Detecting FOXM1 gene targets in K562 cells through the integration of ATAC-seq, ChIP-seq, and RNA-seq data

Neelang Parghi

May 11, 2021

Abstract

Using different types of publicly available biological sequencing data, this experiment identifies which genes are likely targets of the FOXM1 transcription factor. Differentially expressed genes were found using CRISPR interference RNA-seq data and untreated control data, along with control ATAC-seq and ChIP-seq data. These different data were integrated to produce lists of genes that are potential targets of FOXM1. The resulting gene lists were further analyzed to look at which gene ontology terms and transcription factors were significantly represented.

Introduction

Transcription factors (TFs) are proteins that can regulate gene expression by recognizing and binding to specific DNA sequences. They are either *activators* to activate transcription, or *repressors* to repress transcription.

Proteins in the forkhead box (Fox) family can act as both activators and repressors. The FOXM1 transcription factor, one of the forkhead box protein TFs, is a key regulator in multiple cell processes and is implicated in multiple cancers due to its importance in cell proliferation and other different cell processes. Overexpression of FOXM1 promotes cell cycle progression[9] and there has been accumulating evidence that cancer patients with overexpression of this gene have a poor prognosis during their treatment. How does FOXM1 control different genes in different cellular contexts?

The K562 cell line is used because it's an immortalized cell line available in the ENCODE database, and is also a tier 1 cell line that grows well and is transfectable. The cells were taken from a female chronic myelogenous leukemia patient. This cell line was chosen because it's a leukemia cell line and FOXM1 is implicated in leukemia [10, 12].

ChIP-seq is a method to analyze protein interactions with DNA by combining chromatin immunoprecipitation and DNA sequencing with the goal of identifying binding sites for proteins, such as the FOXM1 transcription factor [3]. ATAC-seq allows the measure of genome regions that are accessible to different proteins binding. These regions can potentially be considered regulatory elements that drive gene expression. Additionally, RNA-seq data is used to determine genes that are differentially expressed between different conditions.

Materials and Methods

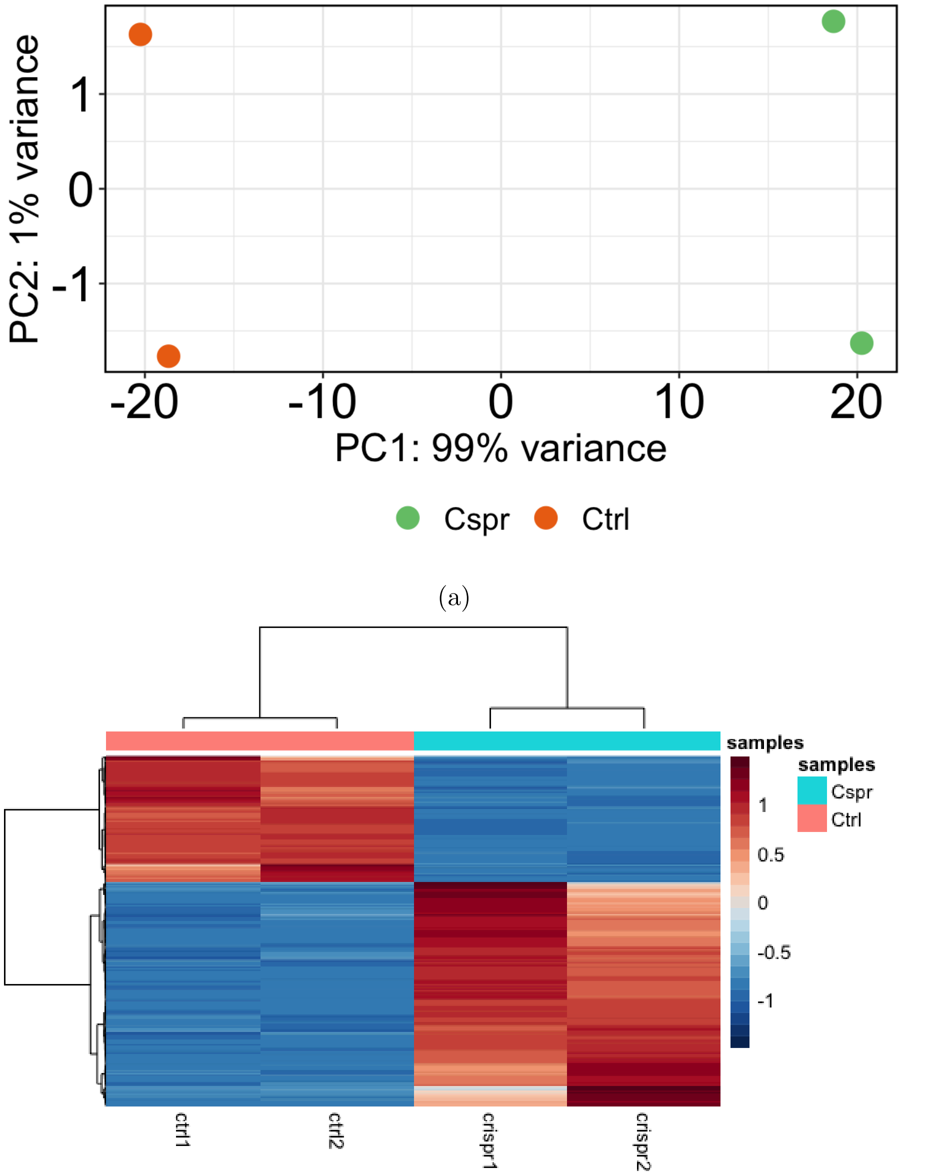
Data preparation

RNA-seq data came from K562 cells treated with CRISPR interference targeting FOXM1 (ENCODE ENCSR701TVL), repressing the gene's expression. Two isogenic replicates were provided. Control RNA-seq data without any type of treatment was also provided (ENCODE ENCSR095PIC). Three replicates of ATAC-seq data for just K562 control cells (ENCODE ENCSR868FGK) and two replicates of FOXM1 ChIP-seq data from K562 cells were also provided (ENCODE ENCSR429QPP). No data from ATAC-seq and ChIP-seq in the target CRISPR interference condition were provided.

FASTQ files for each dataset were trimmed for adapter content and base quality using `Trim Galore`. The trimmed RNA-seq FASTQ files were aligned to the HG38 human reference genome using `HISAT2` once the corresponding index was created. `HISAT2` was used because it is a splice-aware aligner specially suited for RNA-seq. ATAC-seq and ChIP-seq data were aligned to HG38 using BOWTIE2. Each aligned dataset was converted from SAM format to sorted BAM format using `samtools`.

RNA-seq

A count matrix was generated using the `featureCounts` function in the `Rsubread` package in `R`. Differentially expressed genes between the two RNA-seq conditions were found using the `deSeq2` package in `Rstudio`. PCA plotting and heatmap visualization showed there was no batch effect taking place (Figures 1a and 1b). A p-value threshold of 0.05 and $\log_2$ fold change threshold of 1 was used to determine which genes were significantly differentially expressed between the control and CRISPR datasets.



(a)

(b)

Figure 1: (a) PCA plot of control and CRISPR datasets. (b) Heatmap of control and CRISPR datasets.

ATAC-seq and ChIP-seq

The pipelines for the ATAC-seq and ChIP-seq datasets were similar to each other, except for small differences. Once the BAM files were produced for each replicate, the `multiBamSummary` function from `deeptools` was used to measure inter-replicate correlation to ensure that the replicates did not diverge significantly.

Peak calling was performed on both datasets using `MACS2` to detect narrow peaks. For ATAC-seq, broad peaks were also found. This was done because ATAC-seq peaks can have variable sizes [14]. Results for both ATAC-seq approaches are reported in the Results section. The peak calling steps were performed for each replicate BAM file, sorted, then merged into a single BED file using `bedtools`. The tracks for each displayed in the Interactive Genomics Browser are shown in Figure 2.

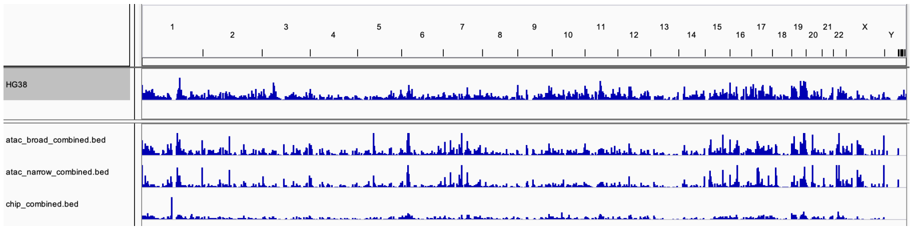


Figure 2: Display of BED files for ATAC-seq (broad and narrow) and ChIP-seq in Interactive Genomics Browser with HG38 human reference genome.

The `annotatePeaks` function in `HOMER` was used to annotate the peaks for both datasets. Once this step was completed, a list of genes associated with ATAC-seq and ChIP-seq peaks was available, as well as a list of differentially expressed genes from RNA-seq. To associate the ATAC-seq/ChIP-seq peaks to genes they may be regulating, the intersections of these genes lists was produced. The `intersect()` function in `Rstudio` was used for this. To predict genes that are likely to be FOXM1 targets, the intersections of the peaks in the MACS2 outputs for ATAC-seq and ChIP-seq were found, then this peak list was annotated and integrated as before with the RNA-seq gene list. The bedtools intersect function was used for this step.

Each list of intersecting genes was input into the gProfiler2 package in Rstudio and the corresponding website [11] to detect and visualize the gene sets and pathways that were significantly represented in each list.

Results

Differentially expressed genes

The RNA-seq analysis revealed 6177 genes that are differentially expressed between the control data and FOXM1 CRISPR interference data. A functional enrichment plot of these genes is seen in Figure 3.

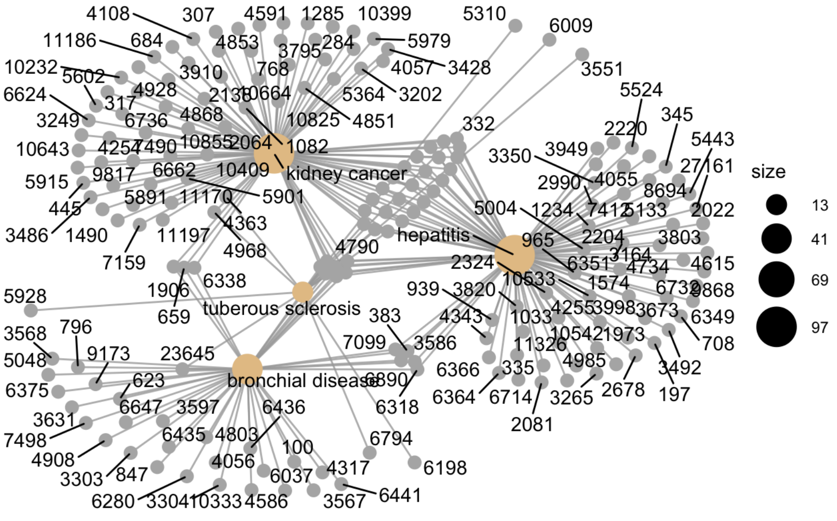


Figure 3: Network plot of differentially expressed genes.

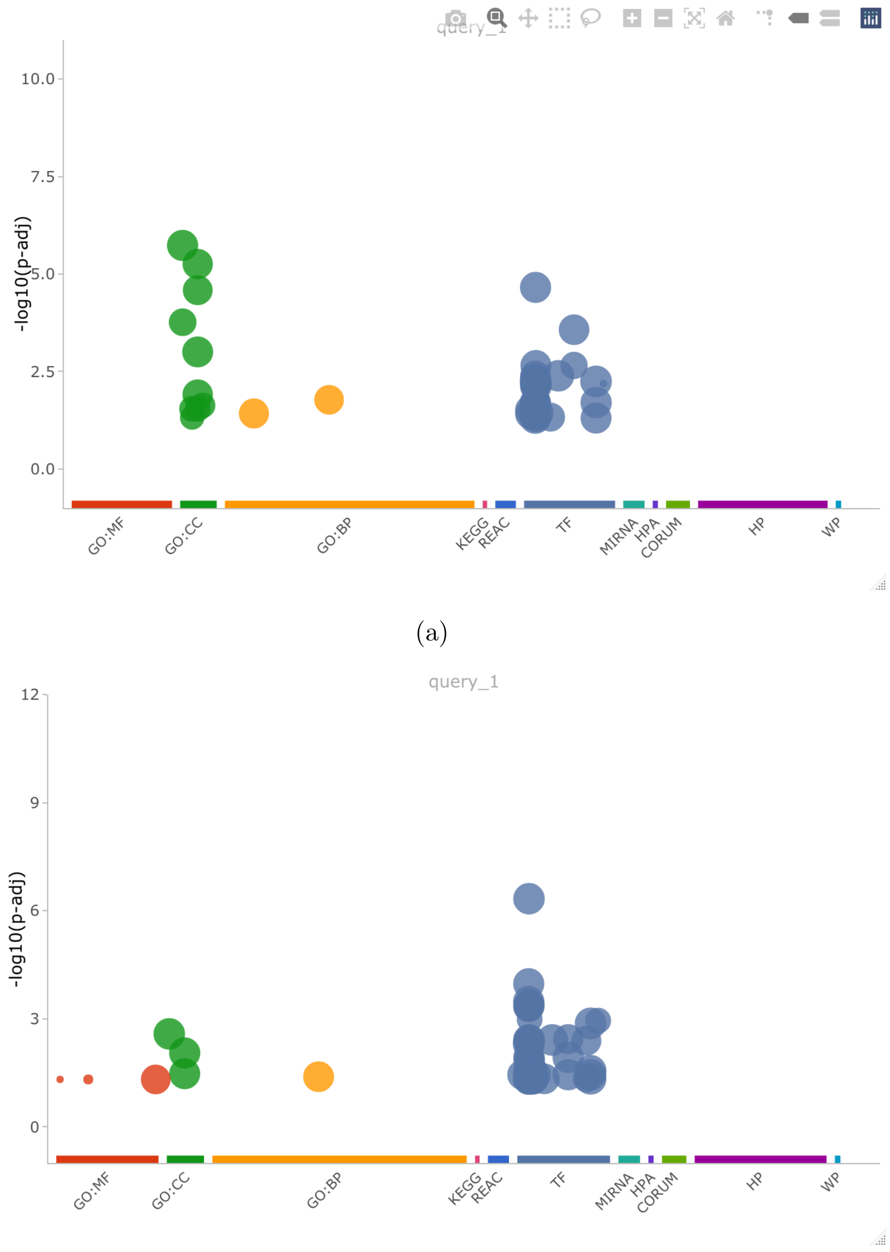
We see that kidney cancer and hepatitis are the two most represented disease categories, with bronchial disease as the third most significant.

Associating ATAC-seq/ChIP-seq with RNA-seq

Peaks from the ATAC-seq and ChIP-seq data were annotated separately for each dataset, integrated with each other, then this list was integrated with the differentially expressed genes from RNA-seq. This step was carried out twice: once for the narrow ATAC-seq peaks and once with the broad peaks. A statistical enrichment analysis of the gene lists was performed. This looks at gene ontology (GO) terms, biological pathways, transcription factors, and other annotation databases. Significance was measured using a false discovery rate (FDR) threshold of 0.05. The intersection of all three datasets using the broad peak ATAC-seq genes yielded a list of 567 genes. Using the narrow peak ATAC-seq genes produced a list of 412 genes.

Looking at the intersection of genes including the annotated ATAC-seq broad peaks, we see that the most significant gene sets are in the GO cellular component (GO:CC) term. The three terms below the FDR threshold were GO:0005622, *intracellular anatomical structure*; GO:0043229, *intracellular organelle*; and GO:0043231, *intracellular membrane-bounded organelle* (Figure 4a). The E2F-2 transcription factor was also significantly represented among the gene list (Figure 4b).

When looking at the intersecting genes including annotated ATAC-seq narrow peaks, the same gene sets and transcription factor as the broad peak results were below the 0.05 FDR threshold.



(a)

(b)

Figure 4: (a) g:Profiler plot for integrated RNA-seq, ChIP-seq, and ATAC-seq (broad peak annotated) genes. (b) Same as (a), but using narrow peak annotated ATAC-seq genes.

Likely FOXM1 targets from integrating ATAC-seq and ChIP-seq peaks with differentially expressed genes

After integrating the peaks for ATAC-seq and ChIP-seq then annotating this combined peak list, the resulting gene list was integrating the the list of differentially expressed genes from RNA-seq. This produces a list of possible FOXM1 target genes based on the ATAC-seq and ChIP-seq peaks. Using broad peaks from ATAC-seq, there were 512 intersecting genes from all three datasets. When using narrow peak ATAC-seq data, there were 381 intersecting genes.

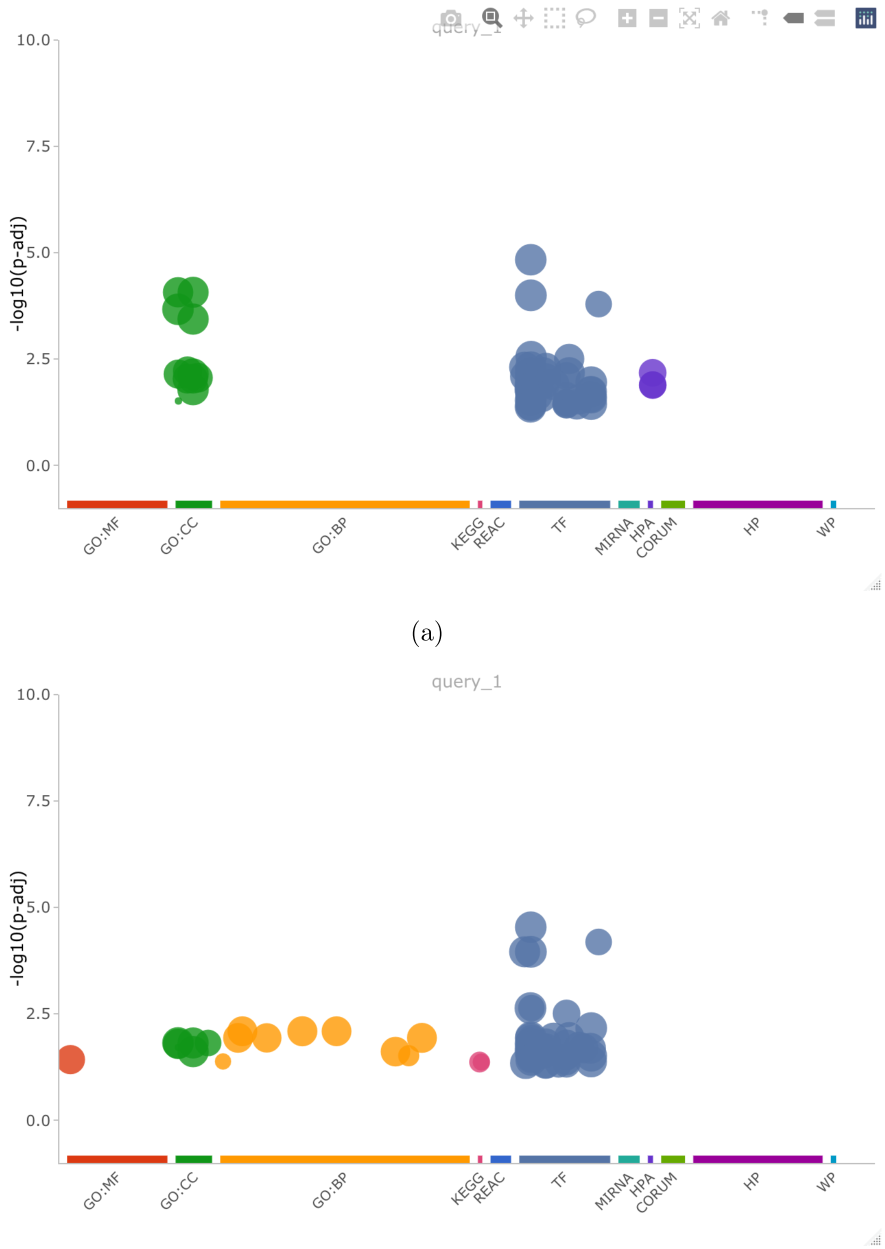
When using the broad ATAC-seq peaks (Figure 5a), two GO gene sets under the FDR threshold were found in the GO:CC namespace. These were GO:0005634, *nucleus*; and GO:0043231, *intracellular membrane-bounded organelle*. The E2F-2 transcription factor was also significantly represented, as was the E2F-3 transcription factor.

In looking at the intersection using narrow ATAC-seq peaks (Figure 5b), no GO terms were within the FDR threshold but three transcription factors were. These were E2F-2, CDX2, and ZNF418.

Discussion

The three most significantly represented disease categories from the differentially expressed gene results (kidney cancer, hepatitis, and bronchial disease) make biological sense. The FOXM1 protein is directly implicated in kidney cancer [13, 2]. Furthermore, hepatitis is a strong risk factor for kidney cancer via “genetic recombination and the activation of oncogenes such as ‘forkhead box protein M1’ (FOXM1)” [4].

Previous published work has also shown that FOXM1 is implicated in



(a)

(b)

Figure 5: (a) g:Profiler plot for integrated RNA-seq, ChIP-seq, and ATAC-seq (broad peak annotated) genes. (b) Same as (a), but using narrow peak annotated ATAC-seq genes.

the proliferation of pulmonary cell types seen in various types of bronchial disease [15, 7], which is the third most represented category.

Results from both conditions (broad and narrow ATAC-seq peaks) of both integration experiments showing that GO terms for various cell components are significant are also not surprising given FOXM1's role in cell cycle progression and as a proto-oncogene.

Among the different experiments, the E2F-2 transcription factor was significantly represented in each condition. The E2F family of TFs "plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses" [5]. This TF warrants further analysis since it was significantly represented across all conditions. The E2F family of transcription factors has been shown to upregulate FOXM1 expression [8].

Another significant transcription factor was E2F-3, another member of the E2F family which regulates the expression of cell cycle genes via the retinoblastoma protein and is implicated in "a number of human cancers" [6]. The CDX2 TF was also represented, which is a regulator of intestine-specific genes involved in cell growth and differentiation and is associated with intestinal inflammation and tumorigenesis when expressed aberrantly [1]. The third was ZNF418, a zinc finger protein coding gene associated with ovarian squamous cell carcinoma and related to pathways for gene expression and herpes simplex virus 1 infection.

All of the transcription factors described above are either directly related to different types of cancer and/or cell cycle regulation. This result also makes biological sense and warrants further research into how FOXM1 affects these diseases and processes.

Further work may involve duplicating the steps above on cell lines other than K562 to ensure that the results are replicated. Such experiments could begin with other abundantly available leukemia cell lines and then expand to other cancer types. A closer look at how FOXM1 overexpression modifies the cell cycle to create such poor prognoses is also warranted. In particular, understanding the connection with the significantly represented transcription factors could play a major role in understanding FOXM1 and how it is a component of many cancer types.

References

[1] CDX2. *GeneCards: the Human Gene Database.* https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDX2

[2] Chen, T.; Xiong, J.; Yang, C. et al. (2014) Silencing of FOXM1 transcription factor expression by adenovirus-mediated RNA interference inhibits human hepatocellular carcinoma growth. *Cancer Gene Therapy* Volume 21, p. 133–138. https://doi.org/10.1038/cgt.2014.8

[3] “ChIP sequencing.” Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 7 May. 2021. Web. 10 May. 2021. https://en.wikipedia.org/wiki/ChIP_sequencing

[4] Chou, L.-F.; Chen, C.-Y.; Yang, W.-H. et al. (2020) Suppression of hepatocellular carcinoma progression through FOXM1 and EMT inhibition via hydroxygenkwanin-induced miR-320a expression. *Biomolecules* volume 10, issue 1, p. 1–15. https://doi.org/10.3390/biom10010020

[5] E2F2. *GeneCards: the Human Gene Database.* https://www.genecards.org/cgi-bin/carddisp.pl?gene=E2F2

[6] E2F3. *GeneCards: the Human Gene Database.* https://www.genecards.org/cgi-bin/carddisp.pl?gene=E2F3

[7] Kalinichenko, VV; Gusrova, GA; Tan, Y. et al. (2003) Ubiquitous expression of the forkhead box M1B transgene accelerates proliferation of distinct pulmonary cell types following lung injury. *The Journal of Biological Chemistry* volume 278, issue 3, p. 37888–37894. https://doi.org/10.1074/jbc.m305555200

[8] Liao, G.-B.; Li, X.-Z.; Zeng, S. et al. (2018) Regulation of the master regulator FOXM1 in cancer. *Cell Communication and Signaling* volume 16, issue 57, p. 1–16. https://doi.org/10.1186/s12964-018-0266-6

[9] Myatt, SS; Lam, E W-F (2007). The emerging roles of forkhead box (Fox) proteins in cancer. *Nature Reviews Cancer* volume 7, p. 847–859. https://doi.org/10.1038/nrc2223

[10] Nakamura, S; Hirano, I; Okinaka, O et al.x (2010) The FOXM1 transcriptional factor promotes the proliferation of leukemia cells through modulation of cell cycle progression in acute myeloid leukemia. *Carcinogenesis* volume 31, issue 11, 2012–2021. https://doi.org/10.1093/carcin/bgq185

[11] Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, Hedi.; Vilo, J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) *Nucleic Acids Research* Volume 47, issue W1, p. W191—W198. https://doi.org/10.1093/nar/gkz369. Available at: http://biit.cs.ut.ee/gprofiler/gost

[12] Sheng, Y; Yu, C; Liu, Y et al. (2020) FOXM1 regulates leukemia stem cell quiescence and survival in MLL-rearranged AML. *Nature Communications* volume 11, 928. `https://doi.org/10.1038/s41467-020-14590-9`

[13] Xia, L.; Huang, W.; Tian, D. et al. (2012) Upregulated FoxM1 expression induced by hepatitis B virus X protein promotes tumor metastasis and indicates poor prognosis in hepatitis B virus-related hepatocellular carcinoma. *Journal of hepatology* Volume 57, issue 3, p. 600-612. `https://doi.org/10.1016/j.jhep.2012.04.020`

[14] Yan, F.; Powell, D.R.; Curtis, D.J. et al. (2020) From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* Volume 21, issue 22, p. 1–16. `https://doi.org/10.1186/s13059-020-1929-3`

[15] Zhang, J.; Zhang, J.; Cui, X. et al. (2015) FoxM1: a novel tumor biomarker of lung cancer. *International Journal of Clinical and Experimental Medicine* Volume 8, issue 4, p. 3136–3140.