is used to create workflows, which facilitates effective rule-based management of computational pipelines, parallelization, conda integration, and improving reproducibility in research projects.

3.2 Methods:

3.2.1 Downloading of RNA-seq samples:

The Sequence Read Archive (SRA) is used by the National Centre for Biotechnology Information (NCBI) to archive and make high-throughput DNA sequencing data accessible. This covers the metadata associated with sequencing experiments, raw sequencing reads, and alignment information. The SRA Toolkit(version 2.11.3), which provides us with sra files, is used to obtain SRA runs for the KG1 cell line to start the research. The prefetch feature in the SRA Toolkit from the NCBI allows users to download and locally save sequencing data from their servers. The downloaded ".sra" files were then transformed into ".fastq" files using the "fastq-dump" program for further processing. While downloading the data, all the ethical considerations were been followed likes Respect the privacy of individuals, Cite the data source and Use the data for its intended purpose.

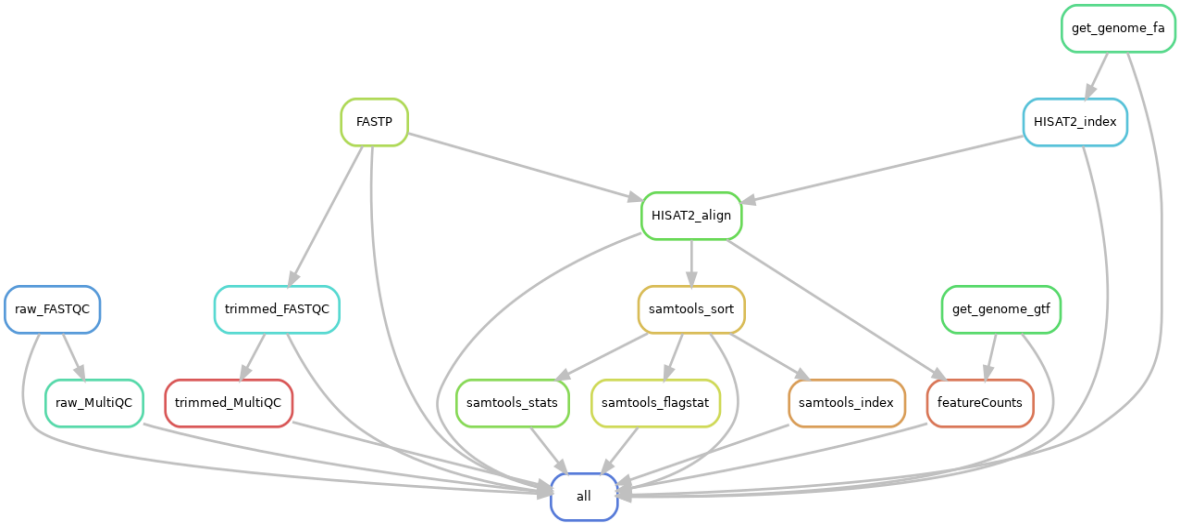


Figure 4 RNA-sequence Pipeline Rulegraph

3.3 Preprocessing

3.3.1 Quality control

FastQC (version 0.12.1) is used in the pipeline's first phase as depicted above in the figure, which involves quality control of the raw RNA-seq samples. Within each cell line, this programme produced a quality report for each unique sample. FastQC offered useful insights into a range of quality indicators, including duplication levels, adaptor content, per-base sequence quality, sequence content, and more. The pipeline uses MultiQC (version 1.12) after FastQC to make data analysis and visualisation simpler. MultiQC created a thorough quality control report for each cell line by combining the various FastQC reports from each sample within a cell line.

After evaluating the initial quality of the RNA-seq samples, pre-processing was done using FASTP (version 0.22.0). The readings were successfully trimmed and verified by FASTP, deleting any remaining adaptor sequences. After that, submitted the trimmed readings to one more round of quality assessment. An aggregated MultiQC report for the cell line was produced as a result of this thorough research, providing a full perspective of the quality metrics. This led to a complete grasp of the integrity and usefulness of the trimmed data for further analysis in the combined report. The workflow was optimised by ensuring the creation of high-quality sequencing data, improving the accuracy and validity of future studies and interpretations.

3.3.2 Alignment:

Aligning sample reads to the reference genome is a critical step in the process of identifying genes and calculating gene expression levels. Utilised HISAT2 (version 2.2.1) to create the required index in order to accomplish this. Ensembl had already provided us with the reference human genome, namely the primary assembly release 106 for the most recent version 38 (GRCh38), which we then uncompressed in our workflow. With the use of the reference genome and the HISAT2 index, it was possible to precisely map the sample reads to their sites on the chromosome, allowing for the identification of gene origins and the measurement of gene expression levels. The foundation for subsequent analysis and interpretation of the RNA-seq data is laid by this crucial alignment phase.

The alignment was completed after getting the trimmed reads with the help of the genome index and HISAT2. The rapid and sensitive alignment strategy used by HISAT2 to provide extremely accurate results is made possible by its hierarchical indexing mechanism (Kim, Langmead and Salzberg, 2015). Samtools version 1.6 was used to convert the aligned readings from the normal SAM (Sequence Alignment/Map) format to the more effective binary format known as BAM (Binary Alignment/Map). The 'view' command, which converts SAM files into BAM files, was used for the conversion. Saving storage space on the computer is a benefit of the conversion to BAM format.

We were able to learn more about the alignment's quality by using Samtools stats and Samtools flagstat functions to produce statistical files from the alignment. The result files showed that every one of our samples had at least eighty-five per cent of their reads mapped to the reference genome, and, astonishingly, none of the samples had any QC-failed reads. As a result, throughout this crucial phase, we did not have to reject or discard any samples. This information demonstrates the high calibre of our

dataset and offers strong assurance that the alignment procedure was effective, laying the groundwork for further analysis and interpretations.

3.3.3 Quantification

We used FeatureCounts version 2.0.3 as the last step in our process for analysing RNA-seq data to measure the levels of gene expression in each sample. Counting the number of reads that accurately mapped to each gene was a vital step in this procedure. The Ensembl Annotation Release 106, which served as the annotation file in .gtf format, and the reads that had previously been aligned to the reference genome were both used to do this. The detailed annotation file included vital details on each gene's chromosomal location and characteristics.

We generated a cell line-specific matrix using FeatureCounts, which we used to methodically count the reads connected to each gene. The abundance of gene expression across all the investigated genes was shown by the raw count for each sample that was displayed in this matrix. This output enables additional downstream analysis, such as differential gene expression, which can provide important details about the underlying biology and regulatory processes in various cell lines. The measured gene expression data derived from this stage provides an essential starting point for further research and identifying genes essential to the biological processes under study.

3.3.4 Normalization

Normalisation is crucial for ensuring the precision and comparability of subsequent analysis in RNA-seq data. We normalised our samples for this purpose using RStudio. This required employing the TMM normalisation approach to translate raw library sizes into scaling factors, which was made possible by the calcNormFactors() function from the edgeR package (version 3.36.0). Using the cpm() function from the same package, we then determined the counts per million (CPM). These normalisation methods allowed us to successfully remove data redundancy and get the data ready for direct expression measure comparisons. This procedure guarantees thorough and trustworthy downstream analysis, enabling insightful analyses and interpretations of the RNA-seq data.

The rma() tool from the Affy package was used to normalise microarray sample data. This elaborate technique included probe data normalisation, background adjustment, and summarization. Effectively removing systematic deviations, adjusting for background noise, and converting unprocessed data into expression values are the functions of rma(). It was made sure that the microarray samples were equivalent and appropriate for further studies by uniting the data in this manner. This process was essential to getting accurate and consistent findings that allowed for accurate interpretation of gene expression patterns in the microarray data.

Using the BioMart software, we enriched the dataset by adding necessary annotations after normalising counts for both RNA-seq and microarray samples. Each sample was annotated by adding pertinent details such as Ensembl gene IDs, gene symbols, and gene types. It was made sure the annotation was appropriate for the particular microarray platform that was utilised with each sample, which is important. By doing this, we were able to create a thorough and accurate annotation of the dataset, allowing for more in-depth analysis and interpretation of the biological importance of the genes discovered in both RNA-seq and microarray investigations. Some microarray probes were duplicated, thus their means were used in place of them.

3.4 Rank Products Analysis:

Breitling et al. developed Rank Product Analysis (RPA) in 2004, an effective statistical technique. Rank Product Analysis is a straightforward, effective technique for finding differentially regulated genes in repeatable microarray data sets. Particularly in the context of microarray data, RPA is intended to identify genes that display consistent differential expression throughout numerous repeated tests. RPA is non-parametric and founded on biological reasoning, unlike conventional approaches that depend on the presumptions of normality and homoscedasticity(Breitling *et al.,* 2004).

The procedure starts with normalising and removing technical variances from the data through preprocessing. Based on the degree of upregulation or downregulation, genes are ranked according to how much their fold varies between experimental conditions. The computation of the rank product (RP) for each gene is the basic concept behind RPA. The rankings of each gene across replicates for each condition are multiplied to produce the RP. RPA reduces the impact of data noise by emphasising the consistency of rankings and selecting genes with consistent and substantial expression changes(Breitling *et al.,* 2004).

A permutation-based method is used to evaluate the importance of the RPs. A null distribution of rank products is produced by random permutations of the data, which enables the estimation of empirical p-values for every gene. Low p-values for genes are regarded as statistically significant, suggesting that changes in their expression are unlikely to be due to chance. For the purpose of adjusting the p-values and preventing false positives, multiple testing correction is used(Breitling *et al.,* 2004).

Small sample sizes, non-normal data distributions, and interdependence between replicates are only a few of the benefits of RPA. It is often used in genomics and bioinformatics to find physiologically significant genes linked to a variety of experimental situations, offering insightful information about underlying molecular processes and prospective treatment targets. Rank Product Analysis is a common option for gene expression analysis in high-throughput research because of its simplicity, robustness, and biological soundness(Breitling *et al.,* 2004).

A normalised count file(KG1) was used for Rank product analysis, which was generated from normalization. Rank Products analysis was performed to identify genes with consistently high expression, consistently low expression, and those displaying inconsistent expression patterns. Analysis was performed in python.

The Rank Product statistical test was conducted using a one-class parameter selection for our data, with a p-value cut-off set at 0.001.The genes that the Rank Products test classified as consistently high, consistently low, or inconsistently have been saved in an Excel files together with the associated statistical data.

3.5 Overlap Analysis:

Using overlap analysis, it is possible to pinpoint the traits or components that different groups or sets of data have in common. It is useful to understand the connections, resemblances and variances between these sets. The primary goal is to measure how

much the sets overlap, which offers useful information on their intersections and uniqueness (*Overlap Analysis.*).

In genomics, overlap analysis compares several gene sets to find shared and distinctive components. Understanding gene interactions, expression patterns, and biological processes are made easier by it. Researchers can identify shared genes between different cell types or situations by examining gene overlaps, making it easier to identify the fundamental genetic reactions. Researchers can identify shared genes between different cell types or situations by examining gene overlaps, making it easier to identify the fundamental genetic reactions. Additionally, overlap analysis might highlight anomalies, such as genes that display both up- and down-regulation at the same time, which may point to problems with the data. In general, overlap analysis in genomics offers useful insights into the dynamics of gene expression, supporting developments in the understanding of illness, the finding of biomarkers, and personalised therapy.

Six distinct cell lines were used for the overlap analysis and each cell line underwent previous rank product analysis. Two separate sets of data were created as a result, one with high consistent genes and the other with low consistent. The overlap analysis led to the identification of common genes, which were then used for further study.

3.6 Correlation analysis:

A statistical method called correlation analysis is used to quantify the magnitude and direction of a linear relationship between two or more variables. Correlation coefficients between -1 and 1 are commonly used to quantify how changes in one variable affect changes in another. An inverse link is implied by a negative correlation, whereas a positive correlation shows that when one variable rises, the other tends to rise as well. Correlation analysis, which is frequently used in disciplines like economics, psychology, and the natural sciences, identifies patterns and dependencies between variables, assisting in understanding linkages and possible predictive links within data sets(Gogtay and Thatte, 2017).

An essential statistical method for examining connections between gene expression patterns and medication response in distinct cell lines is correlation analysis. A matrix with consistent gene expression levels and associated HXR9 IC50 values for each cell line is built in order to carry out this investigation. The intensity and direction of putative correlations are assessed by computing correlation coefficients, such as Spearman's rank correlation, for each consistent gene against the HXR9 IC50 values. This procedure identifies if variations in HXR9 drug sensitivity across several cell lines are related to the expression levels of specific genes. In order to be sure that observed links are statistically relevant rather than just coincidences, statistical tests are also used to evaluate the significance of these correlations. Finally, correlation analysis aids in the discovery of possible biomarkers or molecular components that contribute to differential sensitivity to the HXR9 drug among various cell lines by offering insightful information about the molecular mechanisms behind drug responses.

After the selection of common genes through the overlap analysis, the subsequent step involved extracting the mean expression data from these common genes. In order to identify potential links between gene expression and drug sensitivity, a correlation analysis was performed to examine the connection between the IC50 values and the mean expression levels of the common genes.

3.7 Statistical Analysis:

Statistical analysis is the act of analysing and interpreting data using statistical methods in order to find trends, connections, and important insights. In order to reach well-informed judgements or draw conclusions, it entails using mathematical and computational approaches to data sets. Research studies, experiments, surveys, and other situations requiring data analysis to address research questions or test hypotheses are frequently mentioned in connection with statistical analysis(Witte and Witte, 2017).

Steps for statistical analysis:

1. Selecting the Appropriate Statistical Test: Based on aspects such as data distribution, sample size, and experimental design, select an appropriate statistical test. T-tests, ANOVA, the Wilcoxon rank-sum test, and the Kruskal-Wallis test are available options.
2. Setting Up Null and Alternative Hypotheses: Based on your study topic, develop the alternative hypothesis (H1) and the null hypothesis (H0), both of which assume the absence of any significant difference or link.
3. Conducting the Statistical Test: Use the selected statistical test to assess the correlation between consistent genes and HXR9 IC50 values. Determine the corresponding p-value by calculating the test statistic.
4. Setting the Significance Level: To determine the cutoff for rejecting the null hypothesis, choose a significance level (alpha), often 0.05. Rejecting the null hypothesis in favour of the alternative is appropriate if the resulting p-value is less than the level of significance.
5. Addressing many Comparisons: When examining many genes or doing multiple tests, reduce the possibility of Type I errors by modifying the p-values using techniques like the Bonferroni adjustment or false discovery rate (FDR).
6. Results Interpretation: According to the p-values obtained from the statistical test. Indicative of a substantial correlation between consistent genes and HXR9 IC50 values if p-value alpha. If the p-value exceeds alpha, there is insufficient support for the null hypothesis to be rejected.

3.9 Gene Ontology:

The Gene Ontology (GO) framework is essential in genomics and molecular biology because it offers a standardized language for identifying and classifying genes and their byproducts. GO systematically characterises the functional properties of genes across species using three hierarchical ontologies: biological process, molecular function, and cellular component. GO directs study into the biological importance of genes and their connections to intricate biological processes by assisting researchers in understanding gene function and simplifying the analysis of massive amounts of biological data(Aleksander *et al.,* 2023).

Biological process: The Process of Biology An ontology in GO covers different biological processes or activities in which a gene or gene product takes part in. It emphasises the larger context of gene activity, highlighting the series of activities and interactions that occur inside a cell, organism, or biological system (Gene Ontology Consortium, 2015).

Molecular functions: The molecular functions or activities carried out by gene products are defined by the Molecular Function ontology. It provides an explanation of the molecular interactions, such as catalysis, binding, or structural functions, that a gene product engages in (Gene Ontology Consortium, 2015).

Cellular component: The goal of the Cellular Component ontology is to identify and describe the subcellular components, spaces, or places inside a cell or organism where a gene product is present or active. It describes the bodily setting in which genes act (Gene Ontology Consortium, 2015).

3.10 Gene Set Enrichment Analysis:

Using large-scale gene expression data, Gene Set Enrichment Analysis (GSEA) is carried out to better comprehend the biological importance and underlying processes. It enables scientists to determine if particular gene sets that reflect well-defined biological pathways, functions, or processes are enriched in a given dataset. GSEA offers insights into the biological context of the data and sheds light on the molecular mechanisms causing observed changes by identifying whether pathways or functions are statistically overrepresented(Subramanian *et al.,* 2005).

GSEA is particularly helpful when working with high-throughput gene expression data, such as microarrays or RNA-seq, because the sheer number of genes makes it difficult to evaluate results individually. Instead, GSEA minimises complexity and assists in the discovery of higher-level biological patterns by analysing sets of genes together. This technique is used to uncover physiologically significant pathways, find possible connections to disorders, and produce ideas for more experimental investigation (Subramanian *et al.,* 2005).

To reveal the biological significance of gene expression data, Gene Set Enrichment Analysis (GSEA) entails many essential steps. First, gene expression data is pre-processed and sorted according to a set metric, frequently indicating differential expression between experimental conditions. The study then moves on to carefully chosen gene sets that reflect certain biological pathways, activities, or processes. In order to assess each gene set's overrepresentation at the top or bottom of the sorted gene list, GSEA creates an enrichment score. Permutation testing is used to ascertain statistical significance by determining if the enrichment scores are greater or lower than would be predicted by chance. By highlighting potential routes and changed functionality between conditions, this method helps uncover gene sets linked to the experimental situation. GSEA is a useful technique for illuminating underlying mechanisms and producing theories for further research since it can give biological context from complicated gene expression data(Subramanian *et al.,* 2005).

Enrichment analysis, made possible by tools like "enrich," is a crucial bioinformatics technique used to reveal the biological importance of a group of genes or proteins within a wider genomic context.

Enrichment analysis by enrich:

Enrichment analysis, carried out using tools like "enrich," is a computational approach crucial in revealing the biological significance of a group of genes or proteins within a larger genomic context. It starts with a list of genes that display particular traits, such differential expression or correlation with a certain phenotype. This gene collection is contrasted with a background dataset that generally represents the complete genome to find statistically enriched biological words, pathways, or Gene Ontology categories. The significance of this enrichment is evaluated using statistical tests, such as hypergeometric or Fisher's exact tests. Researchers may determine which phrases are most pertinent to their dataset by looking at the findings, which include enriched terms and any related statistical scores, such as p-values and q-values. Enrichment analysis is a useful method in the study of genomics and systems biology because it helps reveal the functional context of experimental results, illuminates underlying biological processes, and directs subsequent research (Khatri, Sirota and Butte, 2012).

4. Results

4.1 Rank Product analysis :

Using rank product analysis, the samples were analysed to examine whether any genes were consistently high and low expressed for each cell line, including my cell line KG1. There are six RNA samples in KG1 cell line, which are give below.

ERR3003550
GSM5114211
GSM5114212
GSM1937937
GSM1937938
SRR8615910

Figure 5 RNA Samples of KG1 cell line which used in this study

Table 1 Results from rank products analysis

RANK PRODUCTS ANALYSIS	HEL92.1.7	K-562	KG1	MONO-MAC-6 (MM6)	A375M	KU812F
NUMBER OF SAMPLES	4	26	6	9	3	6
TOTAL GENES	31013	44505	38225	54059	23721	35731
HIGH CONSISTENT GENES	1215	7798	2412	5097	637	2318
	3.91%	17.52%	6.31%	9.42%	1.94%	6.48%
	82	1790	207	448	14	307

LOW CONSISTENT GENES	0.28%	4.03%	0.55%	0.84%	0.81%	0.86%
INCONSISTENT GENES	29716	34917	35605	48513	23069	33105
	95.81%	78.45%	93.14%	89.74%	97.25%	92.65%

Six distinct samples were examined for gene expression as part of the examination for the KG1 cell line. A total of 38,225 genes were evaluated in this examination to see how they were expressed in the KG1 cell line. 2,412 genes in the KG1 cell line were consistently expressed at a high level in 6 samples, while 207 genes were consistently expressed at a low level. These gene groupings make up approximately 6.31% and 0.55%, respectively, of all the genes examined for KG1.

The results table shows that K562 has a significantly greater proportion of important genes discovered by the rank products analysis. The large number of samples a total of 26 RNA-seq samples that are readily available for this particular cell line may be credited with this finding. Comparatively, the sample sizes for the remaining cell lines, which range from 3 to 9, are much smaller.

4.2 Overlap analysis:

4.2.1 High consistent genes:

```
Cell Line: HCG_HEL92.1.7.csv
Common genes with other lines: 75
Cell Line: HCG_K-562_finalCounts.csv
Common genes with other lines: 75
Cell Line: HCG_KG1.1_finalCounts.csv
Common genes with other lines: 75
Cell Line: HCG_MM6_finalCounts.csv
Common genes with other lines: 75
Cell Line: HCG_A375M_finalCounts.csv
Common genes with other lines: 75
Cell Line: HCG_KU812_finalCounts.csv
Common genes with other lines: 75
```

Figure 6 Results of Overlap analysis for high consistent genes

As shown in above figure, there are 75 genes that are shared by all cell lines. Meanwhile, the number of unique genes varies between each cell line, from 562 in HCG_A375M to 7723 in HCG_K-562, and is not shared by the others. These unique genes imply that each cell line may have specific genetic features or activities. The level of genetic variety and the distinctive genes that are unique to each cell line could be a sign of their distinctive behaviours, reactions, and possible functions in biological processes.

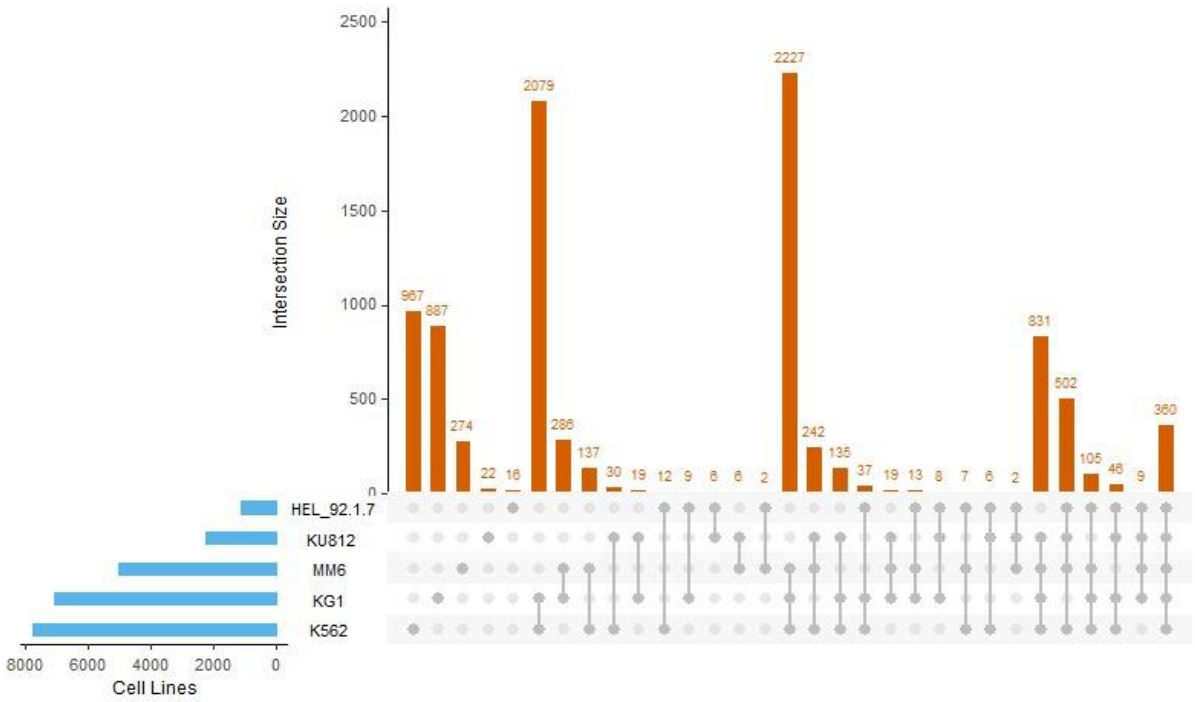


Figure 7 Upset plot for High consistent Genes

In genomics and data analysis, upSet plots are useful for highlighting similarities and differences between gene sets or medical problems. They are used by researchers to investigate both common and distinctive traits, assisting in the discovery of key genes or patterns.

Figure 6 shows that certain genes display similarity among the various cell lines in addition to the set of 75 genes that are shared by all cell lines. It is clear that the KG1 cell line and other cell lines like K562 and MM6 share certain genes.

4.2.2 Low consistent genes:

```
Cell Line: LCG_HEL92.1.7.csv
Common genes with other lines: 0
Cell Line: LCG_K-562_finalCounts.csv
Common genes with other lines: 0
Cell Line: LCG_KG1.1_finalCounts.csv
Common genes with other lines: 0
Cell Line: LCG_MM6_finalCounts.csv
Common genes with other lines: 0
Cell Line: LCG_A375M_finalCounts.csv
Common genes with other lines: 0
Cell Line: LCG_KU812_finalCounts.csv
Common genes with other lines: 0
```

Figure 8 Results of Overlap analysis for low consistent genes

As illustrated in Figure 7, an analysis of six distinct cell lines reveals that none of them exhibit any shared genes, suggesting that each cell line possesses its own unique

genetic characteristics.

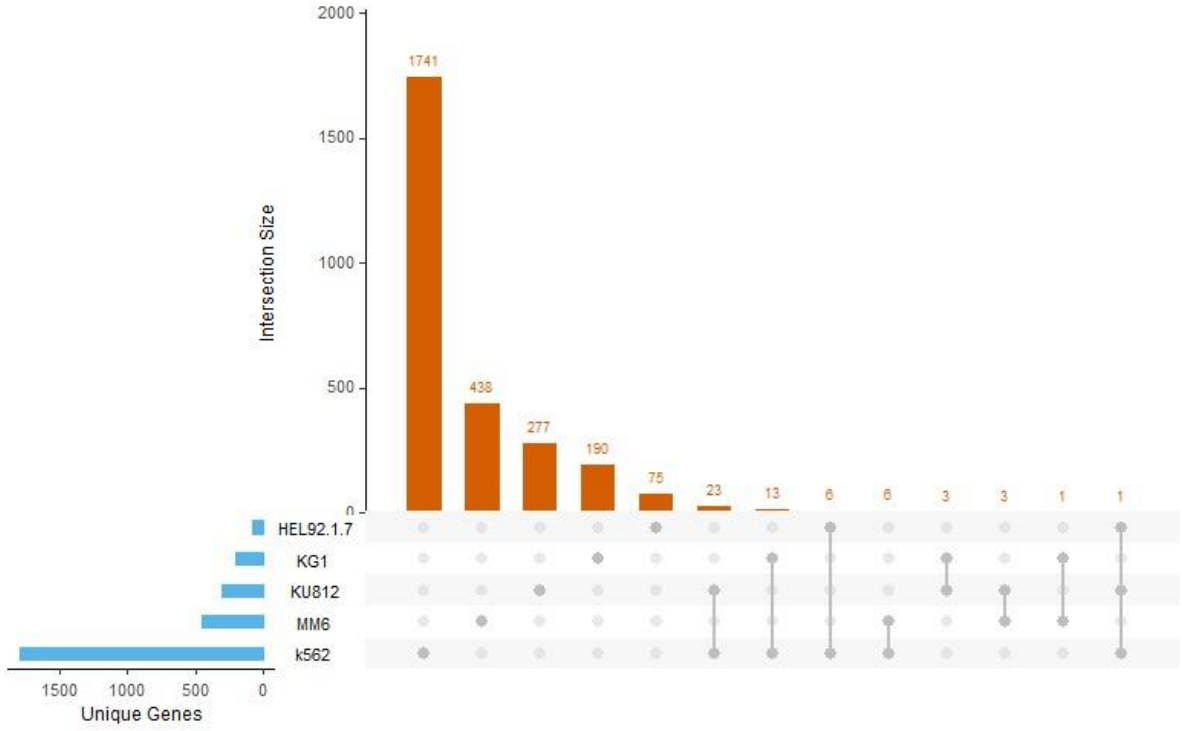


Figure 9 Upset plot for Low Consistent Genes

In contrast to the other five cell lines, the A375M cell line exhibits a unique pattern in Figure 8 where it does not share any low-consistency genes. It's interesting to note that there are far fewer genes that have low consistency than with high consistency. However, the KG1 cell line, which shares genes with K562 and MM6, exhibits behaviours shown in the high-consistency genes (Melus *et al.,* 2021).

4.3 Statistical Analysis:

A two-sample t-test with an alpha level of 0.05 was used in this investigation to examine whether there was a significant relationship between IC50 values and gene expression. The estimated p-value was compared to alpha, which indicates statistical significance, and this comparison produced a clear result: the null hypothesis was rejected if the calculated p-value was smaller than alpha, indicating statistical significance. In contrast, if the p-value was less than alpha, there was no evidence of a significant difference between the two datasets, and the null hypothesis could not be disproved. This extensive statistical analysis offers critical insights into the potential correlations or discrepancies between gene expression and IC50 values.

4.4 Correlation Analysis:

For this investigation, we particularly used Pearson correlation coefficient to determine the association between IC50 values and the mean expression of common genes.

The linear link between two continuous variables is measured by the Pearson correlation. The degree and direction of the linear relationship between two variables is quantified. From -1 to 1, the correlation values are available.

1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship

The correlation values are color-coded on the heatmap to make it simple to see the strength and direction of the associations between the variables.

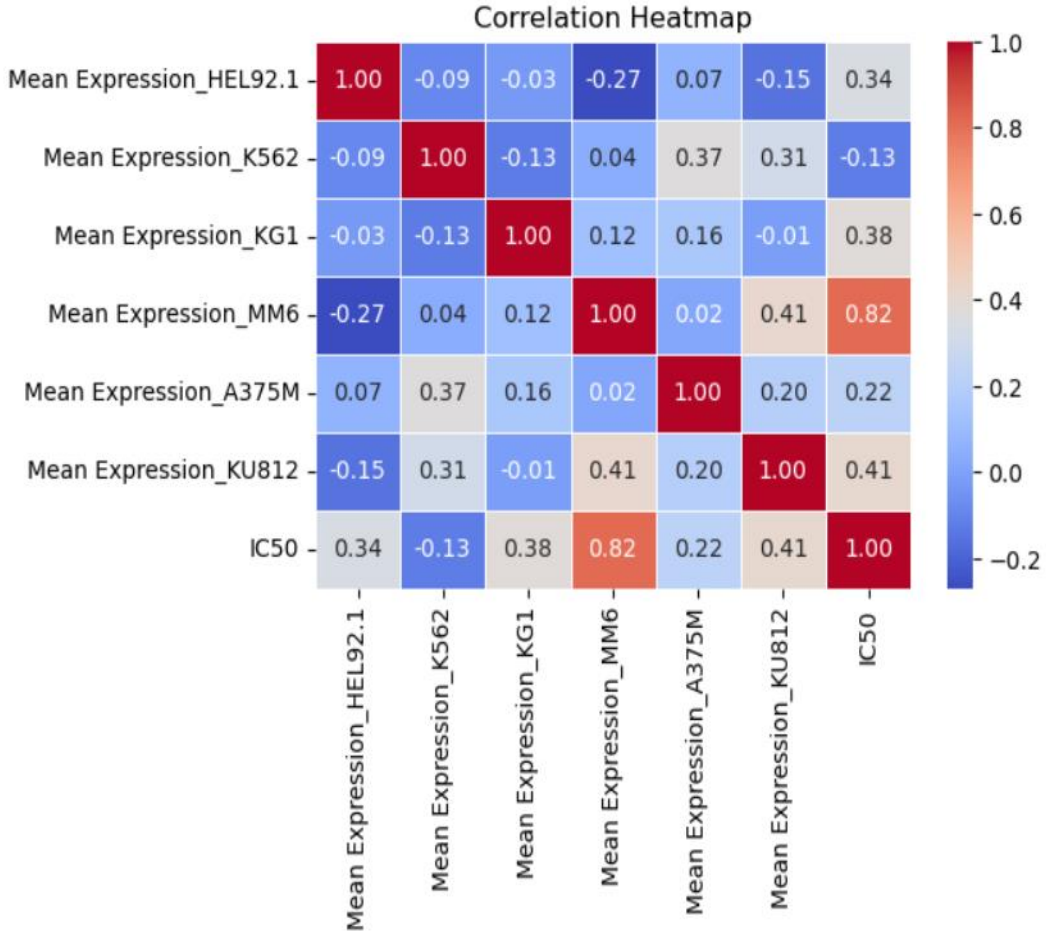


Figure 10 Heatmap of correlation analysis between IC50 values vs Six Cell lines

The correlation heatmap sheds light on the relationships between the mean expression of common genes across different cell lines and the IC50 values, which represent drug sensitivity. Mean Expression_MM6 and IC50 have a significant positive correlation (about 0.815), indicating that the MM6 cell line's gene expression and drug responsiveness are strongly correlated. Additionally, IC50 and Mean Expression_HEL92.1, Mean Expression_K562, Mean Expression_A375M, and Mean Expression_KU812 show weaker but unique associations, offering information on how gene expression may possibly affect drug sensitivity across different cell lines.

Table 2 Result of Correlation Analysis showing six cell lines correlation coefficient

Cell Lines	Correlation Strength
Mean Expression_HEL92.1.7	Moderate Positive (Approx 0.338)
Mean Expression_K562	Weak Negative (Approx -0.125)
Mean Expression_KG1	Moderate Positive (Approx 0.382)
Mean Expression_MM6	Strong Positive (Approx 0.815)
Mean Expression_A375M	Weak Positive (Approx 0.224)
Mean Expression_KU812	Moderate Positive (Approx 0.410)

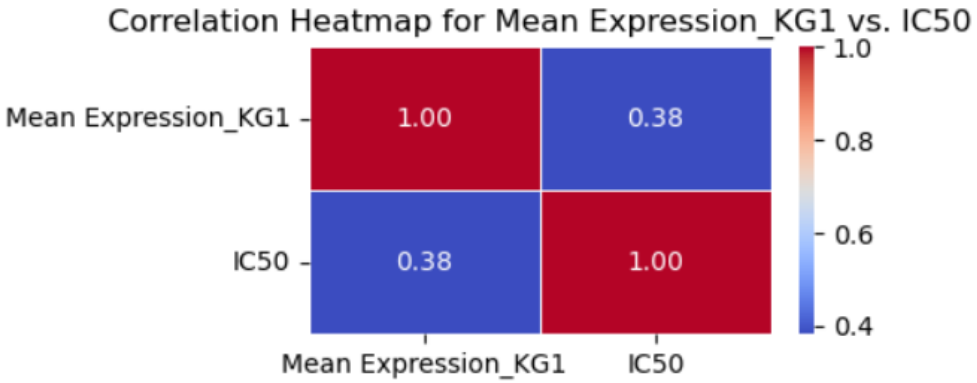


Figure 11 Heatmap of correlation analysis between IC50 values vs KG1 Cell line

IC50 and Mean Expression_KG1 show a moderate positive link with a correlation coefficient of around 0.382. This indicates that the Mean Expression_KG1 tends to grow along with an increase in the IC50 values (which get greater), and vice versa. This moderately positive association raises the potential of a connection between the IC50 response to therapy and the particular gene expression patterns of the KG1 cell line.

4.5 Gene ontology(GO) and Gene Set Enrichment Analysis(GSEA) of Consistent Genes

4.5.1 Gene Ontology (GO) High consistent gene:

Table 3 RESULT TABLE OF GENE ONTOLOGY OF BIOLOGICAL PROCESS OF HIGH CONSISTENT

ID	Description	GeneRatio
GO:0097305	Response to alcohol	87/1635

GO:1903522	Regulation of blood circulation	88/1635
GO:0009410	Response to xenobiotic stimulus	119/1635
GO:0050878	Regulation of body fluid levels	105/1635
GO:0035296	Regulation of tube diameter	58/1635
GO:0097746	blood vessel diameter maintenance	58/1635
GO:0035150	Regulation of tube size	58/1635
GO:0120254	Olefinic compound metabolic process	61/1635
GO:0006936	Muscle contraction	94/1635
GO:0006631	Fatty acid metabolic process	101/1635

The above table depict the result of biological process of High consistent genes(HCG) which generated from gene ontology process. Here, the "ID" indicates a special identification code given to each GO term. These codes serve as standardised identifiers for particular biological process or functions. In the context of Gene Ontology (GO) analysis, the "Description" column offers textual labels or descriptions for each GO ids. These explanations are meant to illustrate the biological processes. The "GeneRatio" value normally denotes the ratio between the number of genes in a dataset or background set that are associated with a given Gene Ontology (GO) word and the total number of genes in the dataset.

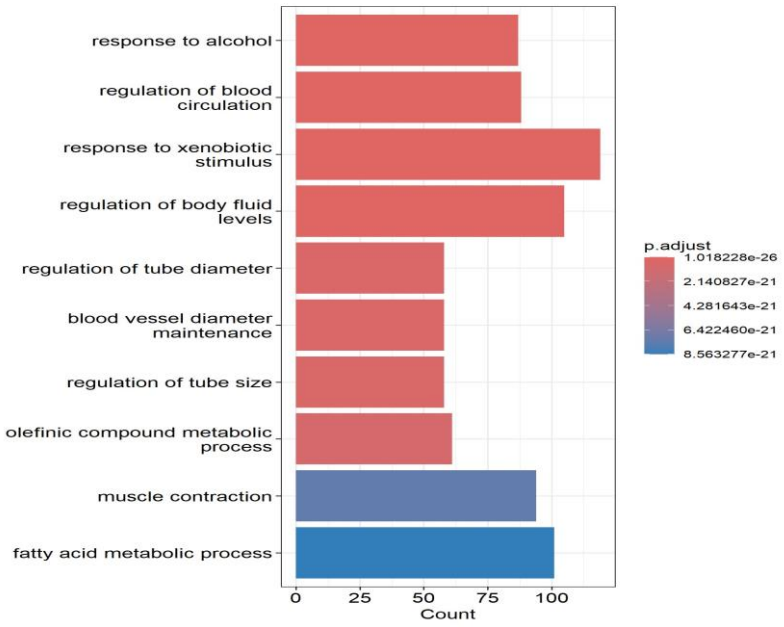


Figure 12 Bar plot of the Gene Ontology (GO) focusing on biological processes of HCG

The outcomes of a Gene Ontology (GO) concentrating on biological processes (BP) are shown in this barplot. The analysis's objective is to locate and classify particular

biological processes that are statistically important in a given dataset. Based on the analysis, the top 10 enriched biological processes are shown in the barplot in Y-axis . Gene counts are represented on the X-axis. The relevant p-values, which denote the statistical significance of each biological process within the dataset, are seen on the right side of the figure. Greater importance is suggested by lower p-values. The importance of p-values in the context of your barplot and Gene Ontology (GO) study is described as "lower p-values suggest greater significance. “Which we can see in figure.10

The bar plot of Gene Ontology (GO) enrichment results reveals more than 100 genes show significant biological processes linked to the "response to xenobiotic stimulus." The term "xenobiotic stimulus" describes the introduction of synthetic or foreign substances, such as pollutants, chemicals, or medications, into the environment of an organism that might cause biological reactions (Duan *et al.,* 2017). This indicates a diverse range of functions and pathways involved in an organism's reaction to foreign substances or chemicals.

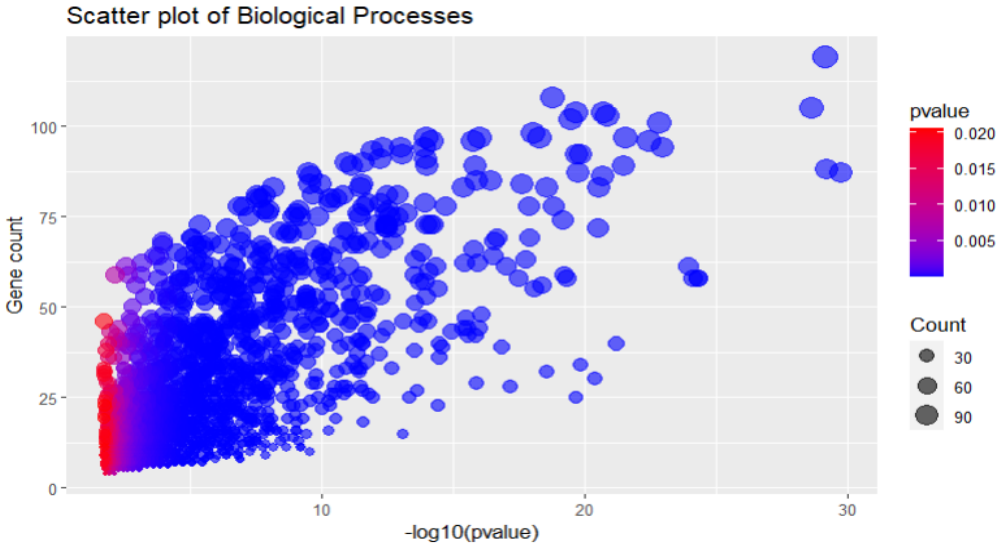


Figure 13 Scatter plot of the Gene Ontology (GO) focusing on biological processes of HCG

This scatter plot depicts Gene Ontology (GO) results for Biological Processes. Each point represents a specific Biological Process category, with the x-axis indicating the significance of enrichment (-log10(p-value)) and the y-axis displaying the gene count associated with each category. Larger points denote more genes associated with a category, while colour signifies the p-value significance, transitioning from blue (less significant) to red (more significant). The plot quickly identifies highly significant Biological Process categories with numerous associated genes, aiding in the identification of biologically relevant processes in the analyzed dataset.

4.5.2 Gene Set Enrichment Analysis(GSEA) High consistent gene:

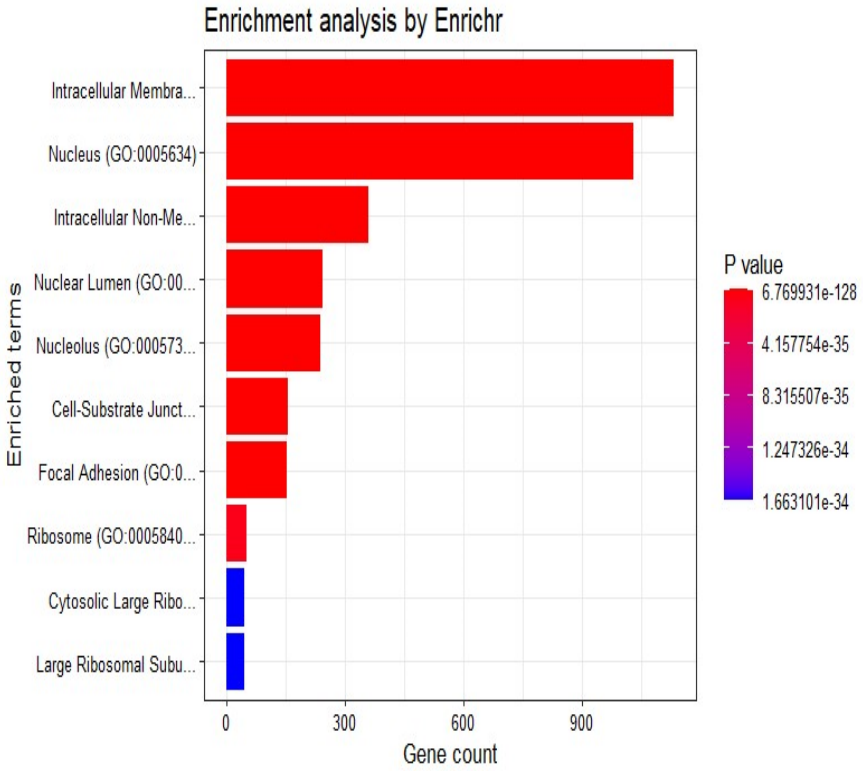


Figure 14 Barplot of Enrichment analysis by Enrichr of high consistent gene (HCG)

Figure 11 depicts the result of enrichment analysis in form of bar plot. The top 10 enriched terms from your enrichment analysis are displayed in the graphic in order of their statistical significance (p-values).

Enriched terms (Left Side): Biological categories or processes that are discovered to be overrepresented in your input gene or protein set when compared to a background set are represented by the words listed on the left side. Each phrase sheds light on prospective uses or connections made by your dataset.

Gene Count (Below): How many genes from your input set are connected to each enriched term is shown by the number of genes provided below it. This total reflects the observed number of genes associated with the word in your dataset.

P-Values (Right Side): The p-values on the right side show the enrichment's statistical significance. They determine the probability that each term's observed gene count might have been discovered by pure chance. A lower p-value denotes a more significant statistical result. A low p-value (e.g., 0.05) in an enrichment study often denotes that the link between your gene set and the phrase is unlikely to have happened by coincidence.

4.5.3 Gene Ontology (GO) of Low consistent genes(LCG):

Table 4 RESULT TABLE OF GENE ONTOLOGY OF BIOLOGICAL PROCESS OF LOW CONSISTENT GENE

ID	Description	GeneRatio
GO:0007189	Adenylate cyclase-activating G protein-coupled receptor signaling pathway	22/143
GO:0007188	Adenylate cyclase-modulating G protein-coupled receptor signaling pathway	24/143
GO:0035296	Regulation of tube diameter	18/143
GO:0097746	Blood vessel diameter maintenance	18/143
GO:0035150	Regulation of tube size	18/143
GO:0009152	Purine ribonucleotide biosynthetic process	20/143
GO:0006171	cAMP biosynthetic process	8/143
GO:0006164	Purine nucleotide biosynthetic process	21/143
GO:0009260	Ribonucleotide biosynthetic process	20/143
GO:0072522	Purine-containing compound biosynthetic process	21/143

The findings of a gene ontology study for biological processes are shown in the above table 4, which includes a total of 50 genes. In this study, the low consistent gene linked to numerous biological processes are identified, and their distribution across multiple cellular components and activities is shown.

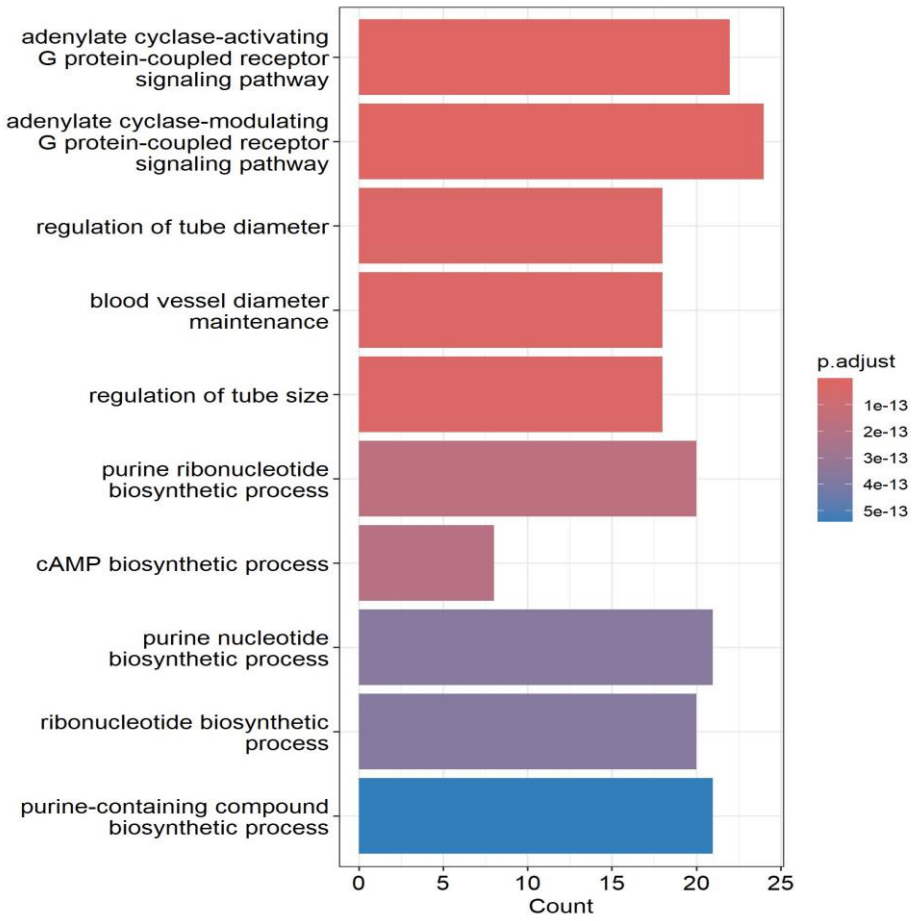


Figure 15 Bar plot of the Gene Ontology (GO) focusing on biological processes of LCG

The barplot provided represents the results of a Gene Ontology (GO) analysis focused on Biological Process (BP) of LCG. It visually displays the findings related to biological processes. Based on the analysis, the barplot illustrates the top 10 enriched biological processes as identified through the Gene Ontology (GO) analysis. The bar plot's p-values for each enriched biological process denote the statistical significance of that process's enrichment within the studied dataset.

The "adenylate cyclase modulating G protein-coupled receptor signaling pathway" involves the regulation of more than 20 genes. These genes collectively orchestrate the complex process of signal transduction initiated by G protein-coupled receptors (GPCRs) and mediated through adenylate cyclase. Within this intricate network, these genes play essential roles in ensuring the accurate transmission of extracellular signals into intracellular responses (El Zein, Badran and Sariban, 2008).

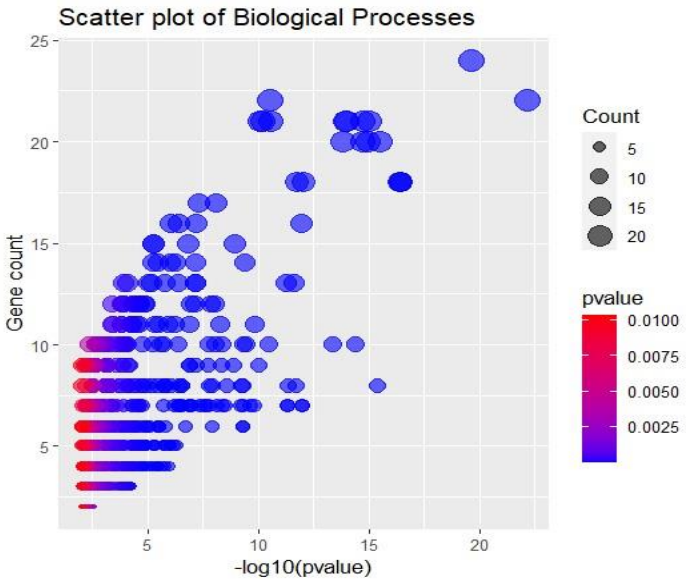


Figure 16 Scatter plot of the Gene Ontology (GO) focusing on biological processes of LCG

This scatter plot visualizes Gene Ontology (GO) enrichment results for Biological Processes. Points represent specific GO terms. X-axis shows significance (-log10(p-value)), y-axis displays gene count, and colour indicates p-value significance (blue to red). It identifies significant processes with many associated genes, aiding in identifying biologically relevant terms.

4.5.5 Gene Set Enrichment Analysis(GSEA) Low consistent gene:

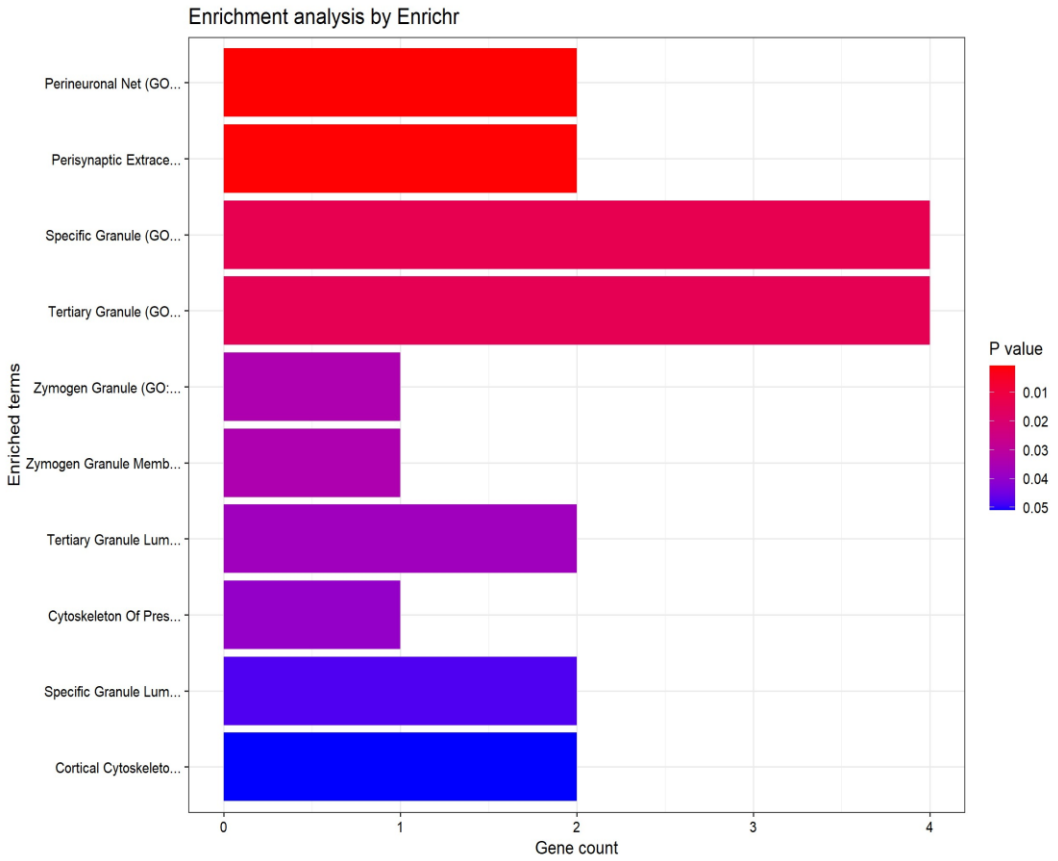


Figure 17 Barplot of Enrichment analysis by Enrichr of low consistent gene (LCG)

This bar plot represents the outcome of an enrichment analysis, showcasing the top 10 significantly enriched terms. These terms have been arranged in descending order of their statistical significance, as indicated by their respective p-values. This graphical representation provides a clear visual insight into the most statistically significant findings from the enrichment analysis.

4.6 Gene ontology(GO) of common genes(KG1)

4.6.1 Biological Process(BP):

Table 5 RESULT TABLE OF GENE ONTOLOGY OF BIOLOGICAL PROCESS

ID	Description	GeneRatio
GO:0006635	Fatty acid beta-oxidation	9/50
GO:0009062	Fatty acid catabolic process	9/50
GO:0019395	Fatty acid oxidation	9/50
GO:0034440	Lipid oxidation	9/50
GO:0016054	Organic acid catabolic process	11/50
GO:0046395	Carboxylic acid catabolic process	11/50
GO:0072329	Monocarboxylic acid catabolic process	9/50
GO:0030258	Lipid modification	10/50
GO:0044242	Cellular lipid catabolic process	10/50
GO:0006631	Fatty acid metabolic process	12/50

The results of a gene ontology investigation focusing on biological processes are presented in the provided Table 5, encompassing a total of 50 genes. This study reveals the presence of common genes with associated with many biological processes. The table illustrates the distribution of these genes across various cellular components and functional activities.

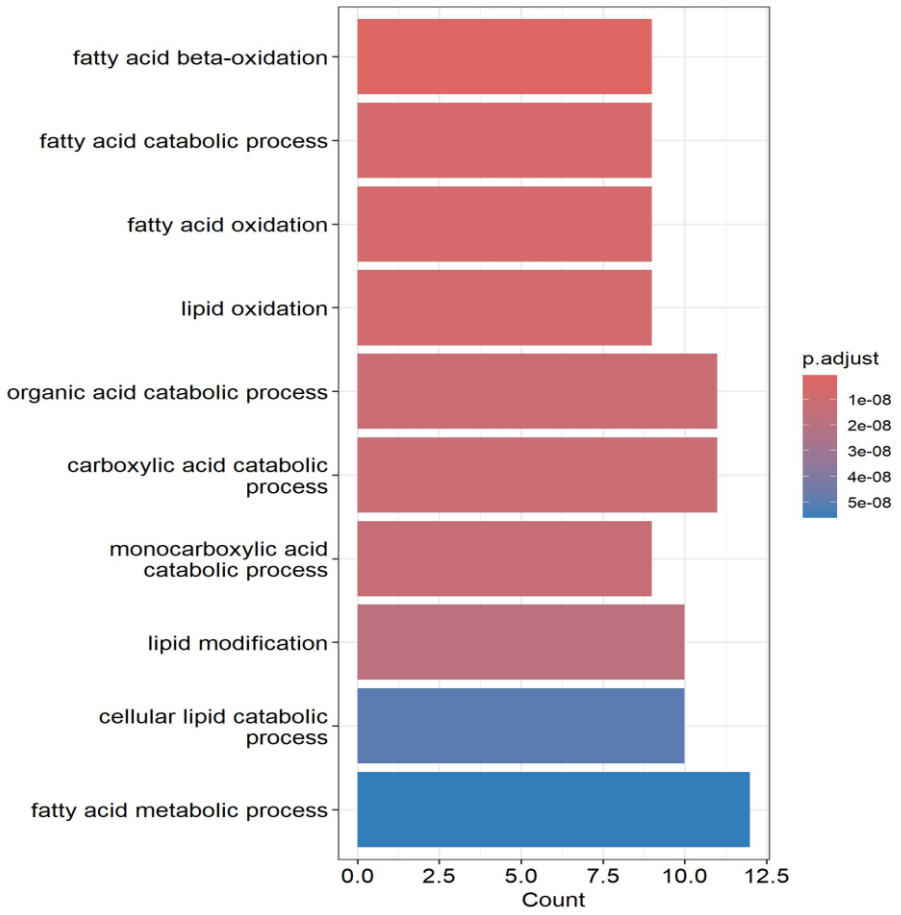


Figure 18 barplot visualizes the results of a Gene Ontology (GO) focusing on biological processes (BP)

Results from a Gene Ontology (GO) analysis with an emphasis on Biological Processes (BP) are shown in the barplot. With accompanying p-values reflecting their statistical significance within the dataset, it displays the top 10 biological processes discovered throughout the investigation.

4.6.2 Cellular Component (CC):

Table 6:RESULT TABLE OF GENE ONTOLOGY OF CELLULAR COMPONENT

ID	Description	GeneRatio
GO:0072562	blood microparticle	8/50
GO:0030175	filopodium	5/50
GO:0005759	mitochondrial matrix	8/50
GO:0015629	actin cytoskeleton	8/50
GO:0030027	lamellipodium	5/50
GO:0034774	secretory granule lumen	6/50
GO:0060205	cytoplasmic vesicle lumen	6/50
GO:0031983	vesicle lumen	6/50
GO:0098858	actin-based cell projection	5/50
GO:0099571	postsynaptic cytoskeleton	2/50

The table presents the results of a gene ontology analysis for cellular components, showing a total of 50 genes. This analysis identifies the common genes associated with various cellular components and provides insights into their distribution across different cellular components and functions.

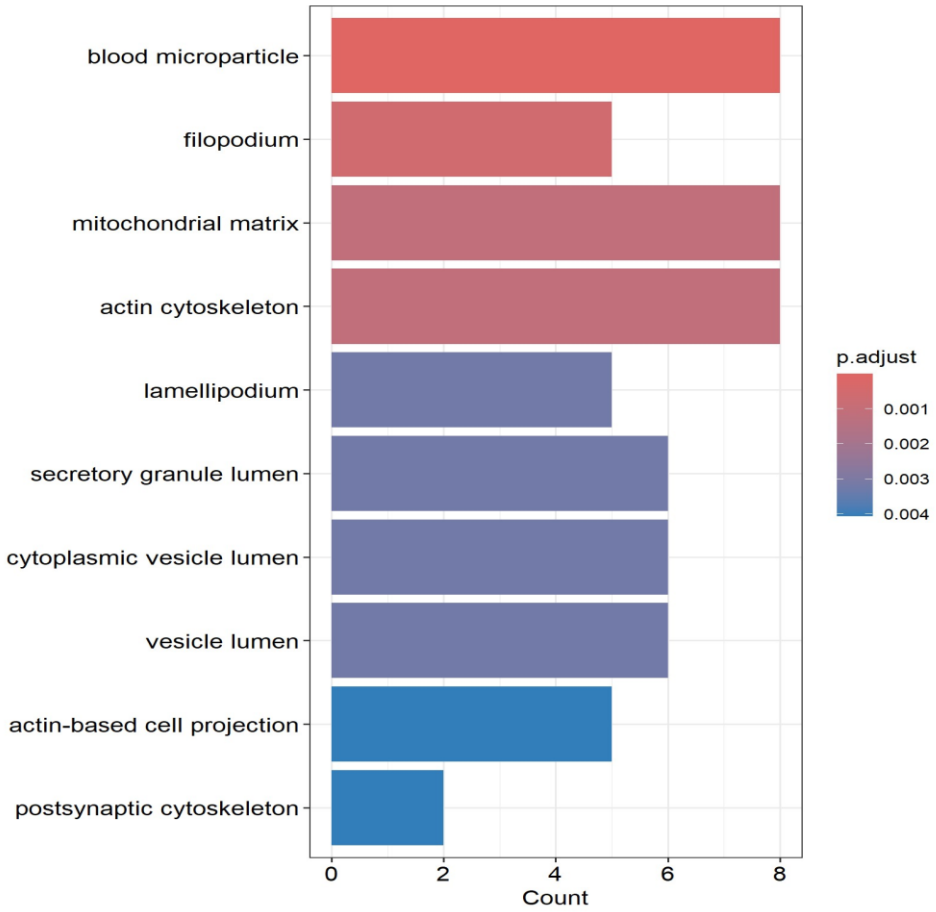


Figure 19 barplot visualizes the results of a Gene Ontology (GO) focusing on cellular component(CC)

The barplot provided represents the results of a Gene Ontology (GO) analysis focused on Cellular Component (CC). It visually displays the findings related to cellular components. Based on the analysis, the barplot illustrates the top 10 enriched cellular components as identified through the Gene Ontology (GO) analysis. The bar plot's p-values for each enriched cellular component denote the statistical significance of that component's enrichment within the studied dataset.

4.6.3 Molecular Functions(MF):

Table 7 RESULT TABLE OF GENE ONTOLOGY OF Molecular Function (MF)

ID	Description	GeneRatio
GO:0003995	Acyl-CoA dehydrogenase activity	5/50
GO:0052890	Oxidoreductase activity, acting on the CH-CH group of donors, with a flavin as acceptor	5/50
GO:0016627	Oxidoreductase activity, acting on the CH-CH group of donors	7/50
GO:0050660	flavin adenine dinucleotide binding	6/50
GO:0016407	acetyltransferase activity	6/50
GO:0140359	ABC-type transporter activity	5/50
GO:0140326	ATPase-coupled intramembrane lipid transporter activity	4/50
GO:0016634	Oxidoreductase activity, acting on the CH-CH group of donors, oxygen as acceptor	3/50
GO:0140303	Intramembrane lipid transporter activity	4/50
GO:0016746	Acyltransferase activity	7/50

The table displays the outcomes of a gene ontology study for molecular functions, illuminating the activities of 50 genes in total. This study reveals the common genes linked to many molecular functions, providing helpful insights into how these genes affect various Molecular Functions and activities.

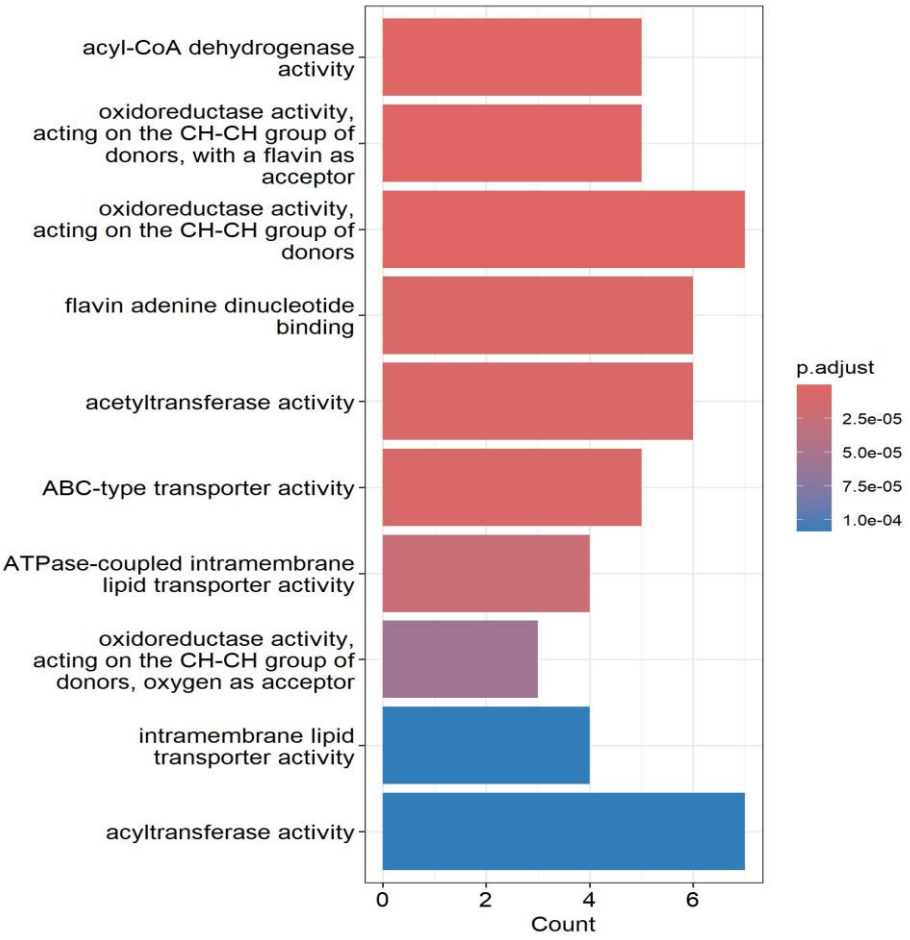


Figure 20 barplot visualizes the results of a Gene Ontology (GO) focusing on Molecular Function(MF)

The barplot represents the results of a Gene Ontology (GO) analysis focusing on Molecular Function (MF). It shows the top 10 enriched molecular functions identified in the analysis. The p-values associated with each function indicate how statistically significant their presence is in the dataset.

4.6.4 Gene Set Enrichment Analysis(GSEA)

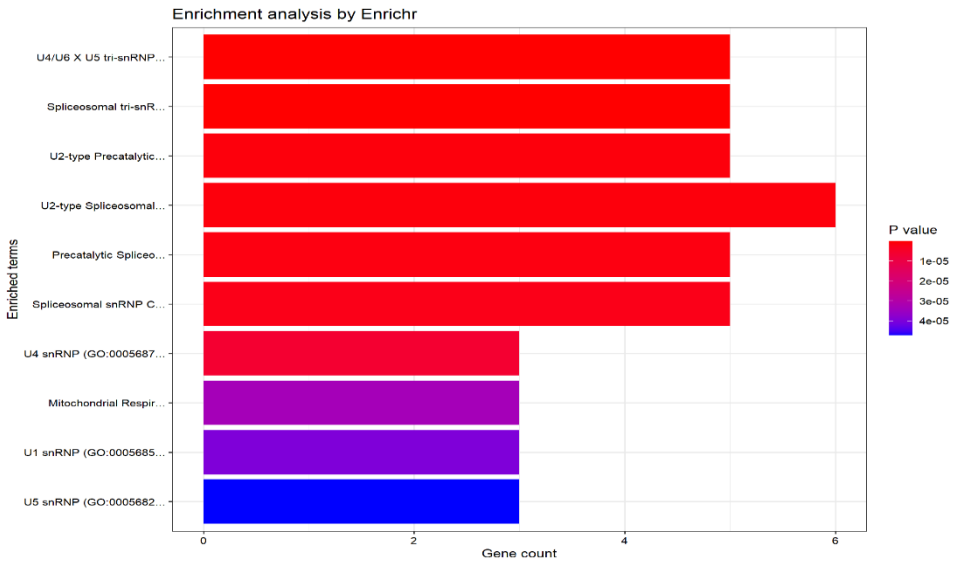


Figure 21 Barplot of Enrichment analysis by Enrichr of Common Genes

This bar chart illustrates the results of an enrichment analysis, highlighting the top 10 markedly enriched terms. These terms have been organized in descending order based on their statistical significance, as denoted by their corresponding p-values. This visual representation offers a concise and intuitive view of the most statistically significant outcomes derived from the enrichment analysis.

5. Discussion

In the initial phase of our individual analysis, it was initiated by the Normalizing the raw data collected from the specific cell line assigned to each of us. In this project, the designated cell line under investigation was kG1. "After data normalisation, the Rank Product approach was utilized to perform Consistent Gene analysis on our normalised data. On our dataset, the Rank Product statistical test was run with a one-class parameter selection, and a strict p-value threshold of 0.001 was used to determine significance. These significant genes were consistently observed with both high and low expression levels, meeting the stringent p-value threshold in our individual analyses. These genes have important biological significance and may be essential for diverse cellular functions and gene control. After completing the Consistent Gene Analysis (CGA) on High Consistent Genes (HCG) and Low Consistent Genes (LCG), Gene Ontology (GO) and pathway studies were carried out on the genes associated with the significant findings from the CGA analysis. This analysis aimed to elucidate the biological processes and enriched pathways within the set of genes high consistent and low consistent associated with significant CGA.

Using a significance threshold (alpha) of 0.05, a two-sample t-test was performed to examine the discrepancy between gene expression and IC50 values. Indicating a substantial difference between the datasets, the analysis produced a p-value that was lower than alpha.

Moving to the group analysis, the data of the genes collected, that were consistently high consistency and low consistency across all six cell lines, taking into account the data that was provided by other group members. After that an overlap analysis was conducted on significant CGA genes(HCG and LCG) ($p<0.01$) from all six cell lines. This analysis revealed a set of genes that consistently exhibited differential patterns

across various cell lines, including both High Consistence Genes(HCG) and Low Consistence Genes(LCG). "As a result, we identified a total of 75 common genes that were consistently present in the high-consistency group (HCG) across all six cell lines. However, there were no genes found to be common in the low-consistency group (LCG) across these cell lines. Then next extracted the mean expression from this collection of genes after using the overlap analysis to identify the common genes from HGC.

We performed a correlation study to look into the connection between the IC50 values and the mean expression levels of these common genes in order to investigate potential correlations between gene expression and drug sensitivity. Figure 8 clearly displays the outcomes, and Table 2 provides a summary of them.

In the last phase of our analysis, we delved into various analyses focused on the common genes within the kG1 cell line that we identified through the overlap analysis. This involved a more detailed exploration of their functional roles and implications. we performed further Gene Ontology (GO) and GSEA(Gene Set Enrichment Analysis) research with a focus on the common genes to learn more about the biological processes and functions they are involved in.

Gene Ontology (GO) serves three key functions: Biological Function(BF), Molecular Function(MF) and Cellular Component(CC).

GSEA was carried out using the enrich tool (library version 3.2) in R to uncover how gene sets enrich specific pathways, providing insights into the broader biological context in which these shared genes function. This analysis played a crucial role in revealing whether specific pathways or networks were significantly influenced by the common genes we identified in our study.

6. Conclusion

Firstly, subsequent Rank Product analysis, with a stringent p-value threshold of 0.001, identified genes consistently expressed at both high-consistent (HCG) and low-consistent (LCG) groups, indicating their crucial roles in diverse cellular functions and gene regulation.

The GO(Gene Ontology of HCG) enrichment analysis highlights over 100 genes associated with significant biological processes related to "response to xenobiotic stimulus," which involves the organism's reactions to synthetic or foreign substances like pollutants, chemicals, or medications, potentially triggering biological responses(Duan et al,2017).

In GO(Gene Ontology of LCG),the "adenylate cyclase modulating G protein-coupled receptor signalling pathway" governs over 20 genes that collaborate to regulate the intricate process of signal transduction initiated by G protein-coupled receptors (GPCRs) and mediated through adenylate cyclase. These genes are pivotal in ensuring the precise conversion of extracellular signals into intracellular responses(El Zein, Badran, and Sariban 2008).

In the enrichment analysis, the High-Consistent Genes (HCG) were notably linked to intracellular membranes, underscoring their importance in cellular structures and processes. On the other hand, the Low-Consistent Genes (LCG) exhibited associations with specific granules and tertiary granules, suggesting specialized roles in the formation or regulation of these cellular structures.

The null hypothesis was thus rejected, indicating a significant difference between gene expression and IC50 values.

According to overlap analysis analysis of High-Consistent Genes (HCG) showed that all six cell lines share 75 common genes, indicating shared genetic features and potential functions. In contrast, among Low-Consistent Genes (LCG), no genes were shared among the cell lines, emphasizing the unique genetic characteristics of each cell line in this group. These findings underscore the distinct genetic profiles and potential functional differences between high and low-consistency genes in these cell lines.

The correlation analysis between IC50 values, representing drug sensitivity, and mean gene expression across different cell lines uncovered intriguing relationships. The study examined the relationship between gene expression in several cell lines and drug sensitivity (IC50 values). The KG1 cell line demonstrated a positive correlation (about 0.382) between gene expression and drug responsiveness. This suggests that as IC50 values increase (indicating reduced drug sensitivity), gene expression patterns in KG1 tend to exhibit higher expression levels. This finding underscores the potential link between drug response and the unique gene expression profiles of the KG1 cell line.

The gene ontology study reveal processes of the 75 common genes in the KG1 cell line. These genes were discovered to be involved in a number of procedures in BP, including lipid oxidation and fatty acid metabolism. Their existence in cellular elements including mitochondria and cytoskeletal structures was discovered through CC analysis. The genes in MF showed oxidoreductase and acyltransferase-related activity. Furthermore, relevant phrases connected to these frequent genes were identified by gene set enrichment analysis (GSEA). Overall, these results reveal the many tasks and responsibilities played by these genes in the KG1 cell line, offering insightful information about its biological operations and molecular processes.

The research findings hold significant promise for advancing cancer medication. By identifying genes consistently expressed in diverse cell lines and unveiling their roles in various biological processes, this study lays the foundation for targeted therapies. The positive correlation between gene expression and drug responsiveness in the KG1 cell line suggests specific genes' influence on treatment sensitivity, potentially enabling the development of personalized cancer therapies. Insights into G protein-coupled receptor signaling pathways offer potential drug development opportunities. Moreover, this research could lead to the discovery of biomarkers for cancer diagnosis and prognosis, and it sheds light on drug resistance mechanisms, providing essential knowledge to create more effective and individually tailored treatments. Ultimately, this work has the potential to significantly improve cancer patient outcomes by advancing precision medicine in oncology.

7 References

1. *Prostate cancer UK* . (2023) Available at: https://prostatecanceruk.org/prostate-information-and-support/risk-and-symptoms/about-prostate-cancer (Accessed: 22 June).

2. *Overlap Analysis.* Available at: https://sparcopen.org/our-work/negotiation-resources/data-analysis/overlap-analysis/ (Accessed: Aug 8, 2023).

3. Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P. and Harris, N.L. (2023) 'The Gene Ontology knowledgebase in 2023', *Genetics,* 224(1), pp. iyad031.

4. Basaldella, M., Furrer, L., Tasso, C. and Rinaldi, F. (2017) 'Entity recognition in the biomedical domain using a hybrid approach', *Journal of biomedical semantics,* 8(1), pp. 1-14.

5. Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P. (2004) 'Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments', *FEBS letters,* 573(1-3), pp. 83-92.

6. Daftary, G.S. and Taylor, H.S. (2006) 'Endocrine regulation of HOX genes', *Endocrine reviews,* 27(4), pp. 331-355.

7. Deguchi, Y. and Kehrl, J.H. (1991) 'Nucleotide sequence of a novel diverged human homeobox gene encodes a DNA binding protein.', *Nucleic acids research,* 19(13), pp. 3742.

8. Drexler, H.G., Quentmeier, H., MacLeod, R., Uphoff, C.C. and Hu, Z. (1995) 'Leukemia cell lines: in vitro models for the study of acute promyelocytic leukemia', *Leukemia research,* 19(10), pp. 681-691.

9. Duan, J., Yu, Y., Li, Y., Jing, L., Yang, M., Wang, J., Li, Y., Zhou, X., Miller, M.R. and Sun, Z. (2017) 'Comprehensive understanding of PM2. 5 on gene and microRNA expression patterns in zebrafish (Danio rerio) model', *Science of the Total Environment,* 586, pp. 666-674.

10. El Zein, N., Badran, B. and Sariban, E. (2008) 'The neuropeptide pituitary adenylate cyclase activating polypeptide modulates Ca2 and pro-inflammatory functions in human monocytes through the G protein-coupled receptors VPAC-1 and formyl peptide receptor-like 1', *Cell calcium,* 43(3), pp. 270-284.

11. Gene Ontology Consortium (2015) 'Gene ontology consortium: going forward', *Nucleic acids research,* 43(D1), pp. D1049-D1056.

12. Gogtay, N.J. and Thatte, U.M. (2017) 'Principles of correlation analysis', *Journal of the Association of Physicians of India,* 65(3), pp. 78-81.

13. Greenhalgh, T. (2019) *How to read a paper: the basics of evidence-based medicine and healthcare.* John Wiley & Sons.

14. Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L. and Chory, J. (2006a) 'RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis', *Bioinformatics,* 22(22), pp. 2825-2827.

15. Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L. and Chory, J. (2006b) 'RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis', *Bioinformatics,* 22(22), pp. 2825-2827.

16. Khatri, P., Sirota, M. and Butte, A.J. (2012) 'Ten years of pathway analysis: current approaches and outstanding challenges', *PLoS computational biology,* 8(2), pp. e1002375.

17. Kim, D., Langmead, B. and Salzberg, S.L. (2015) 'HISAT: a fast spliced aligner with low memory requirements', *Nature methods,* 12(4), pp. 357-360.

18. Kodama, Y., Shumway, M. and Leinonen, R. (2012) 'The Sequence Read Archive: explosive growth of sequencing data', *Nucleic acids research,* 40(D1), pp. D54-D56.

19. Melus, C., Rossin, B., Aure, M.A. and Mahler, M. (2021) 'Biomarker and data science as integral part of precision medicine'*Precision Medicine and Artificial Intelligence* Elsevier, pp. 65-96.

20. Morgan, R., Boxall, A., Harrington, K.J., Simpson, G.R., Michael, A. and Pandha, H.S. (2014) 'Targeting HOX transcription factors in prostate cancer', *BMC urology,* 14, pp. 1-9.

21. Morgan, R., El-Tanani, M., Hunter, K.D., Harrington, K.J. and Pandha, H.S. (2017) 'Targeting HOX/PBX dimers in cancer', *Oncotarget,* 8(19), pp. 32322.

22. Morgan, R., Pirard, P.M., Shears, L., Sohal, J., Pettengell, R. and Pandha, H.S. (2007) 'Antagonism of HOX/PBX dimer formation blocks the in vivo proliferation of melanoma', *Cancer research,* 67(12), pp. 5806-5813.

23. Paço, A., de Bessa Garcia, S.A. and Freitas, R. (2020a) 'Methylation in HOX clusters and its applications in cancer therapy', *Cells,* 9(7), pp. 1613.

24. Paço, A., de Bessa Garcia, S.A. and Freitas, R. (2020b) 'Methylation in HOX clusters and its applications in cancer therapy', *Cells,* 9(7), pp. 1613.

25. Qudsi, J., Tajuddin, M., Hidayat, S. and Yusuf, S.A.A. (2023) *Best wavelet decomposition channel determination for speech processing application using two-way ANOVA*. AIP Publishing, .

26. Quinonez, S.C. and Innis, J.W. (2014) 'Human HOX gene disorders', *Molecular genetics and metabolism,* 111(1), pp. 4-15.

27. Rawla, P. (2019) 'Epidemiology of prostate cancer', *World journal of oncology,* 10(2), pp. 63.

28. ResearchGate *Figure 1: HXR9 mechanisms of action. In the absence of inhibition HOX...* Available at: https://www.researchgate.net/figure/HXR9-mechanisms-of-action-In-the-absence-of-inhibition-HOX-and-PBX-dimerize-enter-the_fig1_314716262 (Accessed: July 13, 2023).

29. Shah, N. and Sukumar, S. (2010) 'The Hox genes and their roles in oncogenesis', *Nature Reviews Cancer,* 10(5), pp. 361-371.

30. Shen, L., Zhou, T., Du, Y., Shi, Q. and Chen, K. (2019) 'Targeting HOX/PBX dimer formation as a potential therapeutic option in esophageal squamous cell carcinoma', *Cancer Science,* 110(5), pp. 1735-1745.

31. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R. and Lander, E.S. (2005) 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences,* 102(43), pp. 15545-15550.

32. Van de Donk, N., Kamphuis, M.M., Van Dijk, M., Borst, H., Bloem, A.C. and Lokhorst, H.M. (2003) 'Chemosensitization of myeloma plasma cells by an antisense-mediated downregulation of Bcl-2 protein', *Leukemia,* 17(1), pp. 211-219.

33. Witte, R.S. and Witte, J.S. (2017) *Statistics*. John Wiley & Sons.

Appendix:

The provided link leads to a directory containing a variety of project-related resources. These resources encompass Python files, R scripts, CSV files for counting, and graphical representations in the form of plotted graphs. This collection of files is integral to the project's documentation and analysis.

Dissertation A3

https://uwloffice365live-my.sharepoint.com/:f:/g/personal/21524331_student_uwl_ac_uk/EmQDmUT563pAkOYbKS4GU-gBXcaYFG0447LCHFa3oVUPXw?e=ejLOmh