Topic 1 Group 3: Drug viability screens for oncological and non-oncological treatments

Cedrik Neber, Lea Ahrens, Lennard Kleemann, Ilya Schneider, Xenia Quaas

19th July 2021

Introduction

Drug development is a time and investment consuming procedure. To tackle this problem, various new strategies have been explored. Remdesivir, originally developed to fight Ebola, has recently been employed to treat COVID-19 infected patients. The process of using an existing drug to combat an alternate disease is known as drug repurposing.

Numerous advantages of this method include the reduced time and decreased likelihood to fail in research and development. Using the previously gathered data from a known drug additionally reduces the overall costs. In this context, computational approaches are gaining in importance, as they allow large amounts of information from repurposing assays to be analyzed (Pushpakom *et al.*, 2019).

According to the Global Cancer Statistics 2020, brain cancer is a rare cancer type, accounting for 2.5% of all new cancer cases. Nevertheless, its mortality rate is comparatively high, which makes drug repurposing a promising approach to treat this disease.

Project structure

This project uses seven data sets generated by the Broad Institute using the PRISM strategy. The aim is to investigate whether it is possible to predict the effectiveness of a drug in brain cancer treatment.

To explore this question, four milestones were determined and explored:

- **How can we distinguish the most effective drugs?**
- **Are there any genetic markers that are specific for brain cancer subtypes?**
- **What are the targets of the effective drugs?**
- **What other factors contribute to drug and effectiveness prediction?**

Data clean-up

As the data sets also contain information about other cancer types, irrelevant cell lines were discarded. Furthermore, the data sets were classified in accordance with the brain cancer subtypes. Cell lines with NA values were either removed or replaced with a mean value, depending on the data set. The third milestone takes a closer look into brain cancer subtypes. To facilitate this step, additional data frames containing the individual subtypes were created.

Identification of the effective drugs

The `prism` data frame gives information on treatments, which were used in the original effectiveness screening. These treatments include 4518 drugs with dosages ranging from 0.00061034 μM to 10 μM. Analysis of the data identified 8 standard dosages, which were then divided into separate data frames and stored in a list.

Several dosages were identified, that did not correspond to the standard set of dosages. These outliers were assigned to the dosage with the least deviation, respectively. The idea of the assignment is to analyse the provided data via the drugs, as opposed to the assigned dosages. In this data set, the lowest values indicate the highest effectiveness. To identify the most effective drugs, the values of the `prism` data frame were analysed. Contrary to the threshold value 0.3 of Corsello *et al.* the threshold in this investigation was set to 0.2. This value was found to correspond to the median of the `brain_cancer` data frame (filtered brain cancer cell lines from the `prism` data frame), allowing the removal of over half of the provided drugs in this step.

In order to reduce the number of drugs and to be able to make accurate predictions later on, only drugs were selected that are effective in all doses. First, the drugs whose mean effectiveness across all cell lines in the dose subset data frames were identified. Their effectiveness was less than or equal to the established threshold. Second, these drugs were compared to each other and only drugs present in all subsets were kept, allowing the identification of **51 drugs**. These drugs were saved in the `effective_in_all_doses` vector. Subsequently, a single dosage was chosen for the further analysis, as the model aims to establish the output of drugs suited for treatments, regardless of their concentration. Dosage 7 (10 μM) was selected due to its high variance, which was much higher than other dosages.

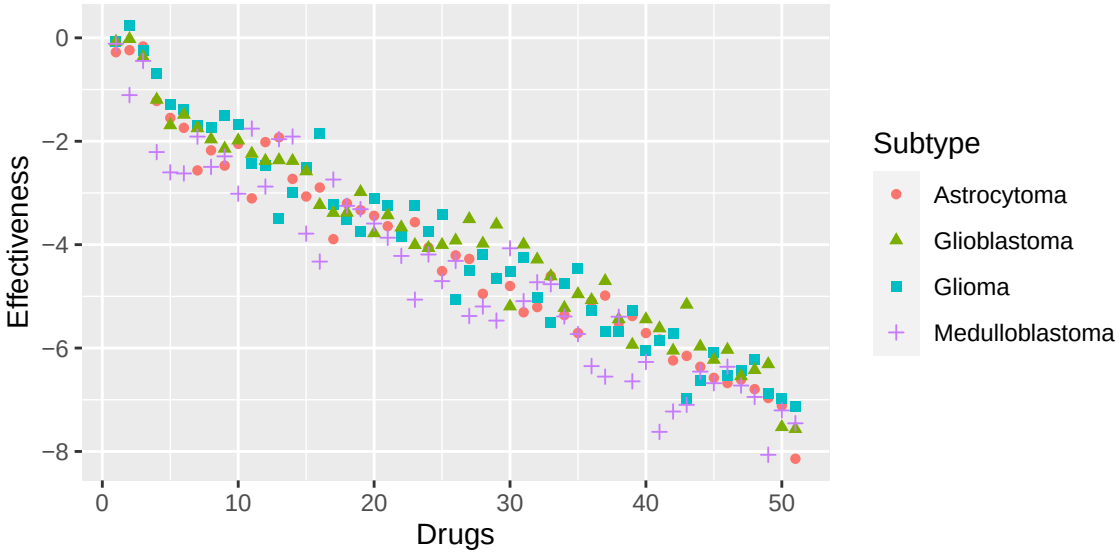

Figure 1: Average subtype effectivness per drug

Figure 1 represents the average effectiveness of each drug for each subtype. The means of the identified drugs have effectiveness scores ranging from 0.2 to approximately -8.2. The graph depicts the drugs ranking in accordance to their individual effectiveness. Overall, there is a general tendency for medulloblastoma cell lines to be more effected by the drug treatment. For 31 out of the 51 drugs, medulloblastoma cell lines show the highest effectiveness scores. The medulloblastoma subtype merely comprises two cell lines, as opposed to the glioblastoma subtype, with over 20 cell lines. Hence, the shown medulloblastoma means are only dependent on two values.

Identification of genetic markers

For this milestone, a further look at brain cancer subtypes was taken. This information could be relevant for the final regression model. The column `disease_subtype` of the `prism.cl` data set gives information about the subtype, whereby a distinction is made between astrocytoma, medulloblastoma, glioblastoma and glioma.

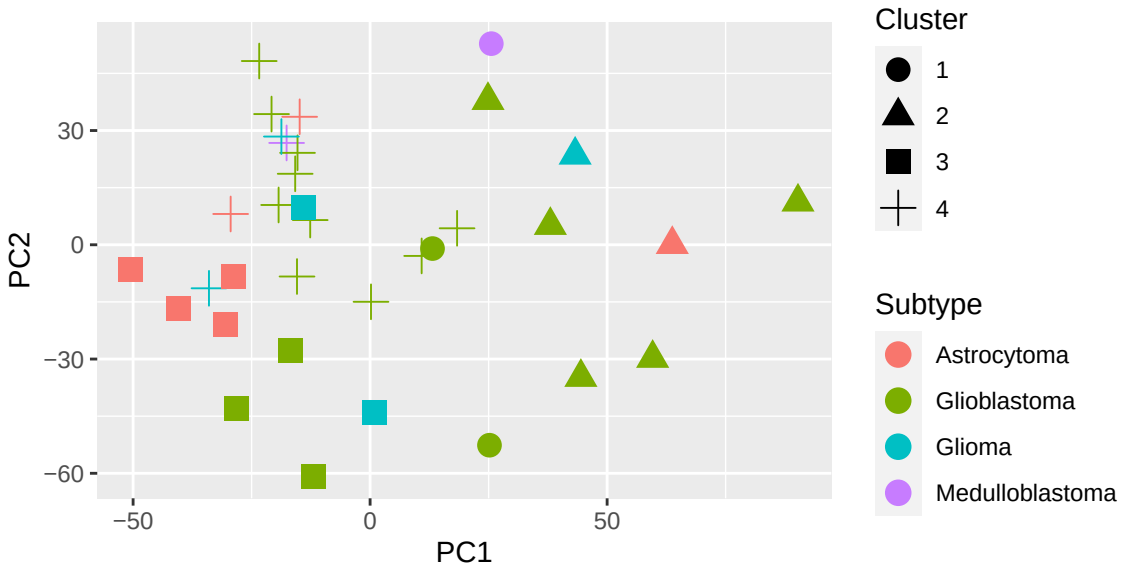

Figure 2: Gene expression clustering

Figure 2 shows the clustering of the two principle components (PC) with the highest variance of the cell lines by subtype. The data used is the expression data from the brain cancer subtypes. In the first step, a principle component analysis (PCA) was conducted to reduce the dimensionality of the data frame before using the kmeans method to cluster the obtained PC's. This method divides observations into clusters by assigning them to the cluster with the smallest euclidean distance to a cluster centroid. The number of expected clusters was predefined as 4. This was chosen as the graph aims to analyse the relationship of the gene expression of the cell lines and the subtypes they belong to.

As can be seen, there is no link between the clusters and the subtypes. This makes the continuous analysis using the subtypes obsolete. Furthermore, the clustering in itself does not seem to be conclusive (see cluster 1). Even an alternate number of itertions in the kmeans did not change that.

Shapiro Wilks test

To find out which genes are expressed significantly differently in brain cancer cell lines compared to other cancers, a statistical test can be applied. For this purpose, it must first be examined whether the available data is normally distributed. A statistical test, the Shapiro Wilks test, can be used for this evaluation. In this case, the H0 hypothesis states that the distribution is a normal distribution. If the p-value is smaller than the significance level, the H0 hypothesis is rejected and it is assumed that the data is not normally distributed. The test was applied to the 19177 genes from the `prism.exp` data set to determine the expression across the 34 cell lines. A significance level of 0.05 was set. As a result, 9799 p-values were less than or equal to this value, thus approximately half of the genes were normally distributed. This is the reason a non-parametric test was conducted.

Wilcoxon Rank Sum test

An unpaired Wilcoxon Rank Sum test was used to identify differential gene expression. In this test two different data sets were compared. The H0 hypothesis is that expression values of both samples come from the same distribution. The first data set was `non_bc_exp`, which has all the cell lines from the `prism.exp` data set, except from the brain cancer disease type. There are 447 cell lines in total from lung, skin, pancreatic and ovarian cancers. The second data set is the `brain_cancer_exp`, which reflects the level of gene expression of the 34 brain cancer cell lines. Although these data sets are not identical in size, this is not problematic as the test is based on ranking the individual values. Some genes have an expression value of 0. To avoid following complications (for example log2-fold-change calculation) the lowest expression value for each data set, so called `pseudo_count`, was added. The pseudo counts were equal for both data sets.

In this non-parametric test the H0 hypothesis is tested simultaneously for all genes. This leads to a multiplicity problem, in which the probability of false positive test results is increased. To adjust the number of significant p-values occurring for this problem, two types of corrections were considered. The Benjamini-Hochberg and the Bonferroni corrections.

The number of genes that are significantly different in their expression differs by a large amount. The Benjamini-Hochberg correction results in 8948 genes, meanwhile the Bonferroni correction in **1859 genes**. The Bonferroni correction was chosen, since it is more conservative and controls the type-I-error. Therefore the number of false positives (keeping genes that are not significant) is avoided.

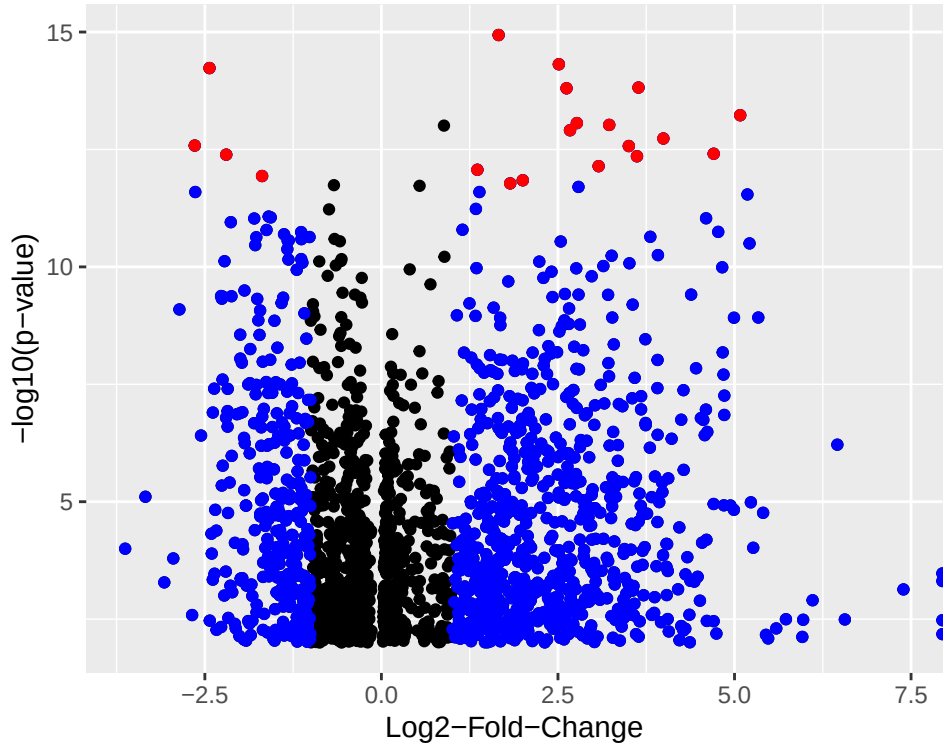

Figure 3: Volcano plot of significant differential expressed genes

To visualize differentially expressed genes a volcano plot can be used, as shown in Figure 3. Each gene expression difference can be represented by the log2-fold-change on the x-axis. The y-axis displays the -log10 of each p-value. First, the genes were filtered according to the log2-fold-change. Only the genes, which value was less or equal to -1 and greater or equal to 1 were kept. These genes are shown in blue on the plot. Second, the top 20 genes with the biggest -log10 p-value were extracted and highlighted in red.

Linear regression model with universal drug

From these top 20 genes, the ones with the most influence on drug effectiveness were selected. An effective drug from dosage 7, that has an overall consistent effect on the cell lines, needs to be determined. Therefore, a "universal drug" was filtered out by dividing the range of effectiveness scores into equally sized intervals. For each drug an interval with the highest amount of responding cell lines was selected. The determined universal drug is the one with the biggest count (Broad ID "BRD-K79145628-001-05-5").

The same method was applied to the non brain cancer cell lines and after that the two universal drugs were compared. They turned out to be different, which indicated that the brain cancer universal drug is specific for this cancer subtype.

This drug was finally used as a reference variable for the linear regression model, in which drug effectiveness is predicted by the expression of the top genes. In this case, each gene is considered as a single variable. By setting up a multiple regression model, it is important to note that no correlated variables should be used. Among the top 20 genes, a high correlation between the genes "FAM83H" and "IQANK1" was found. The gene with a more significant p-value in the Wilcoxon test was kept, the other removed.

The application of the linear regression model involves feature selection. It consists of a reduction in the number of variables, so that only those that contribute the most to the prediction are included in the model. The p-value of the F-test was used for the evaluation, since it compares the linear model with the null model. In the F-test the H0 hypothesis is that the tested model does not perform better than the null model, therefore a small p-value leads to the rejection of the H0 hypothesis.

```
# linear regression model
initial_regression_1<-summary(lm(Drug~.,genes_regression)) # 1. (model A)

# feature selection
repeat{
  end_regression_1<-initial_regression_1

  pvx<-pf(end_regression_1$fstatistic[1],end_regression_1$fstatistic[2],
          end_regression_1$fstatistic[3],lower.tail=FALSE)

coeffs=as.data.frame(end_regression_1$coefficients)
coeffs=coeffs[-c(1),]

coeffs=coeffs[order(coeffs$`Pr(>|t|)`),] # 2.
coeffs=coeffs[-c(nrow(coeffs)),] # 3. (lowest one removed)
genes_regression=cbind(brain_cancer_exp[,rownames(coeffs)],ae_d7[,uni_drug])
colnames(genes_regression)[ncol(genes_regression)]="Drug"

initial_regression_1<-summary(lm(Drug~.,genes_regression)) # 3.(model B)
pvy<-pf(initial_regression_1$fstatistic[1],initial_regression_1$fstatistic[2],
        initial_regression_1$fstatistic[3],lower.tail=FALSE)

if(pvy>=pvx){ # 4.
  break # 5.
  }
}

end_regression_1

final_genes=rownames(as.data.frame(end_regression_1$coefficients)) # 6.
final_genes=final_genes[-1]
```

The following strategy was applied:

1. All 19 genes were the input for the model (model A).
2. Genes were ordered according to their own p-values, from the highest to the lowest.
3. The lowest one was removed and the new linear model was executed (model B).
4. If the overall p-value of model A was bigger than p-value of model B, the procedure was repeated starting from the second step.
5. This loop stopped the moment the overall p-value of model A was smaller than p-value of model B.
6. The model reached its optimal point.

7 final genes arise from this model with a p-value of **0.01353**. All 447 cell lines underwent the same strategy with the linear regression. As a result, a drastically lower r-squared value with only 3 genes was determined to be the optimal model in this case. This indicates the specificity of these 19 genes in the brain cancer cell lines.

To further check this result, the full model was compared with the reduced model using the `anova` function. It compared the variances of the residuals considering the introduced degrees of freedom. The H0 hypothesis was that the two models are equivalent. A p-value of **0.9774** was obtained, which suggested that the reduced model performs significantly better in comparison to the full one.

Identification of drug targets

The 51 effective drugs, which were identified in question 1, were used to determine their specific targets. The `prism.treat` delivered this information. There were drugs which attack more than one target or had no target available. Some targets occur for different drugs. In total, **124 targets** for all effective drugs remained.

To explore these targets, three other data frames where used. The `brain_cancer_achilles` gives information on gene knockdown scores, which indicates the importance for cell survival. The `brain_cancer_cnv` contains gene copy number values and therefore reflects genetic alterations. Last, the `brain_cancer_exp` data set was used, which was already used for the identification of genetic markers. Not all values were available for the targets, therefore "ATP5A1", "MMP12" and "MMP23A" were not utilized for further analysis.

Analysis of the targets

To find out whether the distribution of the target values from one of the data sets had the same distribution as one of the complete data sets, three QQ-plots were created. This is a graphical analysis which compares two probability distributions by their quantiles.

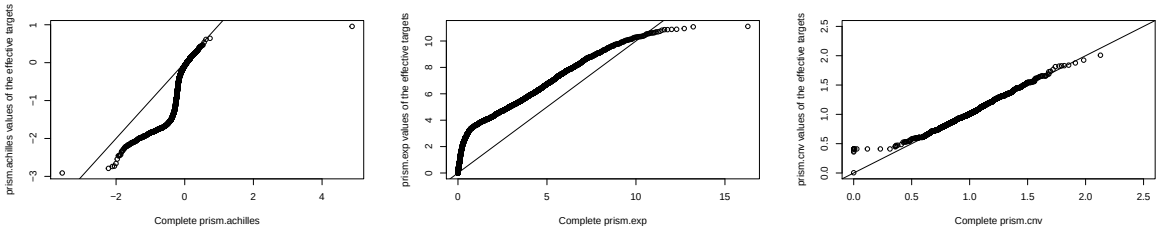

Figure 4: Comparison of the distribution of target knockout, expression values and copy number values of the effective targets with all targets

The visualizations made clear, that the distribution of the effective targets from .exp and .achilles are not representative for the complete data set. A similar shape of both distributions can only be accepted if the QQ-plot is a straight line. Therefore, these two parameters seem to have a correlation to the effectiveness.

To further analyse this correlation, another Wilcoxon Rank Sum test was applied. For the .achilles a one-sided test was employed and resulted in 59 targets out of 121, which derive from the distribution of the complete data set. The one-sided test was used, because only very negative proliferation values are relevant in gene knock-outs. After applying the Bonferroni correction, only **44 targets** were left. For the .exp and the .cnv data a two-sided Wilcoxon test was applied. The corrected values are **98 targets** for .exp and **6 targets** for .cnv, which did not confirm the H0-hypothesis. In conclusion, the effectiveness correlates, in a certain way, with the .achilles and the .exp data. However, there are still other parameters, which contribute to the effectiveness for specific target.

Linear regression model for effectiveness prediction

The first idea was to predict the effectiveness based on a chosen cell line and a drug. However, his idea was discarded because the three data sets `prism.achilles`, `prsim.exp` and `prism.cnv` depend only on the targets and not the drugs. Therefore, the decision was made to predict the effectiveness based on one cell line and one target. Hence, the value for one target of each of the three data sets is specific and depends on a single cell line. The `prism` values, which give information about the effectiveness, were averaged because more than one drug can have the same target. The focus was only set on the highest drug dosage.

To reduce the amount of input values effectively with constant p-values and constant proportion of the explained variance, a PCA was applied. The calculated PC values were used to train the model. It is visible that the model only describes a small proportion of the variance, with a adjusted r-squared value of **0.1852**. However, the regression model is still better than the null model, as indicated by the p-value of the F-statistic of **<2.2 e-16**.

```
##
## Call:
## lm(formula = prism ~ IQANK1 + RAB11FIP1 + FOXG1 + LPIN3 + ERVMER34.1 +
##     PTN + VRK2 + achilles + exp + cnv + prism, data = dat.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8611 -1.2014  0.0963  1.2673  4.6105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.10904    0.44058  -9.326  < 2e-16 ***
## IQANK1      -0.08639    0.03224  -2.680  0.00745 **
## RAB11FIP1   -0.07422    0.04827  -1.538  0.12438
## FOXG1        0.01014    0.06036   0.168  0.86664
## LPIN3        0.05367    0.07831   0.685  0.49326
## ERVMER34.1   0.21559    0.04487   4.804 1.71e-06 ***
## PTN          0.20544    0.17612   1.166  0.24361
## VRK2         3.99364    0.97331   4.103 4.30e-05 ***
## achilles     0.97372    0.07801  12.482  < 2e-16 ***
## exp         -0.05962    0.01839  -3.242  0.00121 **
## cnv          0.29402    0.19266   1.526  0.12720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.753 on 1468 degrees of freedom
## Multiple R-squared:  0.1907, Adjusted R-squared:  0.1852
## F-statistic: 34.59 on 10 and 1468 DF,  p-value: < 2.2e-16
```

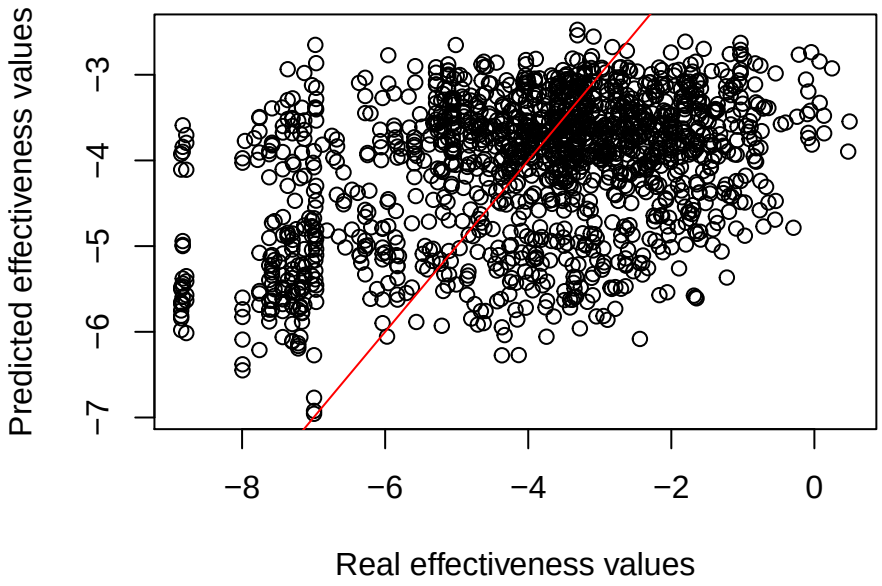

Figure 5: Comparison between real and predicted values

Discussion

This project unites drug repurposing and precision medicine. In the context of this report, precision medicine is defined as the individualized treatment for brain cancer patients, depending on their genetic profile. Meanwhile, the main question of whether it is possible to predict the effectiveness of a drug by means of various variables is exploited.

The foundation of this investigation was to analyse different contributing factors, that will become the variables of the final linear regression model.

Although looking at the dosage as a factor holds a lot of potential, this report is limited by the fact that only the dosage with the highest variance was utilized. More in depth exploration of the dosage contribution could lead to an improvement of the final model, since the data is limited to only 8 identified dosages.

Figure 1 underlines the superior effectiveness of medulloblastoma cell lines in comparison to the other subtypes with the evident quantitative limitations for this subtype. In summary, there is no conclusive evidence to support the claim that medulloblastoma cell lines are more likely to positively respond to drug treatment.

To assess the informational yield of the brain cancer subtypes, a principal component analysis, coupled with a kmeans clustering, was conducted. The overall goal of this clustering was to take a closer look at the relationship of the subtypes to another. During these analytic steps, it became apparent that there is no correlation between the subtypes, regarding their gene expression. An explanation might be that the brain cancer subtype glioma can be divided in astrocytoma and glioblastoma. At this point, the theory is that the description “glioma” is employed whenever there is a classification missing. As these three subtypes belong to the same class of cancers and the clustering did not lead to another result, it was decided to not take the subtypes into account.

As an additional contributing component, cell lines were examined regarding their genetic markers. A “universal drug” was discovered along the way. The universal drug is called **nimorazole** and is a non-infectious

drug, which acts as bacterial DNA inhibitor. It can be used as radiosensitizer to improve radiotherapy in head and neck squamous cell carcinoma. Nimorazole enhances cell death by binding to damaged DNA and inhibits its repair mechanisms. Despite the fact that the drug shows consistent effectiveness throughout the cell lines, it shows very little effect in regards to treating cancer in the context of the other effective drugs. As a result, all seven genes used in the final regression model are solely based on this drug and could similarly create a bias towards this particular drug.

Analysis of the drug targets implicates a correlation between the .exp and .achilles values. This correlation makes these targets an interesting factor for the final regression model. As stated by Maughan (2017), the main difficulty in treating brain cancer are the interactions and relationships between all genetic factors involved. Finding an effective and specific cancer drug like Trastuzumab for breast cancer treatment is extremely rare and not very likely. Another problem is the quick adaptation of cancer cells to treatment options, in order to develop a drug resistance.

Comprising the previously described factors, the final regression model was created. The model, as shown in the report, is severely limited by the minimal amount of variance it shows. One reason for this could be that the data offers known information and drug repurposing is mostly based on unknown effective mechanisms. Maybe a more complex method of machine learning could find a higher correlation between our given parameters. Because such a model is based on a specific number of parameters, it can only support real experiments.

The relevance of this regression model to predict the effectiveness of a drug is to minimize costs in drug redevelopment, which can sum up to $100,000 annually in the US alone (in 2012), as described by Workman *et al.*. This can not only reduce the costs, but potentially also avoid the risks of side effects in cancer patients by providing a personalized treatment plan.

The acquired data suggest that the regression model is not suitable to provide reliable treatment options or identify potential drugs by itself. It could, however, be a very efficient guide in the selection process of drug redevelopment. In addition, exploration of specific brain cancer genes could prove to be a more reliable variable in the developed regression model. In conclusion, precision medicine is the rising star in oncology and holds great promise for its treatment. However, this report highlights the limitations of the approach. Even when combining the six data frames used in this analysis, no reliable regression model could be created.

Literature

Chen, R., Smith-Cohn, M., Cohen, A.L., and Colman, H. (2017). Glioma Subclassifications and Their Clinical Significance. Neurotherapeutics 14, 284-297.

Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., ... Golub, T. R. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. Nature Cancer, 1(2), 235–248.

DeWeerdt, S. (2018). The genomics of brain cancer. Nature, 561(7724), 54-55.

Fang, Z., Du, R., and Cui, X. (2012). Uniform approximation is more appropriate for Wilcoxon Rank-Sum Test in gene set analysis. PLoS One 7, e31505.

Liang, C., Tian, L., Liu, Y., Hui, N., Qiao, G., Li, H., Shi, Z., Tang, Y., Zhang, D., Xie, X., Zhao, X., A promising antiviral candidate drug for the COVID-19 pandemic: A mini-review of remdesivir. (2020). Eur J Med Chem., 201, 112527.

Maughan, T. (2017). The Promise and the Hype of 'Personalised Medicine'. New Bioeth 23, 13-20.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., ... Pirmohamed, M. (2019). Drug repurposing: progress, challenges and recommendations. Nature Reviews Drug Discovery, 18(1), 41–58.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71, 209-249.

Tan, Y.D., and Xu, H. (2014). A general method for accurate estimation of false discovery rates in identification of differentially expressed genes. Bioinformatics 30, 2018-2025.

Tsimberidou, A.M., Fountzilas, E., Nikanjam, M., and Kurzrock, R. (2020). Review of precision cancer medicine: Evolution of the treatment paradigm. Cancer Treat Rev 86, 102019.

Workman, P., Draetta, G.F., Schellens, J.H.M., and Bernards, R. (2017). How Much Longer Will We Put Up With $100,000 Cancer Drugs? Cell 168, 579-583.