Tutorial 2a: Digging Deeper with Cell Line and Drug Subgroups

Introduction

- Large-scale pharmacogenomic studies (like the CCLE and GDSC) measure the effects of *many* anti-cancer drugs on various cell lines.
- The primary interest is not to simply determine which cell lines are most susceptible to which drugs, but rather **what are the characteristics of the cell lines that are susceptible** to certain drugs?
- By matching these characteristics to those of patients' tumor cells, investigators can improve cancer treatment regimens.

If the hope is to improve treatment regimens in cancer patients, why would we study cell lines?

- We have to start in cell lines (looking at the *in vitro* response) when screening many drugs at multiple concentrations, since we need to be able to **compare the effects of drugs and doses on cells of the same type**.
- Cell lines can be grown continously (even in different labs) to generate comparable samples, whereas clinical patient samples are fixed amounts of cells that would be used up after the experiment.

What are the *characteristics* to examine when attempting to relate the best treatment regimen with a cell line or patient?

- Genomic factors (e.g. gene expression, copy number, mutation)
- Pharmacologic (e.g. AUC or IC50 of response to various drugs)
- Whether similar drugs (in terms of mechanism of action) have similar effects

Figure from “Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics” by Andrew Goodspeed et al. (Molecular Cancer Research, 2016).

- The CCLE and GDSC studies both investigated genomic profiles for all the cell lines that were used to study drug response.
- The genomic profiles collected included…
  - gene expression measurements for more than 10,000 genes in every cell line (measured using microarrays)
  - mutation status for a subset of 64 genes in every cell line (measured using targeted sequencing).

We will not have time to explore this data, but it is available through the R package `PharmacoGx` if you wish to explore it for a future project. Some code is also included in the `downloadData.R` script to extract and clean up some of this data (gene expression measurements) to help you get started.

One of the main finding of Haibe-Kains et al. (2013) (the reanalysis paper) was that the genomic data of the cell lines used in both studies were highly correlated.

- However, based on our analysis so far, this *doesn't* seem to be the case with the pharmacological data.
- Part of this may be explained by differences in protocol (e.g. different drug concentrations) between the studies,
- It may also be due to our current approach to measuring "agreement" between the studies.

In the rest of this tutorial, we'll investigate how classifying the cell lines and drugs into subgroups might help explain some of this lack (or not!) of agreement.

Setup Workspace

We start by loading the tidyverse family of packages and setting the default `ggplot2` theme to “`theme_bw`”.

```r
library(tidyverse)
theme_set(theme_bw())
```

Load Summarized Dataset

First, we’ll read in the RDS file that contains the summarized pharmacological data (including the IC50 and AUC values for each drug and cell line combination, as described above).

```r
summarizedData <- readRDS(file.path("..", "data", "summarizedPharmacoData.rds"))
str(summarizedData)
```

```
'data.frame':   2557 obs. of  6 variables:
 $ cellLine : chr  "22RV1" "5637" "639-V" "697" ...
 $ drug     : chr  "Nilotinib" "Nilotinib" "Nilotinib" "Nilotinib" ...
 $ ic50_CCLE: num  8 7.48 8 1.91 8 ...
 $ auc_CCLE : num  0 0.00726 0.07101 0.15734 0 ...
 $ ic50_GDSC: num  155.27 219.93 92.18 3.06 19.63 ...
 $ auc_GDSC : num  0.00394 0.00362 0.00762 0.06927 0.02876 ...
```

Resistance of Cell Lines

- Something that was not considered in Haibe-Kains et al. (2013) was that different cell lines may more or less **resistant** to drug treatment than others.
  - Resistance means that the cell line does not respond to treatment by a drug
- Let’s explore first why that might be an important factor, and then directly compare cell line sensitivity in the two studies.

Consider the following situation: say we have a set of cell lines that are resistant to a particular drug (they don’t respond to it). Then we would expect that no matter what dose we give to the cell lines, their viability results would not change (they would stay near 100%). In this case, the AUC (area above the response curve) value would be near zero since the dose-response curve would look flat:

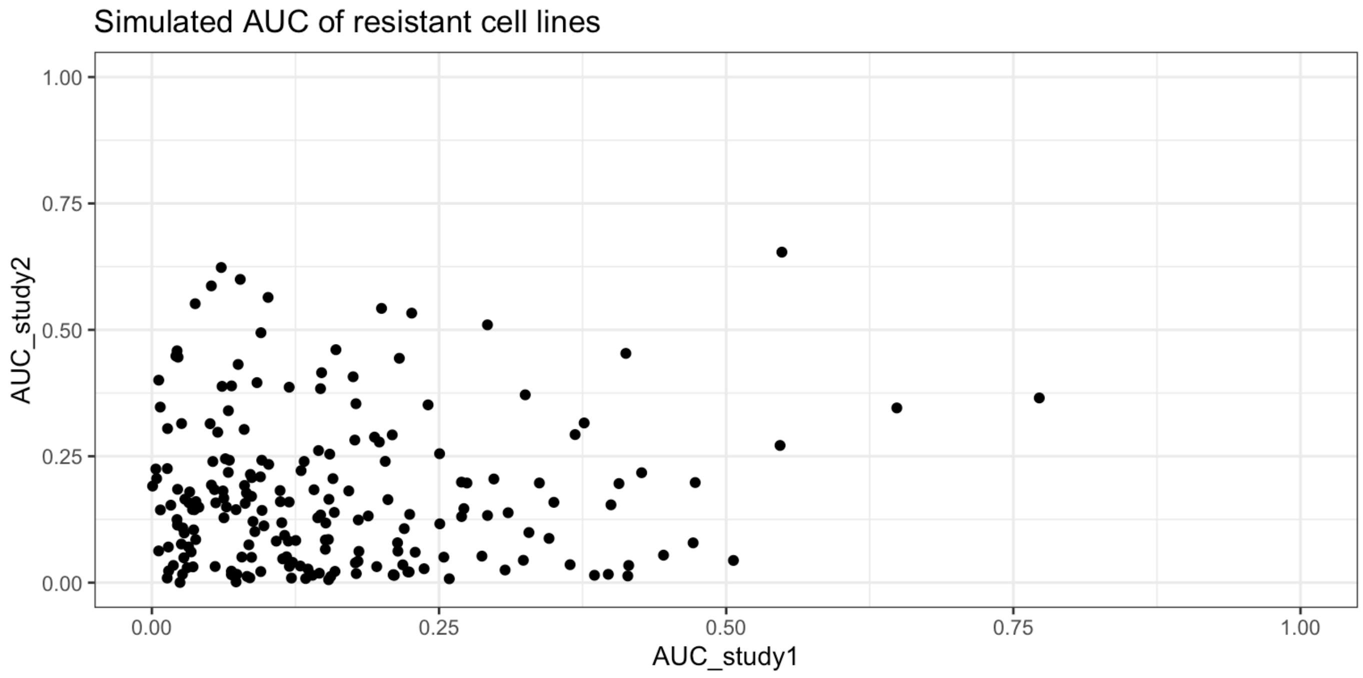
How would you calculate the IC50 value in this case?

> Place your answer here

Assuming the results in both studies were consistent, then a scatterplot of the AUC values in this case would look like this (allowing for experimental error - we don’t expect exactly 100% viability due to variations in experimental conditions and cellular growth rates):

```r
AUC_study1 <- rbeta(200, 1, 5)
AUC_study2 <- rbeta(200, 1, 5)
resistant <- data.frame(AUC_study1, AUC_study2)

ggplot(resistant, aes(y = AUC_study2, x = AUC_study1)) +
    geom_point() +
    xlim(0, 1) +
    ylim(0, 1) +
    ggtitle("Simulated AUC of resistant cell lines")
```



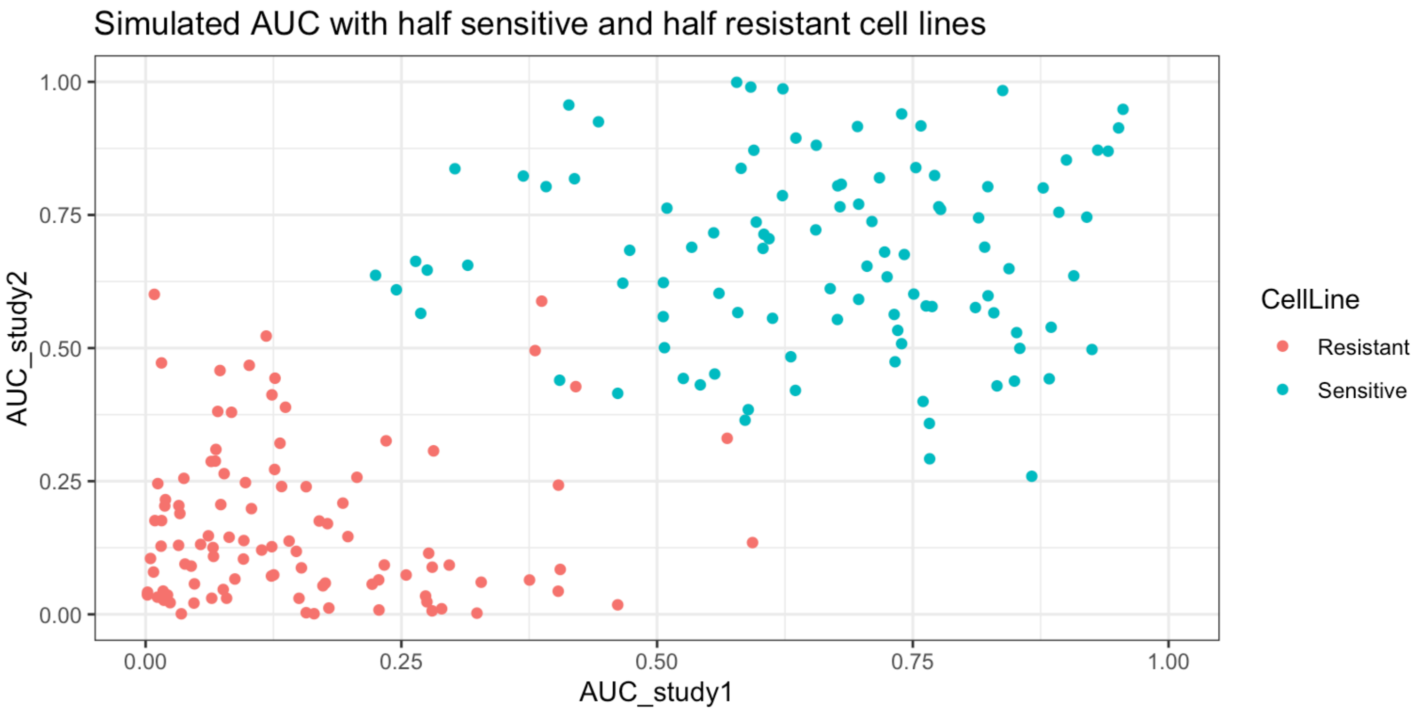
Here we have simulated the AUC values in each study to be closer to zero but with some random variability (this is done here using the Beta distribution). Do these AUC values look correlated? Calculate a correlation coefficient to quantify the correlation.

> Place your answer here

Now, consider the situation where only half of the cell lines are resistant (with AUC values close to zero) and half are sensitive (with high AUC values).

```r
AUC_study1 <- c(rbeta(100, 1, 5), rbeta(100, 4, 2))
AUC_study2 <- c(rbeta(100, 1, 5), rbeta(100, 4, 2))
resistant <- data.frame(AUC_study1, AUC_study2,
                        CellLine = c(rep("Resistant", 100), rep("Sensitive", 100)))

ggplot(resistant, aes(y = AUC_study2, x = AUC_study1, color = CellLine)) +
    geom_point() +
    xlim(0, 1) +
    ylim(0, 1) +
    ggtitle("Simulated AUC with half sensitive and half resistant cell lines")
```



Here we have simulated half the AUC values in each study to be closer to zero and half to be closer to one but with some random variability (using the Beta distribution again). Do these AUC values look correlated? Calculate a correlation coefficient to quantify the correlation.

> Place your answer here

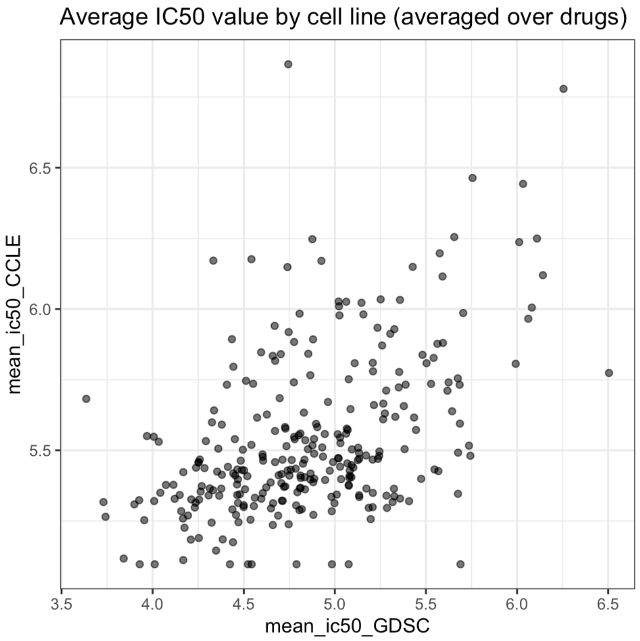
If all cell lines are resistant to a particular drug, would you expect to find a high correlation between the two studies for that drug? Would that result imply that the two studies are not replicable?

> Place your answer here

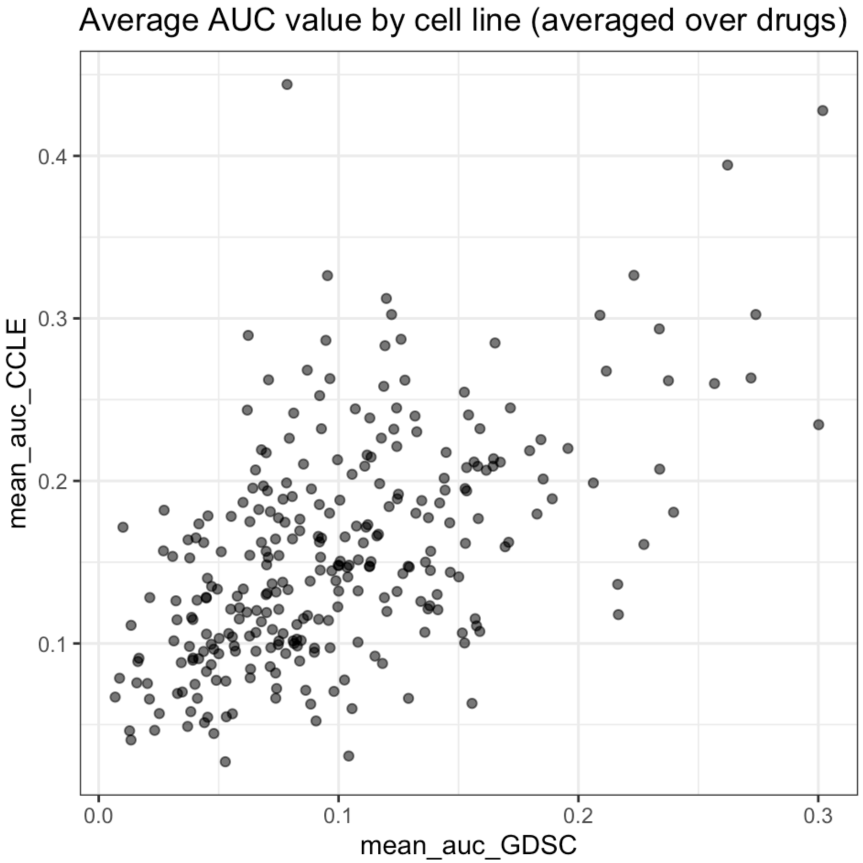
Now, let’s explore the average drug responses by cell line. First, we’ll compute the average drug responses by cell line (averaged over all drugs). Then we’ll create a scatterplot to compare these averages in the two studies.

```r
drugAvg <- summarizedData %>%
              group_by(cellLine) %>%
              summarise(mean_ic50_CCLE = mean(-log10(ic50_CCLE / 10^6)),
                        mean_ic50_GDSC = mean(-log10(ic50_GDSC / 10^6)),
                        mean_auc_CCLE = mean(auc_CCLE),
                        mean_auc_GDSC = mean(auc_GDSC))

ggplot(drugAvg, aes(x = mean_ic50_GDSC, y = mean_ic50_CCLE)) +
    geom_point(alpha = 0.6) +
    ggtitle("Average IC50 value by cell line (averaged over drugs)")
```



```r
ggplot(drugAvg, aes(x = mean_auc_GDSC, y = mean_auc_CCLE)) +
    geom_point(alpha = 0.6) +
    ggtitle("Average AUC value by cell line (averaged over drugs)")
```



What is the most sensitive cell line?

> Place your answer here

Resistance and Replicability

- So far we’ve seen that the sensitivity of cell lines can have major implications for the perceived replicability of studies if we use standard correlation measures for assessment.
- For this reason, some of the followup articles indicated that sensitivity should be taken into account in this kind of analysis.
- How can we consider sensitivity when assessing replicability of the two studies? One idea is to assess the agreement between the studies on which cell lines were sensitive or resistant (instead of agreement on the actual values of AUC or IC50).

To assess agreement between the studies on which cell lines were sensitive or resistant, we first have to define criteria for deciding whether a cell line is resistant or not.

- We’d like to choose an AUC cutoff that divides sensitive and resistant responses. This choice is somewhat arbitrary, we’ll start by choosing the following cutoffs:
- AUC values 0.1 or higher indicate that the cell line is sensitive to the drug
- AUC values below 0.1 indicate that the cell line is resistant
- If no cell lines had observed AUC values above 0.1 in either study, then use 0.4. This is the case for the broadly cytotoxic drug paclitaxel. You are encouraged to try other cutoffs as well.

Let’s go ahead and add sensitivity variables to our dataset, and explore their basic properties.

```r
summarizedData <- summarizedData %>%
              mutate(cutoff = ifelse(drug == "paclitaxel", 0.4, 0.1),
                     sensitivity_GDSC = factor(ifelse( auc_GDSC < cutoff, "Resistant", "Sensitive")),
                     sensitivity_CCLE = factor(ifelse( auc_CCLE < cutoff, "Resistant", "Sensitive")))

table("GDSC" = summarizedData$sensitivity_GDSC,
      "CCLE" = summarizedData$sensitivity_CCLE)
```

```
           CCLE
GDSC         Resistant Sensitive
  Resistant       1289       486
  Sensitive        217       565
```

What proportion of the drug-cell line combinations are in agreement of sensitivity/resistance between the two studies?

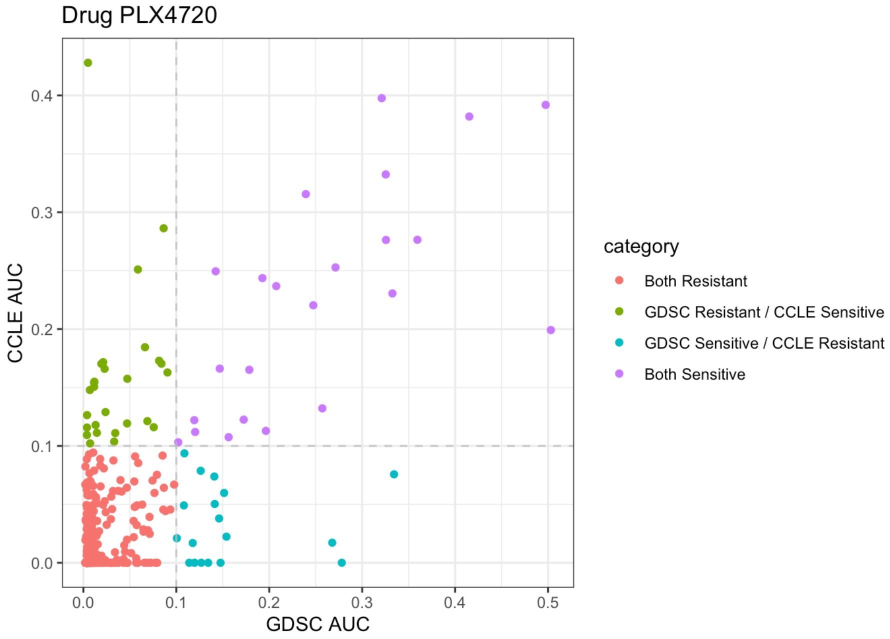
> Place your answer here

Next, we’ll create a scatterplot of the AUC values, colored by sensitivity. We’ll start with just the drug “PLX4720”.

```r
summarizedData <-
    summarizedData %>%
    mutate(category = paste(sensitivity_GDSC, sensitivity_CCLE),
           category = fct_recode(category,
                                 "Both Resistant" = "Resistant Resistant",
                                 "Both Sensitive" = "Sensitive Sensitive",
                                 "GDSC Resistant / CCLE Sensitive" = "Resistant Sensitive",
                                 "GDSC Sensitive / CCLE Resistant" = "Sensitive Resistant"))
table(summarizedData$category)
```

```
                 Both Resistant GDSC Resistant / CCLE Sensitive
                           1289                             486
GDSC Sensitive / CCLE Resistant                  Both Sensitive
                            217                             565
```

```r
summarizedData %>%
    subset(drug == "PLX4720") %>%
    ggplot(aes(x = auc_GDSC, y = auc_CCLE, colour = category)) +
    geom_point() +
    xlab("GDSC AUC") +
    ylab("CCLE AUC") +
    geom_hline(aes(yintercept = cutoff), colour="grey", alpha=0.75, lty=2) +
    geom_vline(aes(xintercept = cutoff), colour="grey", alpha=0.75, lty=2) +
    ggtitle("Drug PLX4720")
```



In this plot, which colors of points correspond to ‘agreement’ between the two studies?

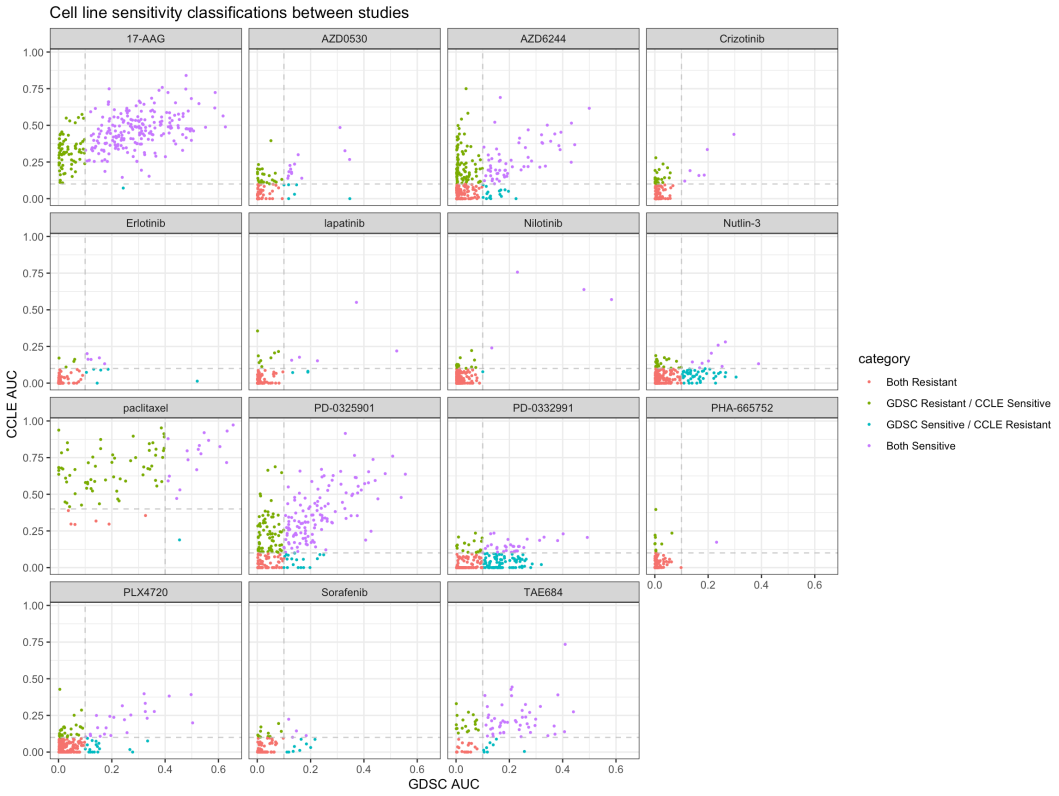
> Place your answer here

Overall, do most cell lines seem to agree or disagree in whether they were sensitive or resistant to PLX4720?

> Place your answer here

Next, we’ll create the same scatterplot above, but for all drugs, so we can compare them side by side.

```r
ggplot(summarizedData, aes(x = auc_GDSC, y = auc_CCLE, colour = category)) +
    geom_point(cex = 0.5) +
    facet_wrap(~ drug) +
    xlab("GDSC AUC") +
    ylab("CCLE AUC") +
    geom_hline(aes(yintercept = cutoff), colour = "grey", alpha = 0.75, lty = 2) +
    geom_vline(aes(xintercept = cutoff), colour = "grey", alpha = 0.75, lty = 2) +
    ggtitle("Cell line sensitivity classifications between studies")
```



Now that we have explored these results visually, is there a way to summarize the agreement for each drug numerically?

- Right away we might think of calculating Pearson or Spearman correlation. However, these measures are intended to measure correlation between either continuous variables or variables with several unique values.
- Here we only have 2 categories: “sensitive” and “resistant”. So we’ll turn to a correlation measure called **Matthews correlation coefficient (MCC)**, which is designed to calculate the agreement based on binary (2-category) classifications. Some details and examples of the MCC statistics are provided in the Supplementary Tutorial, **Supplement: Correlation Measures**.

Since the base R does not include a function for computing the MCC, we will write one now.

```r
mcc <- function (study1, study2) {
    BS <- sum(study1 == "Sensitive" & study2 == "Sensitive")
    BR <- sum(study1 == "Resistant" & study2 == "Resistant")
    SR <- sum(study1 == "Sensitive" & study2 == "Resistant")
    RS <- sum(study1 == "Resistant" & study2 == "Sensitive")

    if (BS+SR == 0 | BS+RS == 0 | BR+SR == 0 |  BR+RS ==0){
        mcc <- ((BS*BR)-(SR*RS))
    }else{
        mcc <- ((BS*BR)-(SR*RS)) / sqrt(exp((log(BS+SR)+log(BS+RS)+log(BR+SR)+log(BR+RS))))
    }
    return(mcc)
}
```

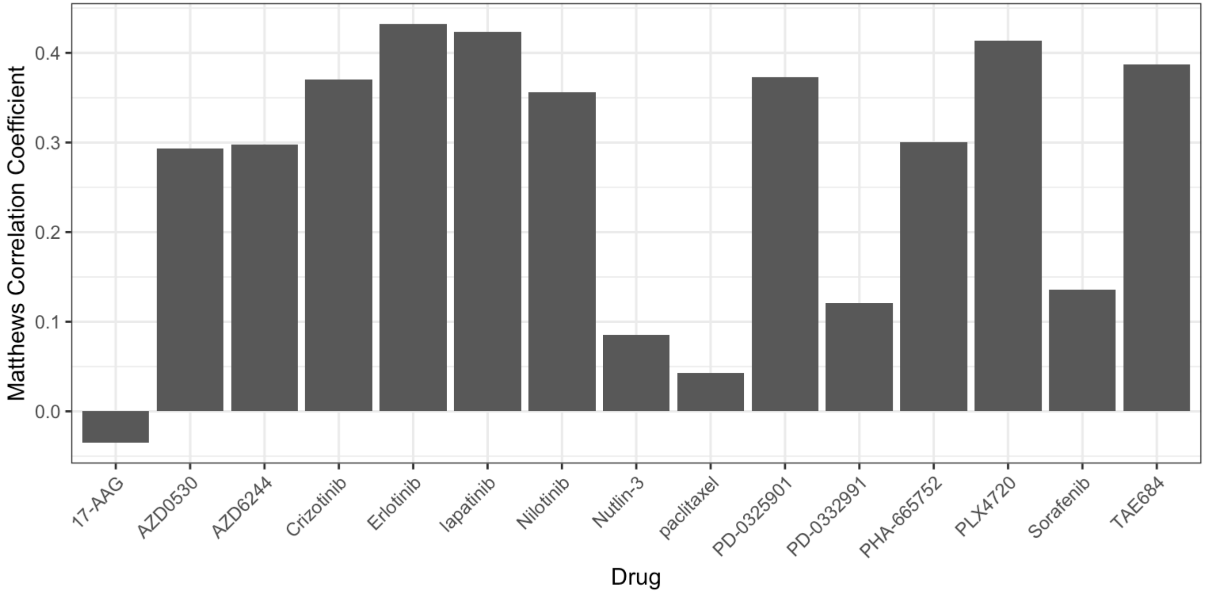
We can now use this function to compute the MCC between studies for each drug separately.

```r
drugCorrs <- summarizedData %>%
    group_by(drug) %>%
    summarise(matthews_corr = mcc(sensitivity_GDSC, sensitivity_CCLE))
drugCorrs
```

```
# A tibble: 15 × 2
   drug       matthews_corr
   <chr>              <dbl>
 1 17-AAG           -0.0350
 2 AZD0530           0.293
 3 AZD6244           0.297
 4 Crizotinib        0.371
 5 Erlotinib         0.432
 6 Nilotinib         0.356
 7 Nutlin-3          0.0856
 8 PD-0325901        0.373
 9 PD-0332991        0.121
10 PHA-665752        0.300
11 PLX4720           0.414
12 Sorafenib         0.136
13 TAE684            0.387
14 lapatinib         0.424
15 paclitaxel        0.0432
```

Let’s now plot the MCC across drugs.

```r
ggplot(drugCorrs, aes(x = drug, y = matthews_corr)) +
    geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    xlab("Drug") +
    ylab("Matthews Correlation Coefficient")
```



Examining the Matthew’s correlation values that take into account cell line categories, which drug do you think shows the most consistency between the studies? How about the least?

> Place your answer here

Does this agree with your visual assessment and answers to these same questions in the previous tutorial (tutorial 2) using continuous correlation measures (Pearson and Spearman)?

> Place your answer here

Targeted Nature of Drugs

In addition to sensitivity of cell lines, another biological factor that might have major implications for the perceived replicability of studies is the **targeted** nature of the drugs used.

- A drug is considered **targeted** if it is expected to have selective activity against some cell lines.
- The mechanism of impact on perceived replicability is very similar to the impact of sensitivity of cell lines: the response in the cell lines that the drug does *not* target may represent random noise (and we don’t expect to observe correlations in random noise).
- For this reason, some of the follow-up articles indicated that drugs should be considered separately based on their targeted nature.

Safikhani et al. describe the following three classes of drugs based on observed responses of the cell lines:

No effect: minimal observed response for all cell lines

- sorafenib
- erlotinib
- PHA-665752

Broad effect: response in a large number of cell lines

- AZD6244
- PD-0325901
- 17-AAG
- paclitaxel

Narrow effect: response in only a small subset of cell lines

- nilotinib
- lapatinib
- nutlin-3
- PLX44720
- crizotinib
- PD-0332991
- AZD0530
- TAE684

Consider, for example, the drugs sorafenib and nilotinib. In the context of these drug classes, which would you expect to have a larger MCC and why? Are the observed MCC values consistent with this?

> Place your answer here