**MATERIALS AND METHODS**

**Materials:**

To conduct our analysis we downloaded RNA-seq data acquired from a previous study[] of MYBL2(table1) from NCMI using the SRATools software. This data was acquired from Lung adenocarcinoma cell lines, in four of which MYBL2 was knocked down using specific siRNAs targeting this gene. The other four were treated with scrambled siRNA and were used as a control group. The original analysis was performed with the Illumina Nova Seq 6000 and all datasets contains paired reads. Additionally we got Chip-seq data from the same study, which was also gained from lung adenochroma cell lines using antibodies against MYBL2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tissue type** | **Sample name** | **Assay** | **Data availability** |
| Lung adenocarcinoma | kdMYBL2\_rep1 | siMYBL2 RNA-seq | GSM6057124 |
| Lung adenocarcinoma | kdMYBL2\_rep2 | siMYBL2 RNA-seq | GSM6057125 |
| Lung adenocarcinoma | kdMYBL2\_rep3 | siMYBL2 RNA-seq | GSM6057126 |
| Lung adenocarcinoma | kdMYBL2\_rep4 | siMYBL2 RNA-seq | GSM6057127 |
| Lung adenocarcinoma | WT\_rep1 | siControl RNA-seq | GSM6057132 |
| Lung adenocarcinoma | WT\_rep2 | siControl RNA-seq | GSM6057133 |
| Lung adenocarcinoma | WT\_rep3 | siControl RNA-seq | GSM6057134 |
| Lung adenocarcinoma | WT\_rep4 | siControl RNA-seq | GSM6057135 |
| Lung adenocarcinoma | MYBL\_Chip | MYBL2 ChIP-seq | GSM6057149 |

**RNA-seq mapping and quantification:**

The RNA-seq reads were first mapped to the human reference transcriptome GENCODE v39 using the Kallisto 0.46.0docker image. This algorithm first maps the reads in a short timeframe by performing a pseudoalignemt using de Bruijn Graphs, then quantifying the identified transcripts in reads per million (tpm). The Kallisto program was run for quantification for double stranded reads, with an estimated average fragment length of 250 and standard deviation length of fragments of 50. The analysis produced 3 files for each sample, of which the one called “abundance.tsv”, containing the target transcript id and their abundance, was used for further analysis.

**Differentially expressed gene analysis:**

Further analysis of the data after quantification of the transcripts was done in R 4.2.2. The files from the previous analysis were imported into R using the “txtimport” package. To facilitate easier interpretation the transcript IDs were replaced with gene names following Gencode v.38 gene annotations. PCA analysis and differential gene expression (DEG) analysis was performed using the “PCAtools” and “edgeR” packages. For the DEG analysis, first low expressed genes were filtered out, then the analysis was performed using the “glmQLFIT” function. Only genes with a p-value<0,05 were considered significant.

**Enrichment Analysis**

In this phase of the research project, our primary objective was to investigate the potential impact of the MYBL2 gene knockdown on various cellular pathways and functions. To achieve this, we conducted an enrichment analysis employing two distinct methods: Over-Representation Analysis (GOA) and Gene Set Enrichment Analysis (GSEA).

The Over-Representation Analysis aimed to compare the entire set of normally expressed genes with the list of differentially expressed genes (DGEs) obtained from the experiment. The purpose was to identify significant differences in specific biological processes, as annotated in the Gene Ontology (GO) database.

Conversely, the Gene Set Enrichment Analysis involved the utilization of all genes, which were classified based on their p-values and fold-changes. This analysis aimed to detect disparities in particular pathways, as defined in the KEGG database.

Both enrichment analyzes were conducted using the Webgestalt software, which facilitated the exploration of gene lists and the identification of functional enrichments associated with the knockdown of the MYBL2 gene.

**RESULTS**

**Most enriched biological processes are related to cellular processes**

In order to investigate the statistically significant and enriched biological processes within the list of differentially expressed genes (DGEs), we conducted an Over-Representation Analysis (ORA). This analysis aimed to determine the extent of enrichment for each biological process by calculating an enrichment ratio, which represents the number of over-represented or enriched genes in relation to the total gene count. The results of this analysis were then visualized in Figure X.

Ein Bild, das Text, Reihe, Diagramm, Schrift enthält.

Automatisch generierte BeschreibungThe ORA revealed a set of highly enriched biological processes among the DGEs. These processes included Mitotic Nuclear Division, Nuclear Chromosomal Segregation, Chromosomal Segregation, Cell Division, Nuclear Division, Organelle Fission, Mitotic Cell Cycle Process, Positive Regulation of RNA polymerase II transcription, upregulation of the RNA metabolic process, and upregulation of gene expression.

Figura 1 - Most enriched biological processes by ORA

To assess the statistically significant and enriched pathways across all genes, not limited to the differentially expressed genes (DEGs), we performed a Gene Set Enrichment Analysis (GSEA). This analysis allowed us to determine the direction and magnitude of enrichment for each gene set using the normalized enrichment score (NES). The NES provides insight into the up-regulation (NES > 0) or down-regulation (NES < 0) of a given pathway. The results of this analysis are presented in Figure X+1. (https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm)

Upon examining Figure X+1, we identified the most enriched up-regulated pathways across all genes. These pathways include Autoimmune thyroid disease, Staphylococcus aureus infection, Allograft rejection, Cell adhesion molecules (CAMs), Phosphatidylinositol signaling system, Endocrine and other factor-regulated calcium reabsorption, Graft-versus-host disease, Intestinal immune network for IgA production, Glycerophospholipid metabolism, and Synaptic vesicle cycle.

Ein Bild, das Text, Screenshot, parallel, Schrift enthält.

Automatisch generierte BeschreibungConversely, the most enriched down-regulated pathways across all genes were found to be Homologous recombination, Ribosome biogenesis in eukaryotes, Terpenoid backbone biosynthesis, Spliceosome, Mismatch repair, Fanconi anemia pathway, RNA transport, Ribosome, DNA replication, and Cell cycle.

Figura 2 - Most enriched up-regulated and down-regulated pathways by GSEA