Insert your team name here

Team members names

Replicathon 2017

Instructions:

In this R Markdown document, you and your team will create a fully reproducible analysis with the goal of assessing and interpreting the replicability of two pharmacogenomic experiments. This document should contain all of the text and code of your analyses, which will allow others to run, interpret, and reuse your work.

The questions below will help guide you in your analyses and interpretation of results. You don't need to answer every question, but for the problems you do complete, make sure that you completely justify your conclusions by explaining your reasoning and including numerical summaries and data visualizations wherever possible. There are four tutorials (also R Markdown documents) that will help you learn new tools to tackle these problems, and the questions are divided into four sections corresponding to the tutorials (though many overlap with more than one tutorial). If questions arise during your analyses that do not fit into these problems, feel free to include those as well.

For each answer, include text by simply typing below the question. Include code in code blocks (include three back ticks at the start and end of each code block):

```
#Your code goes here
```

You may find it helpful to use the version control and code sharing system called GitHub to work together with your team so that all of you can edit the same document and keep track of its changes. Here is a setup guide and brief introduction to Git and GitHub from another course. The mentors will be able to help if you run into problems.

Questions:

Exploratory analysis of pharmacogenomic data

1. How many cell-lines are contained in the data?
2. What drug concentrations were used in each study?
3. Histograms, appart from telling how is the data distributed, can also make evident potential problems with the data. Plot a histogram of drug viabilities. Does it look as one would expect from the description of the data?
4. How many viability data points are within the expected range according to the definition of viability (e.g. above 0 and below 100)? Try to come up with explanations about the values that are out of range. Are these due to technical issues? Biology?
5. Read the csv file containing the summarized data files. What kind of variables are in the data? What does each column represents?
6. Plot a histogram of the viability scores separating the different drug doses. Are stronger drug concentrations consistent with lower viability scores?

Using Correlation Measures to Assess Replicability of Drug Response Studies

1. Create a scatterplot for each drug comparing the AUC in GDSC and CCLE for all cell lines (hint: code from Tutorial 2 may help).
2. Calculate correlation coefficients of the AUC in GDSC and CCLE for each drug (hint: code from Tutorial 2 may help).
3. Based on the scatterplot and correlation values, would you say that they tend to agree? Why or why not?
4. Does the AUC or IC50 suggest more agreement between the studies?
5. Which drug do you think shows the most consistency between the studies? How about the least?
6. If you calculated more than one type of correlation coefficient (for example Pearson and Spearman), how do they differ? Which do you think is a better summary of the consistency between the two studies?
7. We have explored Pearson and Spearman correlation, but what about other correlation measures? For example, you could try out distance correlation, which is sensitive to nonlinear relationships. You can find this measure in the `energy` R package, which you'll need to install and load with the following commands:

   ```
   install.packages("energy")
   load(energy)
   ```

   Then, you can compute distance correlation with the `dcor()` function. How does this correlation measure compare to Pearson and Spearman? Do your conclusions about the agreement between the two studies change at all using this measure?

Identifying biological factors that influence replicability of pharmacogenomic studies

1. Are there any cell lines that seem to be consistently sensitive? (Hint: look for the 5 cell lines which seem the most resistant in both studies according to the average drug response by AUC; code from Tutorial 3 may help) What effect might this have on your conclusions and interpretations of the agreement between the studies? For example, do you observe any increase in replicability (as measured by correlation of drug response) if you exclude the most resistant cell lines?
2. Classify each cell line as resistant or sensitive to a drug based on its AUC value (Hint: choose a cutoff for which cell lines with AUC below the cutoff are considered sensitive and explain your choice of cutoff(s); code from Tutorial 3 may help). Compare the resistance status of cell lines in GDSC and CCLE for each drug using either a scatterplot of AUC values colored by resistance/sensitivity in each study or a table.
3. Compute the Matthews correlation coefficient for sensitivity status (from #2) of cell lines in GDSC and CCLE for each drug (Hint: code from Tutorial 3 may help).
4. Are there any drugs for which most or all cell lines seem to be resistant in both studies (Hint: for each cell line, look at its average response across all drugs; code from Tutorial 3 may help)? If so, what are the correlation values for these drugs? What do these results imply about the replicability of these studies?
5. Compare the Matthews correlation coefficient values by drug classes defined in Tutorial 3 (No effect, Narrow effect, Broad effect). Which drug class shows the most agreement between the studies?
6. Would you say that the sensitivity results for the two studies tend to agree?
7. For one of the targeted drugs, examine the cell lines that were sensitive ih the CCLE and/or GDSC. See if you can find out what types of cells these are by searching the online Cancer Cell Line Encyclopedia http://www.broadinstitute.org/ccle (this will prompt you to register with a username, password,

and email address. If you prefer, you can also search the cell line on other websites). See if you can find out what types of cancers this drug is targeted for using the NCI cancer drug database at https://www.cancer.gov/about-cancer/treatment/drugs. Does the list of cell lines found to be sensitive in the two studies agree with this?

Modeling the relation between two variables (drug concentration vs viability)

1. Explore the response curves for several drug-cell line combinations. How many drugs-cell line combinations contain viability response values that would potentially enable them to be summarized into an IC50 value? You can answer this, for example, by checking whether there are viability values below 50%.
2. Analyze the re-calculations of IC50 and AUCs from the drug 17-AAG in the H4 cell-line and the drug Nilotinib cell-line in the 22RV1 cell-line. See the figure below and asnwer: which statistic is more roboust, IC50 or AUC? Which statistic is more generalizable, IC50 or AUC? Justify your answer with examples and/or using the whole data recalculations from the *mySummarizedData* variable.

Modified from Kirstie Whitaker.

3. Are the results more replicable if one uses the same code to calculate IC50 or AUC on the different datasets? Hint: you can use code from tutorial #3 to evaluate this.
4. Summarize the viability curves of all the drugs in each cell-line using the slope of linear model. Is the slope of the linear regression informative of the drug response? Can we improve the replicability of the studies by using the value of the slope instead of IC50s or AUCs?

Discussion:

Summarize the main findings of your analyses in the previous four sections here.