Drug sensitivity in cancer cell lines

D'Antoni, Marquard, Neumann, Zymela

19 7 2021

Introduction:

Developing new drugs against certain diseases always is combined with high costs, hard work and complicated requirements before even reaching the market. Therefore an alternative approach of repurposing already approved drugs to new diseases displays an attractive method to save time and resources. On top of that by researching further with these already licensed drugs most safety concerns can be eliminated, since the medications have been thoroughly tested in previous investigations. All in all the potential that drug repurposing possesses helps researchers to create new therapy approaches without the excessive use of follow-up investigations (Pushpakom et al., 2019).

In our research, we worked with multiple datasets that display among other things the growth-inhibitory activity of 4518 different drugs against 481 human cancer cell lines. Here we directed our attention towards pancreatic cancer cell lines, of which there were 33 found in the data provided (Corsello et al., 2020).

The most important datasets we focused on in our research were:
- the prism Dataset -> containing the treatment effects of the different cell lines
- the prism.exp Dataset -> containing the gene expression levels for every cell line gene
- the prism.achilles -> containing the gene knockdown score for all cell line genes

The following questions we tried to answer over the course of our research:
- can we identify certain genes that are related to pancreatic cell lines
- can we determine treatments against these pancreatic cell line genes

Data Clean-up:

The first step in our research was to clean up the Datasets. This meant to get rid of all N/A values that would interfere with our findings. The prism.exp dataset did not contain any of these missing values, hence no further clean up steps were necessary. The prism and the prism.achilles dataset on the other hand did possess some N/A values. Here we [. . . . . .]

Descriptive Statistics:

Afterwards we tried to visualize the given datasets to obtain a greater knowledge about their distribution using different dicriptive statistical methods.

Prism dataset

Here we tried to display [. . . . . .]

Using a statistical test to greatly reduce the number of genes of interest

After basic data cleanup and evaluation of the data set prism.exp, which contained expression data of the cell lines on genes, we were looking for meaningful criteria with which to reduce the data set. It was clear that our group should initially focus on the pancreatic cancer cell lines. A reduction by pancreatic cell lines diminished the expression data set from 477 to 33 rows. To achieve this, we first used the metadata set prism.cl to get Depmap IDs, which are assigned to pancreatic cancer. Since the row names of the prism.exp data set received the Depmap ID's, we could cut the data set down.

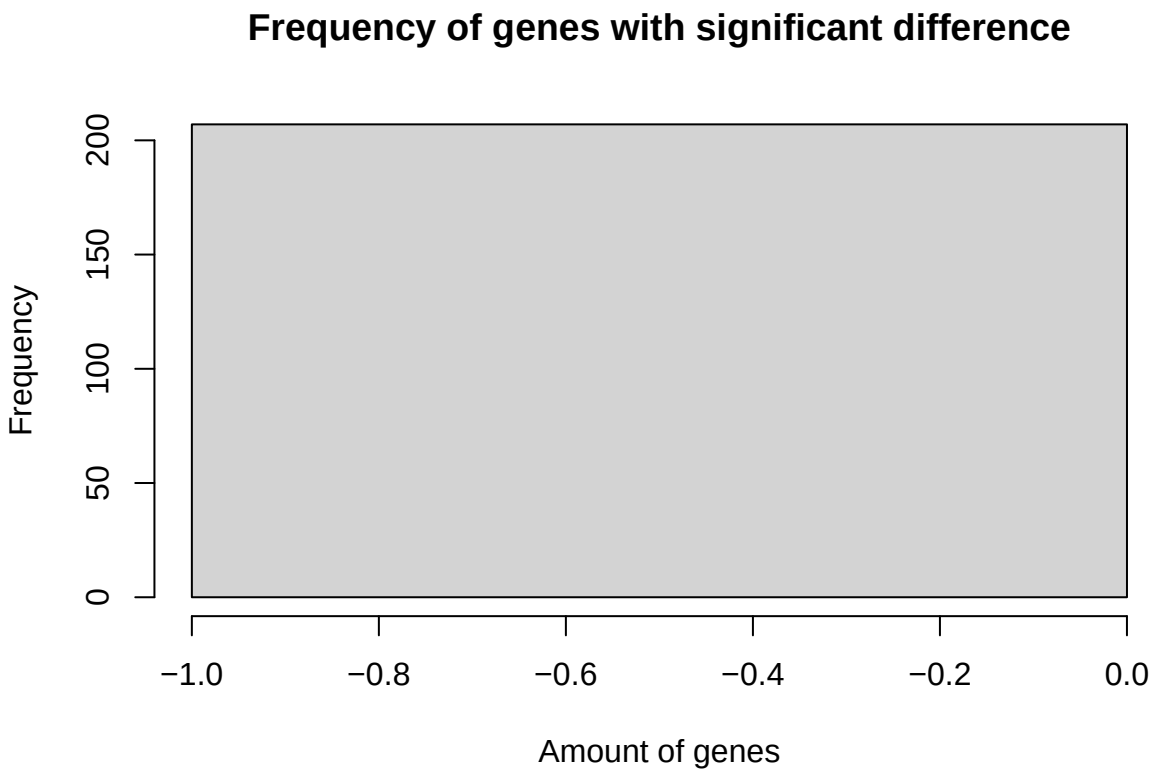
In contrast, identifying interesting genes that are relevant to us turned out to be more difficult. First, it had to be clear what exactly was being searched for. Above all, the aim was to find genes that had distinctive expressions. Such genes whose expression differed significantly from other cancer cell lines. For such a task there are competent methods in statistics, which were also discussed in our bioinformatics lecture. With the help of statistical tests, it is possible to test hypotheses between values of two cohorts and to calculate a p-value, which represents a probability with which a hypothesis can be rejected. For this reason, a corresponding H0 hypothesis was set up for the expression data, paying particular attention to the mean values.

H0: "The mean expression of gene X between pancreatic cancer cell lines and non-pancreatic cancer cell lines is equal"
H1: "The mean expression of gene X between pancreatic cancer cell lines and non-pancreatic cancer cell lines is different"

One gene from the cohort of 33 pancreatic cancer cell lines and a cohort of 33 randomly selected other cancer cell lines were examined for significant deviations from the mean. Since we were using non-normalized values, we had to use a non-parametric test. Furthermore, we are going to use two-samples, which are unpaired. For that reason, we were using the Wilcoxon Rank Sum Test. In this process, the values of the two groups are combined and given a rank, whereby the information on group membership is not lost. A test statistic is calculated from this, which takes the sum of the ranks in a group into account. The result of the test statistic is compared with critical values, the smaller the result, the more significant the H0 hypothesis can be rejected.

First, in addition to the data frame, which contained cell lines and genes, a data frame with the same number of genes in the same order but with 33 other randomly selected cancer cell lines had to be made. A prism.exp data frame was created for this, which did not contain any pancreatic cancer cell lines. 33 cell lines were randomly selected from this new data frame using the sample function. So, we got two cohorts.

33 values of a gene in one group were compared with 33 values of the same gene in the other pancreatic cancer-free group using the Wilcoxon rank sum test. That alone was not enough to find interesting genes. We had to do the same with all 19177 genes in the data frame. So, a multiple hypothesis test. For this reason, a for loop was used, which took over the procedure column by column. For each gene, the calculated p-value was put into a variable, so that in the end a variable with 19177 values was obtained. Now we can convert the p-values into Boolean values, according to whether these values reach a given significance level alpha or not. It must be noted that the significance level must be adjusted in the case of multiple hypothesis tests, as otherwise the results will be greatly changed by false positive values at high implementation rates. We simply used the Bonferroni correction, whereby the significance value alpha is divided by the number of times, in this case 19177. Any NA values have been replaced by FALSE. This made it possible to produce a vector that contained the names that were significant. The colnames vector of one of the cohorts that had the same sequence of gene names was used for that.



We can see in this distribution that there is a small group of genes that regularly differs in the wilcoxon test. Therefore we can reject our previously established H0 hypothesis.

```
head(sorted.df.interesting.genes)
```

```
##   interesting.genes Freq
## 1              ACTB    1
## 2            ACTRT2    1
## 3             AGAP1    1
## 4           ALDH3A2    1
## 5              ALG8    1
## 6            ALKBH3    1
```

The names of the top interesting genes can be viewed in this data frame. Here for each gene is the number of times the H0 hypothesis was discarded for this gene after carrying out onehundred.