Group 4

Reunion 1: **October25, 2021**

Recorder: Nelson Rodriguez

Reflector or Strategy Analyst: Eliam Ruiz Agosto

Manager or Facilitator: Arvin Torres

Speaker or Presenter: Luis Ortiz

Professor's Email(pattiordonez@gmail.com)

Links:
1. areyesq89/PR2017replicaton (github.com)
2. Pharmacogenomics: DNA, drugs and dosage - YouTube
3. Use native R on Google Colab for Data Science - YouTube
4.https://drive.google.com/file/d/1rfD3yMQ6zg9XNN2mp8Mf_hzyMVAGyzUB/view?usp=sharing

Adds:
- %%R
-https://github.com/areyesq89/PR2017replicaton
-https://www.youtube.com/watch?v=huAWa0bqxtA
-https://sci-hub.se/

Notes:

Reunión #1 (Recorder: Nelson Rodríguez) - We began by discussing and volunteering for the roles we wanted, afterwards we allocated the remaining roles in a fair manner. We started to read over the articles provided for us in the git repository. Afterwards, we discussed on how to go about doing the project; after skimming over the parts of the project we (for the moment) have decided to assign a "lead" for each part and decided that they will be in charge of writing/doing their assigned part while receiving support from the other teammates, and so on so forth for each part.

Reunion 2

Notes:

(Recorder: Nelson Rodríguez) - **3 de diciembre de 2021**. For this meeting, we'd all worked on our respective parts separately after our first meeting. The purpose of this meeting was to discuss the date to record our presentation and to have a general overview of the code so far and what everyone has worked on. Additionally, we provided assistance to the members that needed it and we discussed the codes to make sure we agreed that it fit the question. We decided to record the very next day and hand it in the very same.

Reunion 3

Notes:

(Recorder: Nelson Rodriguez) - **4 de diciembre de 2021.** This meeting was solely for working out some last minute fixes on the code that we noticed needed fixing, and we gave it a practice run to decide how we would present in under 5 minutes. After some technical difficulties were quickly solved, we got on to record the presentation and it went smoothly; and so, this concludes our teamwork on this project.

Analysis template completion:

1. Exploratory analysis of pharmacogenomic data: ELIAM S RUIZ-AGOSTO

1.There are 288 unique cell lines in the data and there are 43427 cell lines even counting the ones that are repeated. (code explained in video)
2. We were unable to obtain these values through R. These values were obtained using the UNIQUE function provided by excel. We selected all the data in the column 'concentrations' for each of the rows that were part of each study and applied the unique function.

GDSC [8, 0.007813, 0.015625, 0.03125, 0.0625, 0.125, 0.25, 0.5, 1, 2, 0.003906, 4, 0.000977, 0.001953, 0.0004, 0.0008, 0.0016, 0.0032, 0.0064, 0.0128, 0.0256, 0.0512, 0.1024, 0.039063, 0.078125, 0.15625, 0.3125, 0.625, 1.25, 2.5, 5, 10]

CCLE [0.0025, 0.008, 0.025, 0.08, 0.25, 0.8, 2.53,8]

3. The Histogram values are skewed to the left and there are values that don't seem to be properly collected since there are viability values that go higher than 100 which goes against its description/definition. (The histogram was generated using R and it is explained in the video)

4. In the data there are 15768 values of viability that go over 100 and below 0 which are incorrect data and 27649 that are in the expected range which means that they are appropriate data. (Obtained through R, code explained in video)

5. The following are the names and descriptions of each variable in the data that was contained in the `summarizedPharmacoData.csv` file. (Represented with R in the collab as a data frame table)

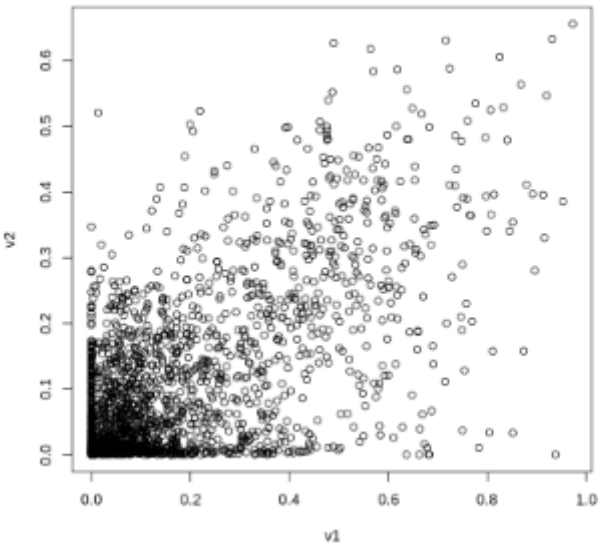
The IC50 a variable that function as a metric for the effectivity of each drug in each study
The AUC is a variable that represents the area under the curve (definite integral).
The cellLine variable is the variable that represents the data in the current cell line.
The drug variable is the variable that stores the drug used.

6. In this question each drug dose ID represents a specific drug concentration. In order to see if the concentration had a relation to the drugs viability we plotted 9 histograms that correspond to the nine doses ID's in the CCLE study. With observation we can see that the stronger drug concentrations do result in lower viability scores. If we observe the concentration values which are sorted in ascending order (in the list below) and we look at the corresponding histograms (in google collab) we can see that the lower concentration drug ID has higher viability scores.

Dose 1	**0.0025**
Dose 2	**0.008**
Dose 3	**0.025**
Dose 4	**0.08**
Dose 5	**0.25**
Dose 6	**0.8**
Dose 7	**2.53**
Dose 8	**8**
Dose 9	**2**

2. Using Correlation Measures to Assess Replicability of Drug Response Studies: Luis Ortiz

Question 1



-

Question 2:

- The correlation between the two datas is 0.6672839.

Question 3:

- I would say yes that they tend to agree because if you look at both the correlation and the scatterplot the information is all grouped in one part instead of separated.

Question 4:

- If one looks at both the IC50 and the AUC data they are not the same because the correlation and the scatterplot are very different in both studies and it looks like there is no agreement between both studies..

Question 5:

- If you could look at the studies you could say that AUC has more consistency than does that in IC50 which could be the least.

Question 6:

- Between the two studies one could see that the correlation of both are closer in the spearman method than in the pearson method.
AUC(pearson = 0.6672839, spearman = 0.5402709), IC50(pearson = 0.3088802, spearman = 0.5554557)

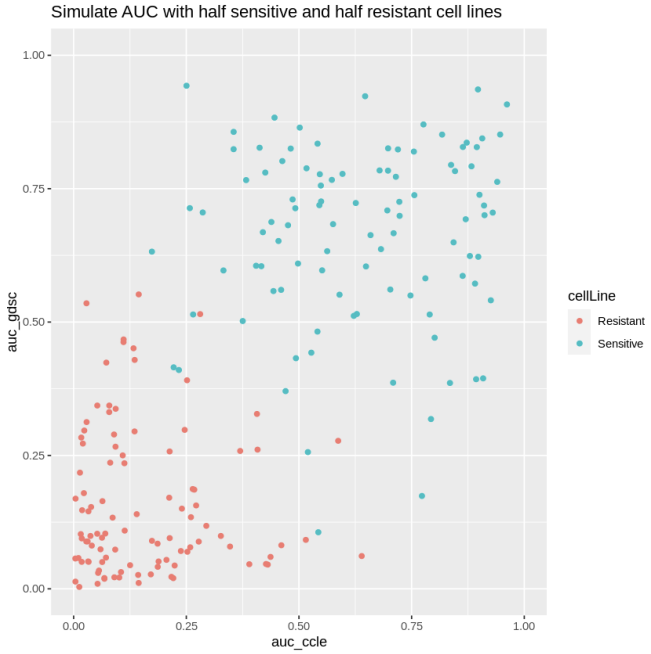
Question 7: There is a problem running the code on r because of the load package to get the results.

3. Identifying biological factors that influence replicability of pharmacogenomic studies: Nelson Rodriguez

Question 1:

- Por lo que he visto, varias celdas presentan gran sensitividad y discrepancia entre resultados. Esto puedes ser un problema ya que si los resultados no son muy consistentes pues es difícil que sea replicable. Siendo este el caso, no es posible indicar que los estudios esten de acuerdo uno con el otro.

Question 2:



- Los resultados resistentes aparentan ser más precisos que los de los sensitivos.

Question 3:

Question 4:

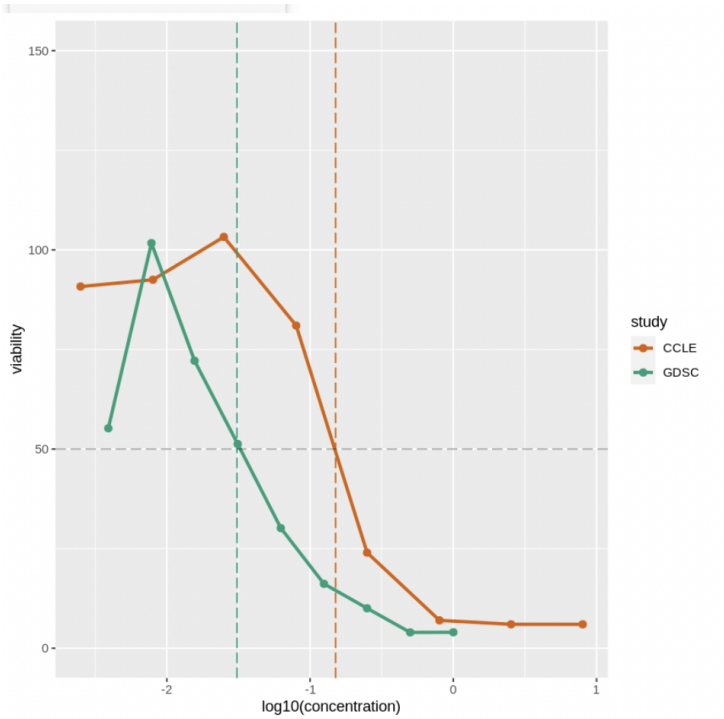
Question 5:

Question 6:

Question 7:

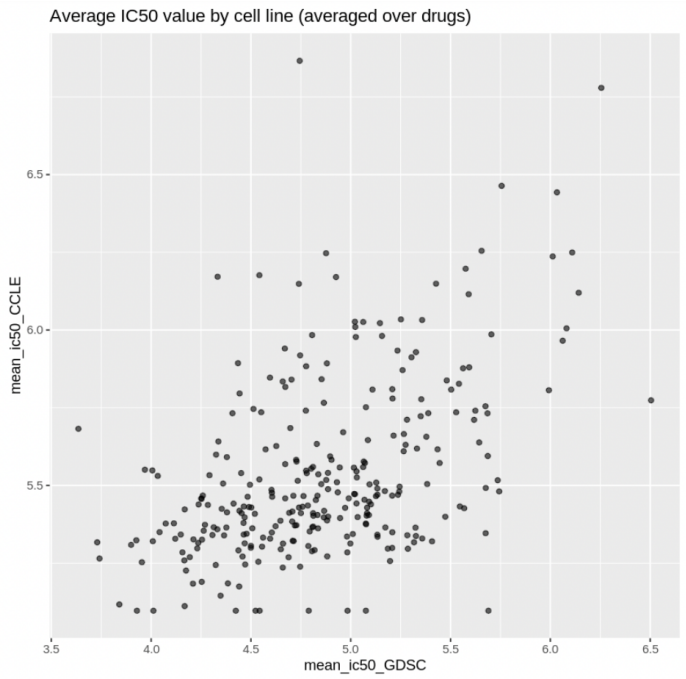
4. Modeling the relation between two variables (drug concentration vs viability) : Arvin Torres Melendez

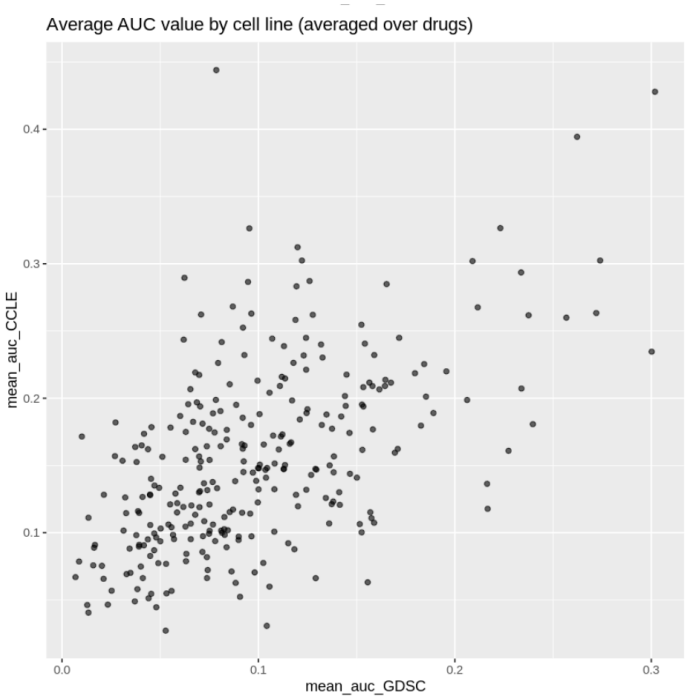
1. **The drug cell combinations that contain viability response values that would potentially enable them to be summarized into an IC50 value (viability values below 50%) are 273. The code has an error on it and it displays 15 instead of the 273.**
2.



By using the plot of viability of these combinations of their drug connection that is already used in tutorial #4 we can see that AUC is more robust than IC50. And that IC50 is more generalisable than AUC.

3.

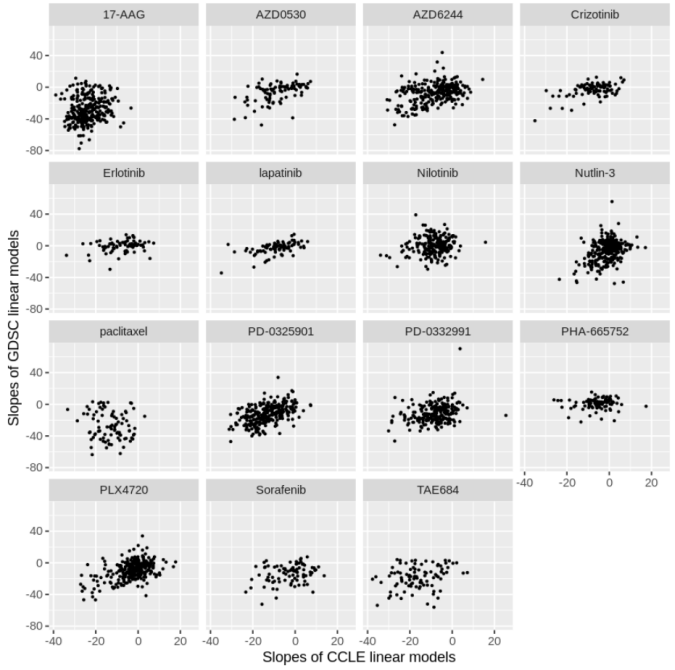




Here we have two scatterplots, one for the means of the IC50 values and another for the means of the AUC values. We can see that the values of AUC are slighty more consistent and that the IC50 values have more outliers. Because of this we can say that the AUC values are more reproducible and that the IC50 values are more replicable.

4.

concentration: -1.43047396498322
concentration: -1.43047396498322
concentration: -25.9304453383784



We summarize the viability curves by calculating the slopes of the models for each cell- line and drug combination for each study and put them in a scatterplot.

From this scatterplot we can say that there is some association between the viability curve slopes, so we are able to gain some information about the viability using these slopes.