DRUG REPURPOSING

Liver Cancer

Drug repurposing increases in popularity due to the benefits it brings with respect to high costs and ressources consumption in the development of new drugs. Our group focuses on liver cancer as a domain of possibilities for drug repurposing

Artan Kryeziu, Eran Levi, Schwarz Robert

Contents

Introduction:

The development of new drugs is linked with massive costs and use of resources. This is why the development of drugs for new diseases is not financially attractive. There are about 7000 rare diseases that still have no treatment. (g[ithubrepo) However, drugs that are already approved or in late stages of development, can possibly target these rare diseases. As the early stages of drug development can be skipped, the repurposing of drugs can be financially attractive and also saves time. More importantly, rare diseases can finally be treated.

The repurposing of drugs is the use of a drug for the treatment of another disease than it is used for in the first place. There are various approaches to how drug repurposing can progress. (nature reviews). Another big advantage of drug repurposing is the lower risk of wasting resources. Many drugs show side effects only in late stages of drug development after a lot of resources have been already invested. Already approved drugs are known for being tolerated by the organism that makes use of it. CIT

Our group's goal is to analyze the success rates of treating cancer with non-oncological drugs. To this end, we have received information of an extensive screening on different cancer cell lines' sensitivity to multiple non-oncological drugs. This drug screening was performed by utilizing the "profiling relative inhibition simultaneously in mixtures"-method (PRISM). This method relies on the relative cell number of a cell line after treatment to measure a specific drug's efficiency. Before treatment with a certain drug, the cell lines are transfected with a unique nucleotide-barcode. After the treatment, the frequency of the occurrence of each unique barcode is quantified, providing information about the relative cell number after subjection to drug treatment with respect to a control pool. A negative value has thus to be associated with effective inhibition of the cancer cell lines' growth and proliferation. Each drug was employed in eight different concentrations. The results of this procedure are stored in the data frame "prism", which we used in our analysis (Citation, nature review etc). In addition to the central "prism" data frame, our group used the data frames "prism.treat", "prism.cl", "prism.exp", "prism.cnv", "prism.snv" and "prism.achilles". These data frames respectively provide information about the utilized drugs, cancer cell lines, gene expression rates, gene copy number variations, gene mutations and proliferation scores of the cell lines after knock-out (by the CRISPR-Cas9 system) of specific genes involved in the screening. Specifically, our group has focused on the analysis of cancer cell lines originating from liver cancer.

By combining the information acquired from the data sets at our disposal, we hope to achieve models for a more sensible approach in targeting liver cancer with non-oncological drugs.

Methods:

Data structuring and cleanup

For the described data frames, structuring and cleaning methods have been utilized. The data has been ordered so that it matches an alphabetical order in gene columns, and certain cell identificators in cell line rows. Furthermore, missing values have been entirely deleted (either individually or by row) or completed, using mean imputation. In effect, five new data frames were created: “new_prism.treat”, “new_prism.exp”, “new_prism.cl”, “new_prism.cnv” and “new_prism.achilles”. Using these new data frames, several options were facilitated, such as conserving the indices of desired elements across data frames, analyzing differences between the data frames locally and testing whether operations on the data frames bring the desired result by providing the same index input.

Graphical representations

The analysis of the data has been visualized using methods that describe their relationship. Therefore, according to the specific goals of the results, a variety of display possibilities like heatmaps, barplots, QQ-plots, histograms or scatter plots have been used. A selection of these representations is also presented in this report.

Performing statistical tests

Statistical tests were employed in order to check the distribution type of the data or to investigate if the differences between to data distributions are indeed statistically true. T-Tests (parametrical tests for testing the true difference in means of two vectors) were conducted in the course of the analysis, but the resulted data was not reliable. Thus, Shapiro-Wilk tests for normality and non-parametrical Wilcoxon Rank Sum tests made up the basis of this method.

Developing a Linear Regression Model

Linear regression models were used to extrapolate discovered relationships between variables to predict unknown values or interactions. R provides a built in function “lm” for the construction of linear models.

Reducing the dimensionality of the data

K-means clustering has been applied in order to create subgroups of genes from the data sets based on similarities and differences of their variance according to different criteria. The optimal number of clusters was determined using the Elbow-method and Silhouette-plots.

Results:

Description of the data set

The prism table provides the sensitivity values of the cell lines to different drugs. Table A.1 in the Appendix shows a selection of active molecules and mechanisms of action of the analyzed drugs. From all the 481 cell lines present in the prism data frame, the 17 cell lines involved in liver cancer are identified in Table A.2 in the Appendix. Mutagenesis is a driving factor of cancer development and progression. Table A.3 in the Appendix shows the numbers of mutations identified in each of the liver cancer cell lines. Interestingly, the distribution of the types of mutations in liver cancer is very similar to the distribution of the mutation types across all cancer cell lines (Figure 1)

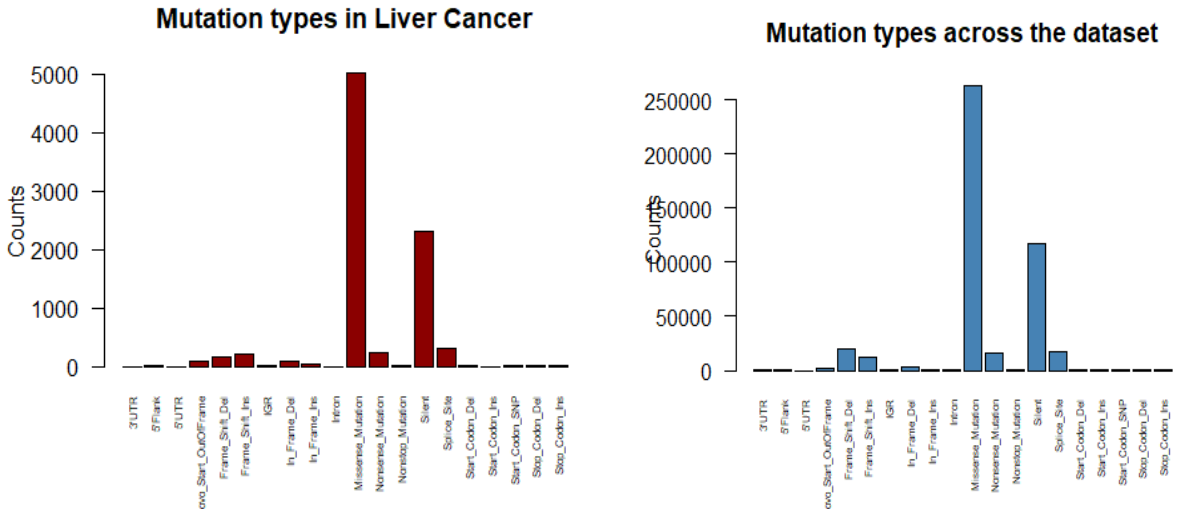


Figure 1: Comparison of the distribution of the mutations between liver cancer and all cancer cell lines

Furthermore, the percentage of deleterious mutations in liver cancer (12.19 %) is also very similar to the overall percentage of deleterious mutations(15.16 %). Also, the genes which are most frequently mutated in liver cancer overlap in significant amounts with those that are generally mutated in cancer.

I General Results

1. Correlation between prism.achilles and prism.exp

The prism.achilles data frame shows the general proliferation rates of the cell lines after specific genes have been knocked out. The data frame prism.exp shows the general expression rates of different genes across the cell lines.

The correlation of these two data frames has been calculated with the goal of finding a connection between a gene's essentiality for a cancer cell line and its overall expression. Figure 5 shows a visualization

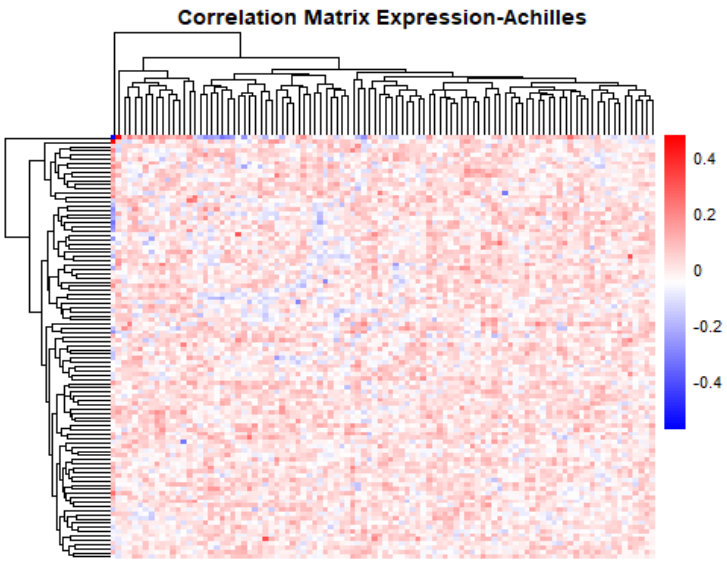


Figure 5: Correlation matrix between the genes of prism.exp and prism.ach. Both the rows and the columns show the common genes of the two data frames. The heatmap shows a random sample of 100 genes from this correlation matrix for the sake of visualization. However, the correlation matrix has been computed for all common genes the two data frames provide information about and comprises 17767 rows and 17767 columns.

We were interested in the extreme high and extreme low correlation values, as they are the ones that provide insights about the analysed genes and their general role in cancer. Therefore, the correlation values above 90% of the maximal correlation value and below 90% of the minimal correlation value were extracted, showing us the genes present in relevant correlation and anti-correlation pairs. First, the general pairs of genes that have a strong correlation or anti-correlation were filtered out. Figure A.6 and Figure A.7 in the Appendix show the relevant gene pairs with high correlation and anti-correlation respectively. The visualization of this data is also shown in the Appendix.

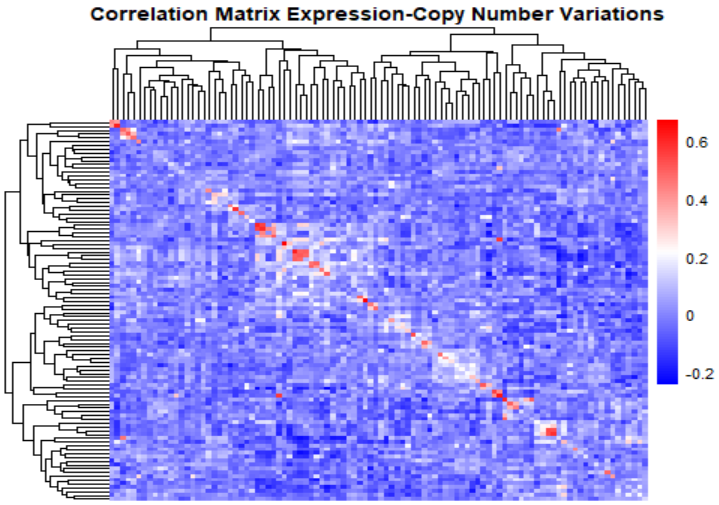
Subsequently, we inspected the high correlation or anti-correlation values of the expression values of the genes with their own Achilles scores. HNF1B and SOX10 were the genes that had an anti-correlation value between their expression values and their Achilles scores which was lower than 90 % of the minimal anti-correlation set in the dataset. The threshold was gradually lowered until we obtained the following genes of interest at a threshold of 75% of the minimal anti-correlation value. TableX shows the genes that showed relevant anticorrelation values between their expression rates and scores in prism.achilles. The visualization of this table can be found in the appendix

```
## [1] "HNF1B" "IRF4"  "PAX8"  "SOX10" "TP63"
```

There is no relevant correlation value lower than the set threshold of 75%.

2. Correlation EXP and CNV

TableX, sample of 100 random genes. Table was computed for all 18559 x 18559. In appendix linearity of selected genes.



Relation between drug sensitivity and gene fitness

We want to see, if there is a correaltion of the drug sensitivity and the gene knockout sensitivity for the gene that the drug targets. For this purpose, we use the prism screen and the CRISPR/Cas9 loss-of-function screen. First, we have to assign a gene knockout to every drug. After the gene knockout scores and the prism scores are isolated, we have two vectors, that contain the information on how sensitive a gene knockout or a treatment is for all cell lines. The vectors, however still do contain missing values. We decided to do the data cleanup so late, because removing the cell lines with at least one missing value would have removed the values of the whole dataset. After the data cleanup, the correlations can be measured with the method "spearman".

Result: Correlations between drug sensitivity and gene knockout sensitivity of drug's target

The correlations between drug sensitivity and gene knockout sensitivity of the drug´s target gene did not show to be high (Fig B1).

The highest negative correlation is -0.17. The highest positive correlation is 0.16. The correlations are normally distributed as shown in Fig. x. To conclude, the given data can not support the hypothesis that gene knockout sensitivity correlates with drug sensitivity.

Finding clusters of genes that have similar effects on cells

We want to investigate if there are genes that have similar effects on the cell survival inorder to find new drug targets. For this reason, for every gene, we calculate the mean expression level, the average copy number variation and the mean gene knockout sensitivity over all cell lines. Then, the data is scaled. After that, we perform a clusering by

kmeans with the preinstalled function "kmeans". The optimum of numbers of clusters was selected with the elbow method. The elbow plot was computed with the function "fviz_nb_clust". In order to make sure, the clusters are clearly defined, we analyzed the clusters with a silhouette plot, performed by the function "fviz_silhouette".

Result: Dimensionality reduction by kmeans clustering

We have multiple approaches, on how we can reduce the dimensionality of data with kmeans clustering: Kmeans clustering of the drug screening, the transcriptome and of copy number variations of cell lines did not lead to clearly defined clusters.

After that, the genes are tried to be clustered by their mean values in expression, copy number variation and their essentiality for cell survival. The genes can be clearly assigned to 4 clusters with an average silhouette width of 0.56. The clusters all have significant differences in their expression, copy number variation and their esentiality for cell survival, as wilcoxon rank sum tests show.

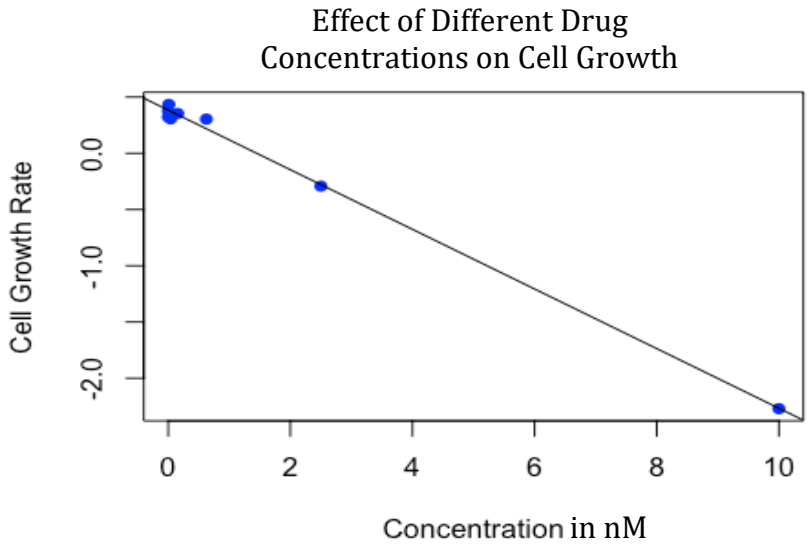
Genes of cluster 1 and 3 seem to be more essential to cell survival than the other 2 clusters. (Fig B2)

II Specific Results for liver cancer

Developing a Linear Regression Model for the prediction of drug concentrations

We utilized linear regression models to predict drug concentrations and we determined an optimal threshold for the R-squared value for each cell line, which served as a criterion for assessing the reliability of linear regression analysis. We set this optimal threshold value as 99% of the maximum R-squared value obtained from the linear regression analysis. For each distinct liver cancer cell line, we filtered the drugs that met the reliability criterion, resulting in a subset of drugs with reliable regression models.

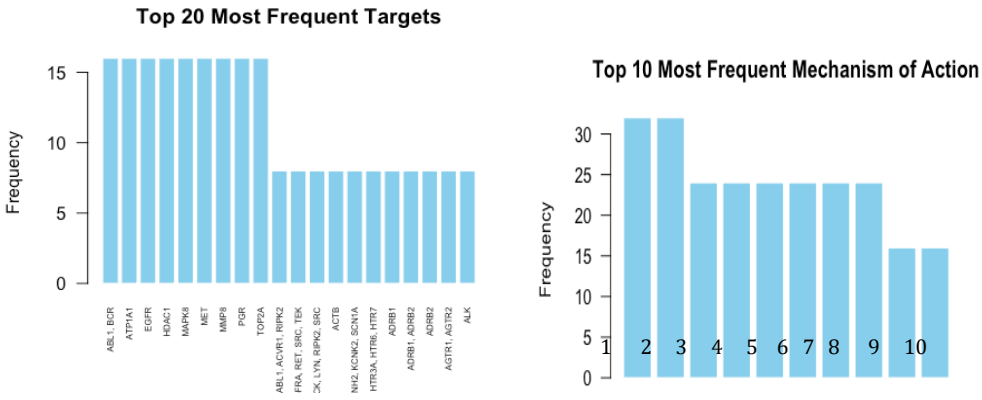
We then examined the frequency of occurrence above the threshold for these reliable drugs across the different liver cancer cell lines. By identifying the most frequently occurring drug, we obtained the linear regression results with the highest consistency and reliability. In our case, for the liver cancer cell lines, our analyses resulted that the drug “BRD-K70358946-001-15-7” occurs 7 times with high reliability (ACH-000393, ACH-000471, ACH-000476, ACH-000480, ACH-000537, ACH-000625, ACH-001318) in 16 cancer cell lines which represents approximately 44% of all liver cancer cell lines. The cell line ACH-000393 was randomly selected as representative. FigX shows the visualization of the linear regression model analysis of the drug “BRD-K70358946-001-15-7” and the cell line ACH-000393. An example of a detailed calculation for the selection of the depicted pair, as well as the function call of the linear model can be found in the appendix (FigX)



FigX: Linear model for the the drug "BRD-K70358946-001-15-7" and the cell line ACH-000393. The slope is described by the following linear equation: y = -0,2653x + 0,384. The cell line ACH-000393 was chosen as a representative, as all 7 cell lines in which the drug "BRD-K70358946-001-15-7" occurs with high R-squared reliability behave similarly.

Determining Therapeutic Strategies:

By analyzing the information in the "prism.treat" data frame, we can identify the targets of each drug and determine which targets are more frequently observed. FigX and FigX show selections of the most frequently targeted structures and the most frequently occurring drug mechanisms of action and in our liver cancer cell lines.



Index References:

1 = adrenergic receptor antagonist

2 = ALK tyrosine kinase receptor inhibitor

3 = HDAC inhibitor

4 = ATPase inhibitor

5 = cyclooxygenase inhibitor

6 = glycogen synthase kinase inhibitor

7 = MEK inhibitor

8 = NFkB pathway inhibitor

9 = opioid receptor antagonist

10 = phosphodiesterase inhibitor

FigX: The top 20 most frequently targeted structures in the 17 liver cancer cell lines.

FigX: The top 20 most frequently targeted structures in the 17 liver cancer cell lines.

2. Differentially expressed genes in liver cancer

In order to construct the base of a specialized treatment for liver cancer, the determinatory characteristics of these liver cancer cell lines have to be determined. Therefore, the differentially expressed genes in liver cancer (the genes which are much stronger or

weaker expressed in liver cancer cell lines than in the other cancer cell lines) have been identified, a random selection of which is shown in Table A.5 in the Appendix.

First, the data distribution of the gene expression vectors has been checked for normality using multiple Shapiro-Wilk tests. Subsequently, the Wilcoxon Rank Sum Test with a significance threshold of 5% was employed in order to determine strong differences in expression, as the data was not consistently normally distributed. The obtained p-values have been corrected using the Benjamini-Hochberg method, thus successfully lowering the false discovery rate.

Finding biomarkers for drug sensitivity

We want to find drugs, that possibly can possibly inhibit liver cancer cell growth. In order to do so, we first extract the genes that are higher or lower expressed in the given liver cancer cell lines than in the remaining cell lines, termed as liver cancer genes. Next, we calculate the correlations between the drug sensitivity and the gene expression of the drug´s target, termed as drug-target gene pair. The correlations can be seen in Fig X. Only 70 drug-target gene pairs have a correlation that is higher than 0.1, which is a low correlation.

Next, there is analyzed, if any of the target genes of the drug-target gene pairs is a liver cancer gene. 3 genes were found to meet the criteria: HRH2, SYK and VDR. HRH2 possibly has a role in cell growth. SYK plays a role in vascular maturation and VDR is a vitamin D receptor. (string database)

These genes can now be assigned to the clusters. HRH2 is assigned to cluster 1 and SYK and VDR are assigned to cluster 2. As genes in cluster 1 seem to be most important to cell survival, further investigations on drugs that target HRH2 can be interesting. However, HRH2 expression still has a very low correlation of 0.14 with the drug sensitivity of BRD-A67516570-001-02-8::0.00976562::HTS002 and the prism scores on liver cancer cell lines of this drug are also above 0 on average, as Fig 3B shows.

Discussion:

1. Description of datasets

TTN is very highly expressed both in liver cancer and generally in cancer cell lines so it appears to have a significant role in the transformation of these cells. This is not surprising, as TTN is often mutated in liver cancer (CIT)

2. Correlation prism.exp and prism.cnv

   The values of the “prism.achilles” data frame describe whether the proliferation rate of a specific cell line has increased or decreased after a specific gene of interest has been knocked out. Thus, the values are inversely proportional to the gene's relevance for the cell’s prosperity. If a gene bears a crucial difference to the cell line, a KO of the said gene will result in lower proliferation rates.

i. In the correlation pairs from TableX, the expression of one gene is strongly correlated with the relevance of the other gene for the cell.

This can hint at *opposite* roles of the two genes and their respective products in cell functionality. A comparison with information from verified studies offers insights into this hypothesis. *FERMT1* proves a negative regulation on cell proliferation (https://amigo.geneontology.org/amigo/term/GO:2000647), while *FERMT2* impacts the proliferation of stem cells positively (https://amigo.geneontology.org/amigo/term/GO:1902462). *TP53* and *EDA2R* were also expected to have opposite functions. Unexpectedly, their function is similar. Normally, this would contradict our hypothesis. However, these two genes directly influence each other, *TP53* activating the transcription on *EDA2R*. This is consistent with the strong correlation of the data for these two genes (Brosh *et al*., 2010). These results could be the fundament of a prediction model for gene pair behavior for yet unknown genes. *RPP25* for example negatively regulates cell proliferation. The function of *RPP25L* is yet unclear CIT, however the interaction with *RPP25* has been proven CIT. Judging after the aforementioned results, we expect *RPP25L* to have a similar function to *RPP25*. This model could be applied up to a certain extent to other unstudied gene pairs with high correlation as well, like *CDS1* and *CDS2*.

ii. In the anti-correlation pairs from TableX, the expression of one gene is strongly anti-correlated with the relevance of the other gene for the cell.

This can hint at *similar* roles of the two genes and their respective products in cell functionality. A comparison with information from verified studies offers insights into this hypothesis. *SOX10* promotes cell proliferation CIT, while *CAPN3* CIT, *CPN1* CIT and *DCT* CIT also have a positive effect on cell proliferation. However, *CAPN3* has also been found to inhibit cell proliferation CIT, which shows the limitations of our hypothesis. Still, the model could be used up to a certain extent to predict the function of an unknown gene, if its high anti-correlating partner has been described.

iii. Anti-correlation of the expression values of relevant genes with their proliferation rates upon their own knock-out in TableX.

The described anti-correlation could be interpreted in two ways. Either a gene's relevance is beneficial for cancer cell lines, which would result in lower proliferation scores upon KO and high general expression rates, or a relevant gene's activity is inhibitory for cancer cell lines, which would manifest in higher proliferation rates upon KO and low general expression rates. Using the values of the "prism.exp" and "prism.achilles", we made a set of predictions and tested them with verified studies. We predicted *HNF1B* and *TP63* to negatively impact cancer cell prosperity and *IRF4*, *PAX8* and *SOX10* to positively impact cancer cells. Our presumption was partially confirmed in the cases of *SOX10* -possible oncogene CIT, *IRF4* – both oncogene and tumor suppressor gene CIT, PAX*8* – oncogene CIT and *TP63* – tumor suppressor gene CIT. However, we were wrong about *HNF1B*, which is a tumor suppressor gene CIT. These findings indicate the potential of the established model, but also its limitations. Identifying important genes for general cancer cell prosperity could represent the basis of the development of new treatment targets against cancer.

EXP CNV

TableX shows clear positive correlation between a gene's expression rate and their copy number variation. We successfully conclude a sure prediction of a gene's expression based on the number of copies it has. This information is important in the context of specific target genes in cancer treatment. TableX in the appendix shows this inter-dependance for the popular oncogene *KRAS.*

Correlations between drug sensitivity and gene knockout sensitivity of drug's target

The correlations between drug sensitivity and gene knockout sensitivity of the drug's target are not high. The results can not give us reliable information the mechanism of action of the drug. As cells are extremely complex, there are hundreds of possibilities on why the correlations are so low. The only thing that we can say is, that there is not a high correlation between the drug sensitivity and its target gene gene knockout.

Dimensionality reduction by kmeans clustering

The dimensionality of the data can be reduced to genes of 4 clusters. Cluster 1 especially does show high importance for cell survival. If a a gene in this cluster is targeted, targeting another gene in this cluster can have similar effects on the cell. However, one can only make assumptions and similar effects are not guaranteed. Nevertheless, the dimensionality of the data can succesfully be reduced.

Finding biomarkers for drug sensitivity

There is a gene, that can possibly be a biomarker for drug sensitivity of liver cancer cell linen, named HRH2. We find that by analyzing if there are correlations of drug sensitivity correlates with the expression level of the drug target gene. The are some drug-target gene pairs with low correlations that are also differently expressed in liver cancer cell lines. These findings can only guide drug development, especially, because the correlation value of HRH2 with its according drug is not high.

Developing a Linear Regression Model

Plotting the linear regression results for this drug allowed us to visualize its impact on cell growth in liver cancer. This integrated approach, combining regression analysis, filtering based on reliability, and examining drug frequencies, provides insights into the drugs that exhibit strong regression relationships and their potential for effective treatment in liver cancer. By predicting the most efficient concentrations for a stronger inhibition, we can have different therapy possibilities for cancer. This allows us to estimate how different concentrations may affect cell growth and identify the optimal concentration for inhibiting cancer cell growth. By combining this knowledge with information on the mechanisms of action and target sites of these drugs, we can explore potential solutions for cancer treatment. This approach enables us to identify promising drug candidates and optimize their dosage strategies, offering new possibilities in the field of cancer therapy. It is also important to indicate that there were 17 cancer cell lines, we can only use 16 of them in our analyses though. The reason is that the liver cancer cell line ACH-000217 has false values in which all the drugs have the same effect (+0.5) on the cell growth for this cell line which doesn't seem to be reliable.

Determining Therapeutic Strategies

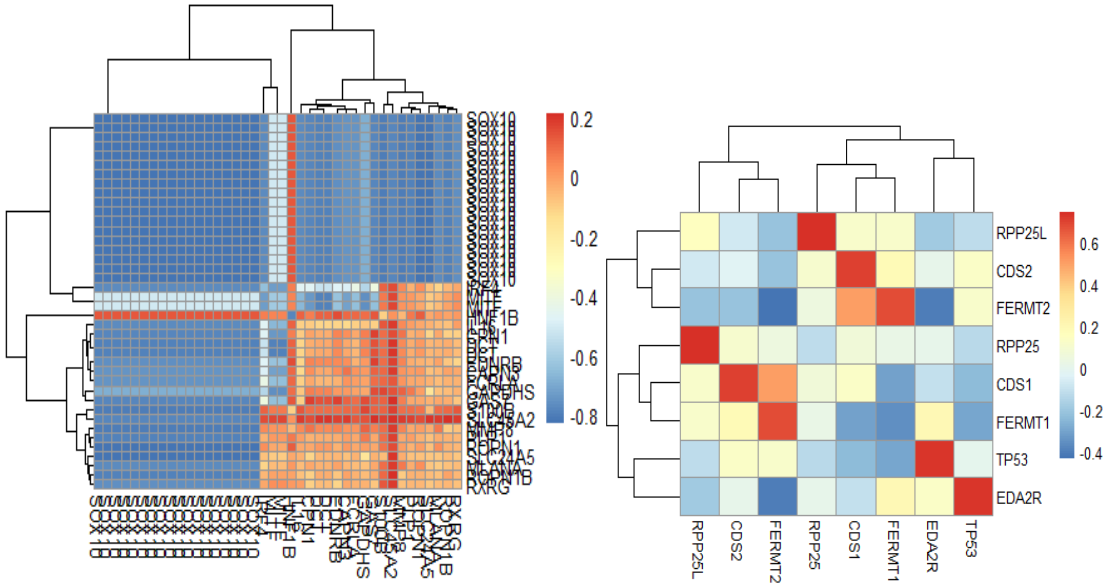
Understanding which targets are more commonly targeted by drugs allows us to identify potential opportunities for repurposing existing drugs for cancer treatment. If a specific target is highly frequent among different drugs, it suggests that this target might play crucial role in cancer biology. It helps us to identify key pathways and biological processes that are commonly dysregulated in cancer cells. Through the filtering of drugs with high R-square values using linear regression analysis, we have identified a set of reliable drugs for liver cancer cell lines. By examining the targets and mechanisms of action of these drugs, we can gain insights into how they affect cell growth in the context of liver cancer.

Among the most frequently occurring targets in liver cancer cell lines are ABL1,BCR, ATP1A1, BCL2, EGFR, HDAC1, MTOR and MAPK8. EGFR, a receptor tyrosine kinase, is frequently overexpressed or mutated in various cancers. Its aberrant activation leads to increased cell proliferation, survival, and metastasis, contributing to tumor progression and resistance to therapy. Inhibition of EGFR can potentially disrupt these pathways and inhibit the growth of liver cancer cells. In liver cancer, dysregulation of HDAC1 has been linked to aberrant gene expression patterns, promoting tumor growth and survival. Inhibiting HDAC1 activity can lead to the reaction of tumor-suppressive genes and hinder the growth and survival of liver cancer cells.

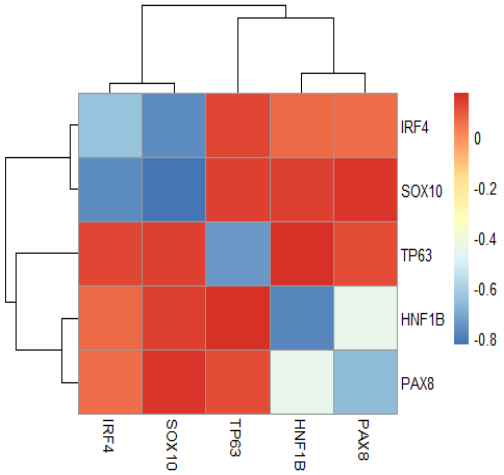
Another most frequent mechanism of action is the adrenergic receptor antagonist which acts by blocking the interaction of adrenergic receptors with their ligands. They can affect downstream signaling pathways involved in cell growth and proliferation. Naftopidil, one of the adrenergic receptor antagonists, has shown the ability to regulate the expression of pro-apoptotic members of the Bcl-2 family, which has potential applications in enhancing the effectiveness of targeted therapies and overcoming resistance to apoptosis in cancer cells.

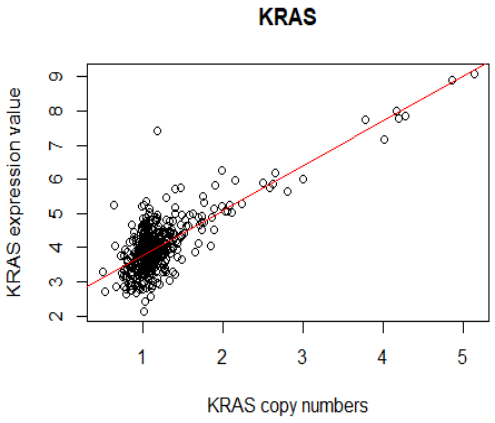
Appendix

Correlation and Anticorrelation of gene pairs corexpach



Anticorrelation genes with themselves





Pr ediciton model exp cnv KRAS

Table A.1: Selection of active molecules (A.) and mechanisms of action (B.) which the screened non-oncological drugs operate by and indices of the drugs which use them. The eight different positions of the same molecule or mechanism of action refer to the eight different concentrations in which a drug has been

employed, thus all eight indices belong to the same drug. If sixteen positions are being shown, then there are two drugs containing the same active molecule or operating by the same mechanism of action.

A.

```
##                       1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
## 8-bromo-cGMP          1  2  3  4  5  6  7  8 NA NA NA NA NA NA NA NA
## noretynodrel          9 10 11 12 13 14 15 16 NA NA NA NA NA NA NA NA
## prednisolone-acetate 17 18 19 20 21 22 23 24 NA NA NA NA NA NA NA NA
## betamethasone        25 26 27 28 29 30 31 32 NA NA NA NA NA NA NA NA
## mepivacaine          33 34 35 36 37 38 39 40 NA NA NA NA NA NA NA NA
## XL888                41 42 43 44 45 46 47 48 NA NA NA NA NA NA NA NA
```

B.

```
##                                                          1  2  3  4  5  6 7  8    9    10   11    12   13  14   15   16
## Abl kinase inhibitor, Aurora kinase inhibitor, FLT3 inhibitor  1  2  3  4  5  6 7  8   NA    NA   NA   NA   NA   NA   NA   NA
## Abl kinase inhibitor, Bcr-Abl kinase inhibitor                 9 10 11 12 13 14 15 16 4025 4026 4027 4028 4029 4030 4031 4032
## Abl kinase inhibitor, Bcr-Abl kinase inhibitor, src inhibitor 17 18 19 20 21 22 23 24   25   26   27   28   29   30   31   32
## Abl kinase inhibitor, src inhibitor, VEGFR inhibitor          33 34 35 36 37 38 39 40   NA   NA   NA   NA   NA   NA   NA   NA
## ACAT inhibitor                                                41 42 43 44 45 46 47 48 5105 5106 5107 5108 5109 5110 5111 5112
## acetylcholine receptor agonist                                49 50 51 52 53 54 55 56  177  178  179  180   181 182  183  184
```

Table A.2 : Cancer cell lines originating from liver cancer

```
##  [1] "ACH-000217" "ACH-000221" "ACH-000316" "ACH-000361" "ACH-000393"
##  [6] "ACH-000420" "ACH-000471" "ACH-000476" "ACH-000480" "ACH-000493"
## [11] "ACH-000537" "ACH-000620" "ACH-000625" "ACH-000671" "ACH-000734"
## [16] "ACH-000848" "ACH-001318"
```

Table A.4: Total mutation numbers in liver cancer cell lines

```
## ACH-000217 ACH-000221 ACH-000316 ACH-000361 ACH-000393 ACH-000420 ACH-000471
##        432        362        327        145        415        441        410
## ACH-000476 ACH-000480 ACH-000493 ACH-000537 ACH-000620 ACH-000625 ACH-000671
##        420        450        437        341        563        320        486
## ACH-000734 ACH-000848 ACH-001318
##        618       1092       1223
```

Table A.5: Differentially expressed genes in liver cancer. 50 of 1201 gene names are being displayed.

```
##  [1] "CEP63"    "TRAF3IP3" "CNTN2"    "MYLK4"    "CCDC113"  "PLCXD2"
##  [7] "DOK1"     "FOXP1"    "CDC42EP1" "ARRB2"    "ETV6"     "SLC2A8"
## [13] "ACP7"     "SLC29A4"  "FGFR4"    "TRIM13"   "SAV1"     "GAMT"
## [19] "DNAH2"    "RPLP2"    "ACSM2B"   "LY6K"     "TMED10"   "VNN3"
## [25] "MTG1"     "F13B"     "VTN"      "SERPINA1" "NID1"     "ADAM15"
## [31] "OR14J1"   "ZNF750"   "CMTM8"    "LYPD3"    "CIPC"     "CD36"
## [37] "ARHGEF19" "ZSCAN5A"  "CRHR1"    "CGAS"     "THEGL"    "GJB5"
## [43] "RTN4RL2"  "CMTR2"    "EIF3A"    "C16orf95" "ZFP3"     "PRM2"
## [49] "ADAT1"    "AGTR1"
```

Figure X shows a comparison between the top 50 most frequently mutated genes in liver cancer and the top 50 most frequently mutated genes across all cancer cell lines:

TableX the relevant correlation pairs

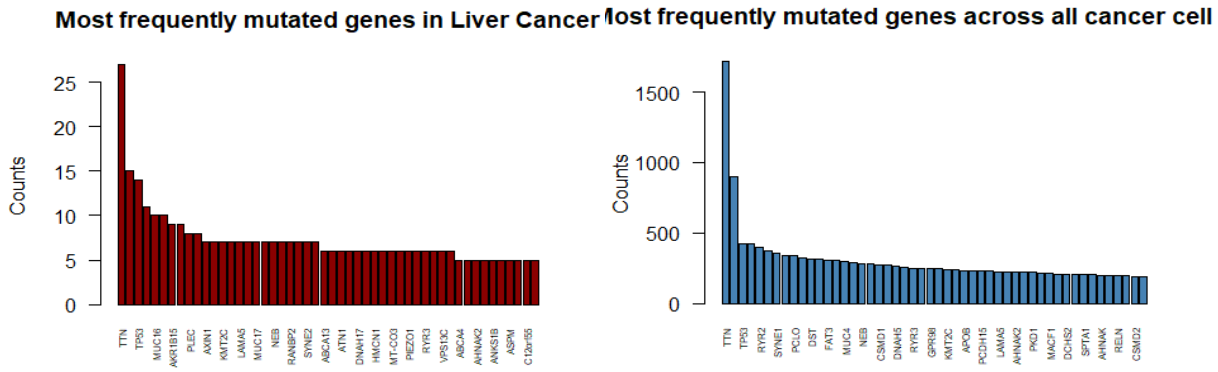
```
##      Pair 1 Pair 2 Pair 3  Pair 4   Pair 5   Pair 6   Pair 7   Pair 8
## [1,] "CDS1" "CDS2" "EDA2R" "FERMT1" "FERMT2" "RPP25"  "RPP25L" "TP53"
## [2,] "CDS2" "CDS1" "TP53"  "FERMT2" "FERMT1" "RPP25L" "RPP25"  "EDA2R"
```

TableX. the relevant anit-correlation pairs

```
relevant_gene_anticorrelation_pairs
```

```
##      Pair 1  Pair 2  Pair 3 Pair 4  Pair 5  Pair 6  Pair 7   Pair 8  Pair 9
## [1,] "CAPN3" "CPN1"  "DCT"  "DCT"   "EDNRB" "FCRLA" "GAPDHS" "GAS7"  "HNF1B"
## [2,] "SOX10" "SOX10" "MITF" "SOX10" "SOX10" "SOX10" "MITF"   "SOX10" "HNF1B"
##      Pair 10 Pair 11 Pair 12 Pair 13  Pair 14 Pair 15 Pair 16 Pair 17 Pair 18
## [1,] "IL16"  "IRF4"  "MITF"  "MITF"   "MLANA" "MMP8"  "PLP1"  "ROPN1" "ROPN1B"
## [2,] "SOX10" "SOX10" "DCT"   "GAPDHS" "SOX10" "SOX10" "SOX10" "SOX10" "SOX10"
```

```
##      Pair 19 Pair 20 Pair 21   Pair 22   Pair 23 Pair 24 Pair 25 Pair 26
## [1,] "RXRG"  "S100B" "SLC24A5" "SLC45A2" "SOX10" "SOX10" "SOX10" "SOX10"
## [2,] "SOX10" "SOX10" "SOX10"   "SOX10"   "CAPN3" "CPN1"  "DCT"   "EDNRB"
##      Pair 27 Pair 28 Pair 29 Pair 30 Pair 31 Pair 32 Pair 33 Pair 34 Pair 35
## [1,] "SOX10" "SOX10" "SOX10" "SOX10" "SOX10" "SOX10" "SOX10" "SOX10" "SOX10"
## [2,] "FCRLA" "GAS7"  "IL16"  "IRF4"  "MLANA" "MMP8"  "PLP1"  "ROPN1" "ROPN1B"
##      Pair 36 Pair 37 Pair 38   Pair 39   Pair 40
## [1,] "SOX10" "SOX10" "SOX10"   "SOX10"   "SOX10"
## [2,] "RXRG"  "S100B" "SLC24A5" "SLC45A2" "SOX10"
```



FigX

Here too we can find similarities. The following genes are not only under the top 50 genes that mostly suffer mutations in liver cancer, but also in all cancer cell lines:

```
##  [1] "TTN"    "OBSCN"  "TP53"   "PCLO"   "MUC16"  "MUC4"   "PLEC"   "KMT2C"
##  [9] "KMT2D"  "LAMA5"  "MUC17"  "NEB"    "SYNE1"  "SYNE2"  "ABCA13" "FAT1"
## [17] "HMCN1"  "RYR2"   "RYR3"   "AHNAK2"
```

APPENDIX

Here is an example for the cell line ACH-000393:

```
## Optimal threshold value: 0.9868978
```

Our output visualizes the optimal threshold value is the optimal R-squared value of the drugs for a specific cell line and the drugs above this threshold are listed below.

```
## Drugs above the threshold of R-squared value for the cell line ACH-000393:

BRD-A55312468-001-04-7, BRD-K00317371-001-06-1, BRD-K15567136-003-27-2, BRD-
K27182532-001-02-3, BRD-K36386086-001-01-1, BRD-K56032964-001-02-1, BRD-
K61443650-236-13-1, BRD-K70358946-001-15-7, BRD-K77908580-001-09-6, BRD-
K87782578-001-01-4
```

```
## Call:
## lm(formula = cell_growth_rates_all ~ concentrations_nM)
##
## Coefficients:
##       (Intercept)  concentrations_nM
##            0.3842            -0.2653
```

```
summary(regression_model_all)
```

```
##
## Call:
## lm(formula = cell_growth_rates_all ~ concentrations_nM)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.067380 -0.024960 -0.004618  0.021702  0.085684
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.384185   0.022172   17.33 2.37e-06 ***
## concentrations_nM -0.265287   0.006072  -43.69 9.63e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0558 on 6 degrees of freedom
## Multiple R-squared:  0.9969, Adjusted R-squared:  0.9963
## F-statistic:  1909 on 1 and 6 DF,  p-value: 9.626e-09
```

Our equation of the line is:

$y = -0{,}2653x + 0{,}3842$

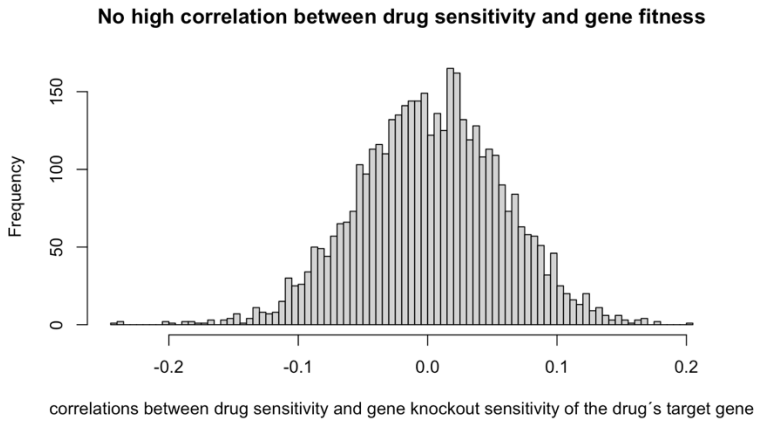


Fig B1: Correlations between drug sensitivity and gene fitness

```
Cluster means:
  achilles.mean  exp.mean cnv.mean
1   -0.57191479 7.3420959 1.026934
2   -0.02076931 0.4643571 1.010401
3   -0.27382636 4.8156963 1.019652
4   -0.08019922 2.8247483 1.010407
```

Fig B2: Results of kmeans clustering

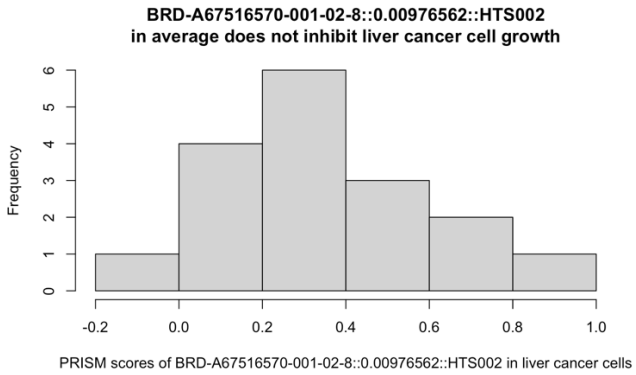


Fig B3: Efficiency of drug on liver cancer cell lines that has correlation between drug sensitivity on liver cancer cell lines and HRH2 expression

REFERENCES

References

Mossmann, D., et al. (2018). "mTOR signalling and cellular metabolism are mutual determinants in cancer." Nat Rev Cancer **18**(12): 744-757.

Zhao, C., et al. (2020). "Histone deacetylase (HDAC) inhibitors in cancer: a patent review (2017-present)." Expert Opin Ther Pat **30**(4): 263-274.

Harrison, P. T., et al. (2020). "Rare epidermal growth factor receptor (EGFR) mutations in non-small cell lung cancer." Semin Cancer Biol **61**: 167-179.

3. Inspect the correlation and anticorrelation of the genes with each other
ANTICORRELATION:

The anti-correlation between the expression and the Achilles scores could be interpreted as following:

i) GENE is under-expressed in the cancer cell line and the cancer cell line proliferates stronger when GENE is knocked out. This hints at the character of GENE to be that of a TUMOR SUPPRESSOR GENE. A tumor suppressor gene would be under-expressed in cancer cells and cancer cells would proliferate stronger if there is no tumor suppressor gene blocking their prosperity.

OR:

ii) GENE is overexpressed in the cancer cell line and the cancer cell line proliferates at a lower rate when GENE is knocked out. This hints at the character of GENE to be that of an ONCOGENE. Oncogenes would be over-expressed in cancer cells and cancer cells would not proliferate as well if there would be no active oncogene promoting their spread.

By knowing the values of one of the two, either the gene's expression or it's Achilles score, we could predict the character of that gene:

Prediction: HNF1B = Oncogene
Truth: HNF1B = Tumor Suppressor Gene
Lu W, Sun J, Zhou H, Wang F, Zhao C, Li K, Fan C, Ding G, Wang J. HNF1B inhibits cell proliferation via repression of SMAD6 expression in prostate cancer. J Cell Mol Med. 2020 Dec;24(24):14539-14548. doi: 10.1111/jcmm.16081. Epub 2020 Nov 10. PMID: 33174391; PMCID: PMC7754016.

Prediction: IRF4 = Oncogene
Truth: IRF4 = Both Tumor Suppressor Gene and Oncogene
Maffei, R., Fiorcari, S., Atene, C.G. et al. The dynamic functions of IRF4 in B cell malignancies. Clin Exp Med (2022). https://doi.org/10.1007/s10238-022-00968-0

Prediction: PAX8 = Oncogene

Truth: PAX8 = Oncogene

Bie LY, Li N, Deng WY, Lu XY, Guo P, Luo SX. Evaluation of PAX8 expression promotes the proliferation of stomach Cancer cells. BMC Mol Cell Biol. 2019 Dec 27;20(1):61. doi: 10.1186/s12860-019-0245-9. Retraction in: BMC Mol Cell Biol. 2021 Jun 01;22(1):31. PMID: 31881968; PMCID: PMC6935224.

Prediction: SOX10 = Oncogene

Truth: SOX10 = Oncogene (possible)

Cronin JC, Watkins-Chow DE, Incao A, Hasskamp JH, Schönewolf N, Aoude LG, Hayward NK, Bastian BC, Dummer R, Loftus SK, Pavan WJ. SOX10 ablation arrests cell cycle, induces senescence, and suppresses melanomagenesis. Cancer Res. 2013 Sep 15;73(18):5709-18. doi: 10.1158/0008-5472.CAN-12-4620. Epub 2013 Aug 1. PMID: 23913827; PMCID: PMC3803156.

Prediction: TP63 = Tumor Suppressor Gene

Truth: TP63 = Tumor SUppressor Gene

Melino, G. p63 is a suppressor of tumorigenesis and metastasis interacting with mutant p53. Cell Death Differ 18, 1487–1499 (2011). https://doi.org/10.1038/cdd.2011.81

=> This model could predict the main function of a gene regarding cell proliferation and tumorigenesis (up to a certain extent)

CORRELATION:

We see no relevant correlation bacause a correlation for the these datasets would mean contradicting effects on cell proliferation for the same genes.
(1 gene at 70% LTO1?)

Hypothesis model:

May a cancer cell proliferate strongly when GENE1 is knocked out and have a low expression score of GENE2. This suggests that GENE1 normally inhibits cell proliferation when it isn't knocked out and GENE2 must have a product that inhibits cell proliferation, as the cancer cell wouldn't promote genes that would inhibit its spread. So, the effect of GENE1 and GENE2 on cell proliferation must be SIMILAR.
INFO:

Example of hypothesis for the genes:

SOX10 promotes cancer cell proliferation:

Cronin JC, Watkins-Chow DE, Incao A, Hasskamp JH, Schönewolf N, Aoude LG, Hayward NK, Bastian BC, Dummer R, Loftus SK, Pavan WJ. SOX10 ablation arrests cell cycle,

induces senescence, and suppresses melanomagenesis. Cancer Res. 2013 Sep 15;73(18):5709-18. doi: 10.1158/0008-5472.CAN-12-4620. Epub 2013 Aug 1. PMID: 23913827; PMCID: PMC3803156.

Depletion of CAPN3 inhibits cell proliferation (so CAPN3 important for cell proliferation promotion):

"In the absence of CAPN3, immune cell influx is attenuated and expression of pro-myogenic factors is reduced. As a result, cell proliferation and myogenic differentiation is compromised, contributing to a growth deficit under reloading conditions observed in C3KO mice." cited from:

Irina Kramerova and others, Calpain 3 and CaMKIIβ signaling are required to induce HSP70 necessary for adaptive muscle growth after atrophy, Human Molecular Genetics, Volume 27, Issue 9, 01 May 2018, Pages 1642–1653, https://doi.org/10.1093/hmg/ddy071

CPN1 promotes cell proliferation:

Yang S, Yang H, Grisafi P, Sanchatjate S, Fink GR, Sun Q, Hua J. The BON/CPN gene family represses cell death and promotes cell growth in Arabidopsis. Plant J. 2006 Jan;45(2):166-79. doi: 10.1111/j.1365-313X.2005.02585.x. PMID: 16367962.

DCT promotes cell proliferation:

Jiao Z, Zhang ZG, Hornyak TJ, Hozeska A, Zhang RL, Wang Y, Wang L, Roberts C, Strickland FM, Chopp M. Dopachrome tautomerase (Dct) regulates neural progenitor cell proliferation. Dev Biol. 2006 Aug 15;296(2):396-408. doi: 10.1016/j.ydbio.2006.06.006. Epub 2006 Jun 8. PMID: 16857183.

But also: CAPN3 inhibits cell proliferation

Moretti D, Del Bello B, Allavena G, Corti A, Signorini C, Maellaro E. Calpain-3 impairs cell proliferation and stimulates oxidative stress-mediated cell death in melanoma cells. PLoS One. 2015 Feb 6;10(2):e0117258. doi: 10.1371/journal.pone.0117258. PMID: 25658320; PMCID: PMC4319969.

=> result may still be unclear at some points

=> If the function of specific genes in the group is known, this model can be used to up to a certain extent to predict the functions of the unknown genes they strongly anti-correlate to in the matrix.

Tp53 and eda2r

Brosh R, Sarig R, Natan EB, Molchadsky A, Madar S, Bornstein C, Buganim Y, Shapira T, Goldfinger N, Paus R, Rotter V. p53-dependent transcriptional regulation of EDA2R and

its involvement in chemotherapy-induced hair loss. FEBS Lett. 2010 Jun 3;584(11):2473-7. doi: 10.1016/j.febslet.2010.04.058. Epub 2010 Apr 29. PMID: 20434500.

RPP25L - function unclear, but interacts with RPP25
Köferle, Anna et al.,Interrogation of cancer gene dependencies reveals paralog interactions of autosome and sex chromosome-encoded genes
Cell Reports, Volume 39, Issue 2, 110636
i) in correlation pairs: the expression of the one gene is strongly correlated with the proliferation of the cell after the KO of the other gene. E.g. one cell proliferates strongly after CDS1 is knocked out and CDS2 is overexpressed. AND: one cell proliferates weakly after CDS1 is knocked out and CDS2 is underexpressed. This can hint at opposite roles of the two genes and their respective products in cell proliferation.
INFO:

This is indeed true for FERMT1 and FERMT2
FERMT1 - negative regulation of stem cell proliferation
https://amigo.geneontology.org/amigo/term/GO:2000647

FERMT2 - positive regulation of stem cell proliferation
https://amigo.geneontology.org/amigo/term/GO:1902462

Hypothesis: The expression of one of the genes promotes the expression of the other. So, even though the gene is KO'd, the other one can act as a substitute for the function of the KO'd gene=> upon KO of EDA2R, the cell proliferates weaker because TP53 is still expressed, even if underexpressed. OR: TP53 is KO'd and the cell still proliferates weaker because EDA2R is still expressed, even if underexpressed

RPP25L and RPP25 also interact with each other. However, the effect on cell proliferation is yet unclear:
RPP25 - negative regulation of cell proliferation
Xiao D, Wu J, Zhao H, Jiang X, Nie C. RPP25 as a Prognostic-Related Biomarker That Correlates With Tumor Metabolism in Glioblastoma. Front Oncol. 2022 Jan 12;11:714904. doi: 10.3389/fonc.2021.714904. PMID: 35096558; PMCID: PMC8790702.

RPP25L - function unclear, but interacts with RPP25
Köferle, Anna et al.,Interrogation of cancer gene dependencies reveals paralog interactions of autosome and sex chromosome-encoded genes
Cell Reports, Volume 39, Issue 2, 110636

Tp53 and eda2rBrosh R, Sarig R, Natan EB, Molchadsky A, Madar S, Bornstein C, Buganim Y, Shapira T, Goldfinger N, Paus R, Rotter V. p53-dependent transcriptional regulation of EDA2R and its involvement in chemotherapy-induced hair loss. FEBS Lett. 2010 Jun 3;584(11):2473-7. doi: 10.1016/j.febslet.2010.04.058. Epub 2010 Apr 29. PMID: 20434500.

SOX10 promotes cancer cell proliferation:
Cronin JC, Watkins-Chow DE, Incao A, Hasskamp JH, Schönewolf N, Aoude LG, Hayward NK, Bastian BC, Dummer R, Loftus SK, Pavan WJ. SOX10 ablation arrests cell cycle, induces senescence, and suppresses melanomagenesis. Cancer Res. 2013 Sep 15;73(18):5709-18. doi: 10.1158/0008-5472.CAN-12-4620. Epub 2013 Aug 1. PMID: 23913827; PMCID: PMC3803156.

Depletion of CAPN3 inhibits cell proliferation (so CAPN3 important for cell proliferation promotion):
Irina Kramerova and others, Calpain 3 and CaMKIIβ signaling are required to induce HSP70 necessary for adaptive muscle growth after atrophy, Human Molecular Genetics, Volume 27, Issue 9, 01 May 2018, Pages 1642–1653, https://doi.org/10.1093/hmg/ddy071

CPN1 promotes cell proliferation:
Yang S, Yang H, Grisafi P, Sanchatjate S, Fink GR, Sun Q, Hua J. The BON/CPN gene family represses cell death and promotes cell growth in Arabidopsis. Plant J. 2006 Jan;45(2):166-79. doi: 10.1111/j.1365-313X.2005.02585.x. PMID: 16367962.

DCT promotes cell proliferation:
Jiao Z, Zhang ZG, Hornyak TJ, Hozeska A, Zhang RL, Wang Y, Wang L, Roberts C, Strickland FM, Chopp M. Dopachrome tautomerase (Dct) regulates neural progenitor cell proliferation. Dev Biol. 2006 Aug 15;296(2):396-408. doi: 10.1016/j.ydbio.2006.06.006. Epub 2006 Jun 8. PMID: 16857183.

But also: CAPN3 inhibits cell proliferation
Moretti D, Del Bello B, Allavena G, Corti A, Signorini C, Maellaro E. Calpain-3 impairs cell proliferation and stimulates oxidative stress-mediated cell death in melanoma cells. PLoS One. 2015 Feb 6;10(2):e0117258. doi: 10.1371/journal.pone.0117258. PMID: 25658320; PMCID: PMC4319969.

Lu W, Sun J, Zhou H, Wang F, Zhao C, Li K, Fan C, Ding G, Wang J. HNF1B inhibits cell proliferation via repression of SMAD6 expression in prostate cancer. J Cell Mol Med. 2020 Dec;24(24):14539-14548. doi: 10.1111/jcmm.16081. Epub 2020 Nov 10. PMID: 33174391; PMCID: PMC7754016.

Maffei, R., Fiorcari, S., Atene, C.G. et al. The dynamic functions of IRF4 in B cell malignancies. Clin Exp Med (2022). https://doi.org/10.1007/s10238-022-00968-0

Bie LY, Li N, Deng WY, Lu XY, Guo P, Luo SX. Evaluation of PAX8 expression promotes the proliferation of stomach Cancer cells. BMC Mol Cell Biol. 2019 Dec 27;20(1):61. doi: 10.1186/s12860-019-0245-9. Retraction in: BMC Mol Cell Biol. 2021 Jun 01;22(1):31. PMID: 31881968; PMCID: PMC6935224.

Cronin JC, Watkins-Chow DE, Incao A, Hasskamp JH, Schönewolf N, Aoude LG, Hayward NK, Bastian BC, Dummer R, Loftus SK, Pavan WJ. SOX10 ablation arrests cell cycle, induces senescence, and suppresses melanomagenesis. Cancer Res. 2013 Sep 15;73(18):5709-18. doi: 10.1158/0008-5472.CAN-12-4620. Epub 2013 Aug 1. PMID: 23913827; PMCID: PMC3803156.

Melino, G. p63 is a suppressor of tumorigenesis and metastasis interacting with mutant p53. Cell Death Differ 18, 1487–1499 (2011). https://doi.org/10.1038/cdd.2011.81

Shen J, Qi L, Zou Z, Du J, Kong W, Zhao L, Wei J, Lin L, Ren M, Liu B. Identification of a novel gene signature for the prediction of recurrence in HCC patients by machine learning of genome-wide databases. Sci Rep. 2020 Mar 10;10(1):4435. doi: 10.1038/s41598-020-61298-3. PMID: 32157118; PMCID: PMC7064516.