Finding Promising Genes in Drug Development

Oliver Spohngellert

Introduction

It is common in cancer drug development to analyze the expression levels of genes within a cancer cell line before and after drug administration. This type of analysis can be used to explain the mechanisms by which drugs work and can be used to make sure the drug does not have unwanted side effects. One example of the use of this type of analysis is in the DeSigN tool. DeSigN is a web tool that predicts a drug's efficacy in treating cancer by comparing the expression profile from the drug to a database of known drugs.

Due to technological advances high throughput screenings of perturbations are becoming a much more common way of identifying drug candidates for a variety of conditions, including cancer. However, when using the data from screenings, it is possible researchers will go down dead ends, as some genes change expression from many different perturbations. These genes should not be prioritized, as the change in their expression level does not likely signal the efficacy of treatment. However, genes that almost never change expression level should be prioritized. Therefore, the goal of this project is to establish a methodology for identifying genes that are most promising to investigate in drug development.

Methods

The data for this project was from a high throughput screen done by the Broad Institute called L1000. The dataset contains measurements for 978 landmark genes, which are then used to impute the expression of 11,350 additional genes at high accuracy. Only the landmark genes were used for the purpose of this project. The expression of these genes was measured under 19,811 small molecule perturbations and 5,075 genetic perturbations over a period of years. These perturbations were given to many different cancer cell lines including skin, lung, breast, prostate, and colon cancers. Only the four cell lines MCF7, HCC515, A375, and HT29 were used in this study.

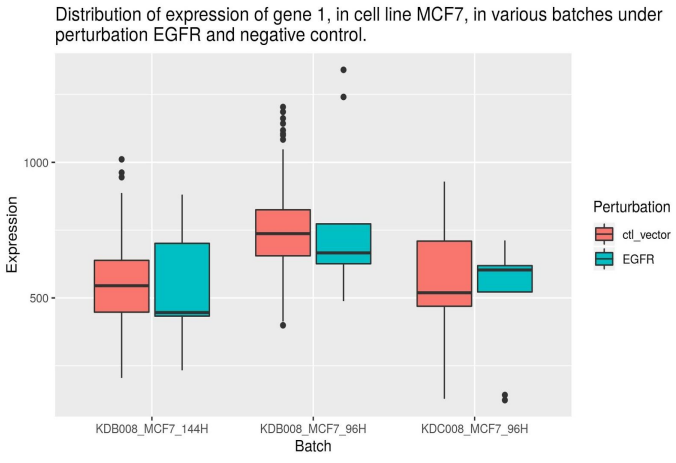

Figure 1

The experiment to generate the dataset was run by measuring expression profiles on 384 well plates in triplicate. This creates a batch effect, as differences in climate, the person conducting the experiment, and the measuring device can cause changes to the measurement. Thus, any study of the data needs to account for batch as a confounding variable. Each plate has control cells that were either unperturbed or given a perturbation that is known to have no effect on gene expression in humans. These can be used as a baseline to compare to the

expression profile under perturbation. In this study, only the negative controls were used. The reason for this is that the vehicle control, being the liquid DMSO which is used to administer perturbations, generally had a much lower distribution than cells under any perturbation. It was therefore decided that the negative control is most appropriate for this experiment.

In order to confirm the existence of a batch effect, boxplots were generated to visualize the distribution of gene expression in various batches under both a perturbation and a control. As can be seen in Figure 1, there is a clear difference in the distribution of expression across batches, with the middle batch having a higher average expression than the others. It was therefore decided that batch would be included as a factor in any models of expression.

To design a methodology for modeling the data, the distribution of replicates across batches must be understood. Thus we create design tables for multiple perturbations vs the controls. As all genes are measured in all wells on the plate, they do not need to be considered in the design table. There are two common cases in the data: designs where there are at least two batches in common between the control and the perturbation, and designs where there are fewer than two batches in common between the control and the perturbation. Table 1 is one example of the former case. Batches 1 and 2 have measurements for both the perturbation and the control, though the design is clearly unbalanced. Therefore in this scenario, a two factor model should be under consideration. This gives us the advantage of considering batch as a factor, while having the disadvantage of forcing us to not use many of the measurements for the perturbation. Table 2 is an example of the latter case, as there is only one batch in common between the control and the perturbation. In this case we cannot consider a 2 factor design, so we have to ignore batch as a confounding variable. This gives us the advantage of having more data to work with, as measurements can be taken from all present batches, though it leaves us with a model that is not true to the experimental design. Thus, based on the design tables we will be using two

Batch Number	Perturbation Replicates	Control Replicates
B1	**6**	**64**
B2	**9**	**96**
B3	24	-
B4	-	93
B5	-	93
...	...	...
B43	-	79

Table 1: Design table for a 2 factor model

Batch Number	Perturbation Replicates	Control Replicates
B1	**9**	**24**
B2	24	-
B3	-	96
B4	-	93
B5	-	93
...	...	...
B43	-	79

Table 2: Design table for a 1 factor design

different models depending on the perturbation.

Based on the analysis of the data above, two models were chosen based upon the design table for a given perturbation. For the two factor model, the following is how the data was modeled:

$$Y_{ijk} = \mu_. + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \epsilon \overset{iid}{\sim} N(0, \sigma^2), \alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$$

In this model α is the effect of the perturbation on the model, where i = 1 for perturbed and 2 for control. β is the batch effect, with j = 1, ... n_batches, with n_batches depending on the perturbation. Finally, $(\alpha\beta)$ is the interaction term. When using batch is not possible, the following model is chosen:

$$Y_{ij} = \mu_. + \tau_i + \epsilon ij, \epsilon \overset{iid}{\sim} N(0, \sigma^2), \tau_1 = 0$$

In this model τ is the effect of perturbation vs control, with i again being either 1 or 2. In both of these models the assumption is made of constant variance. It is common in biology to see variance increase as expression levels increase, so this assumption may not be valid. However, this was a simplifying assumption chosen to be made for the purposes of this project, and should be investigated further in future work.

Finally, for both of these models, the following testing framework is chosen:

$$H_0 : \mu_{1.} - \mu_{2.} = 0, H_a : \mu_{1.} - \mu_{2.} \neq 0$$

This is a simple comparison of factor level means implemented as a contrast function. However, as roughly $978 \cdot 4000 = 3912000$ tests are being done per cell line, using a bonferroni correction for multiple testing would produce very few, if any, results. Thus, instead of controlling for the Family Wise Error Rate (FWER), it is decided that the False Discovery Rate (FDR) will be controlled. FDR is the expected proportion of null hypotheses that are rejected that should not have been. FDR is controlled by the Benjamin-Hochberg procedure. In the Benjamin-Hochberg procedure, all p-values are ordered, and then you find the smallest k which satisfies the following equation:

$$P_{(k)} \leq \frac{k}{m}\alpha, \; m = total\ number\ of\ tests, \; \alpha = FDR$$

For the purposes of this project, α was set to be 0.05, though more work could have been done to find an optimal FDR. This procedure was applied per cell line, so each cell line has a different corrective value. After correction, the null hypothesis was tested at 95% confidence. Finally, the number of tests passed per gene were summed on a per cell line basis. The genes that had the most rejected null hypotheses would be considered the most overactive, while the genes that rejected the fewest would be considered the most inactive. The genes could be ranked by the number of tests passed in either model.

Results

The results for the interaction model were a mixed bag. As can be seen in Table 3, in the cell line MCF7, it was clear to see that some genes had significantly different expressions in

many perturbations versus controls, and thus were more overactive. However, there were many genes that had had no perturbations impact their expression. Further, in the other cell lines, very few tests were run, so the results for these cell lines cannot be seen as significant.

Cell Line	**MCF7 - Breast Cancer**	**HCC 515 - Lung Cancer**	**A375 - Skin Cancer**	**HT29 - Colon Cancer**
Top 5 genes by index and proportion of tests passed	**188**, 0.363 231, 0.292 76, 0.274 83, 0.262 319, 0.228	**188**, 0.020 458, 0.020 414, 0.017 127, 0.014 317, 0.014	Irrelevant (fewer than 10 perturbations had no replicate)	Irrelevant (fewer than 10 perturbations had no replicate)
Bottom 5 genes by index and proportion of tests passed	Many genes (> 50) had 0 tests pass.	Many genes (> 50) had 0 tests pass.	Irrelevant (fewer than 10 perturbations had no replicate)	Irrelevant (fewer than 10 perturbations had no replicate)

Table 3: Most and least active genes based on tests using the interaction model

As can be seen in Table 4, the model that did not include batches produced many results, and seems to be promising in terms of ranking genes by activity. No genes within the top 5 in terms of activity in any cell line appear in the top 5 of another cell line, as is the case with the bottom 5. Further, in the cell lines MCF7 and A375, some genes seem to be extremely active, rejecting the null hypothesis over 75% of the time. Nevertheless, this model does not account for the confounding variable of batch, so the results should be taken with more scrutiny. It is possible that using a higher confidence would be better in this case, but as of now, these are the results. Therefore, it is clear that there are genes in each cell line that are overactive, and genes that are almost never active. Further, it is clear which genes are most active and to what extent differs significantly by cell line.

Cell Line	**MCF7 - Breast Cancer**	**HCC 515 - Lung Cancer**	**A375 - Skin Cancer**	**HT29 - Colon Cancer**
Top 5 genes by index and proportion of tests passed	617, 0.766 666, 0.695 208, 0.683 663, 0.656 430, 0.650	381, 0.419 157, 0.411 201, 0.393 401, 0.389 608, 0.385	499, 0.766 272, 0.762 553, 0.756 110, 0.755 26, 0.738	816, 0.291 622, 0.288 307, 0.278 421, 0.272 965, 0.266
Bottom 5 genes by index and	458, 0.000 568, 0.000	945, 0.008 505, 0.009	483, 0.004 579, 0.008	Many genes (> 50) had 0 tests

proportion of tests passed	592, 0.000 619, 0.000 630, 0.000	920, 0.014 933, 0.015 779, 0.021	674, 0.009 736, 0.009 340, 0.014	pass.

Table 4: Most and least active genes based on tests using the model without batch

Finally, it was important to understand the standard error associated with the tests done on each model. As can be seen in Figure 2, the interaction model produced lower standard errors than the non-interaction based model. This is likely due to lower variance of the control group due to the use of batches. Based on this result, it would be advantageous to do more tests based on an interaction model. In order to do this, more work would have to be done understanding the original dataset's experimental design.

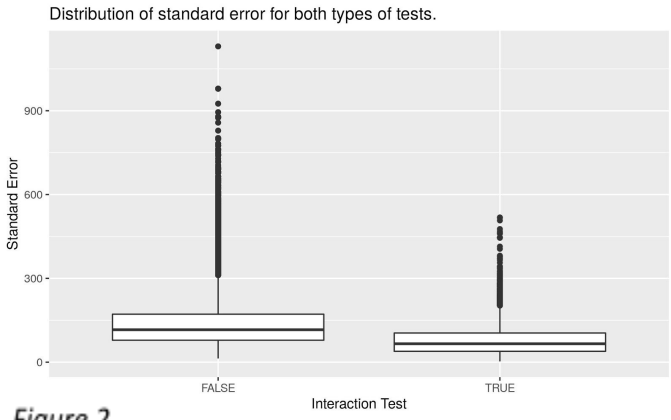

Figure 2

Discussion

This analysis has shown that there are clearly genes that are reactive to arbitrary perturbations, and also genes that are reactive to almost no perturbations. Further, it has developed a methodology for determining the statistically significant change in expression of a gene in a perturbed cell line versus a control cell line. It is clear that including batch as a variable in the two factor model leads to more accurate results, and should be used whenever possible.

There are many areas that can be explored in the future with regards to this data. The primary area of exploration is the use of plate as a block instead of batch. This would allow for more uses of an interaction based model, and likely would produce more accurate results. Another key area of exploration is the assumption of constant variance. As stated previously, it is common in biological settings to see non constant variance, especially having variance increase with Y. This should be explored, and measures should be taken if this is the case. Finally, it is possible that some changes in genetic expression that are statistically significant have no biological significance. This should be accounted for, but would require some background in biology to determine.

References

Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-1452.e17.

Chengalvala MV, Chennathukuzhi VM, Johnston DS, Stevis PE, Kopf GS. Gene expression profiling and its practice in drug development. Curr Genomics. 2007;8(4):262-70.

Lee BK, Tiong KH, Chang JK, et al. DeSigN: connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics. 2017;18(Suppl 1):934.

Benjamini, Yoav & Hochberg, Yosef. (1995). Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. J. Royal Statist. Soc., Series B. 57. 289 - 300. 10.2307/2346101.