Tutorial 1b: Exploring Replicability with the `summarizedPharmacoData` Dataset

Introduction

In the previous tutorial (**Tutorial 1a**) we explored some of the features of the CCLE and GDSC drug response datasets.

In contrast to the raw data which include the viability at each drug concentration for each cell line, the summarized dataset contains numerical summaries of each cell line's response to each drug over all concentrations.

In this tutorial we'll first learn more about summary measures of drug response, and then use scatterplots and correlation measures to assess the agreement of these summary measures in the two studies.

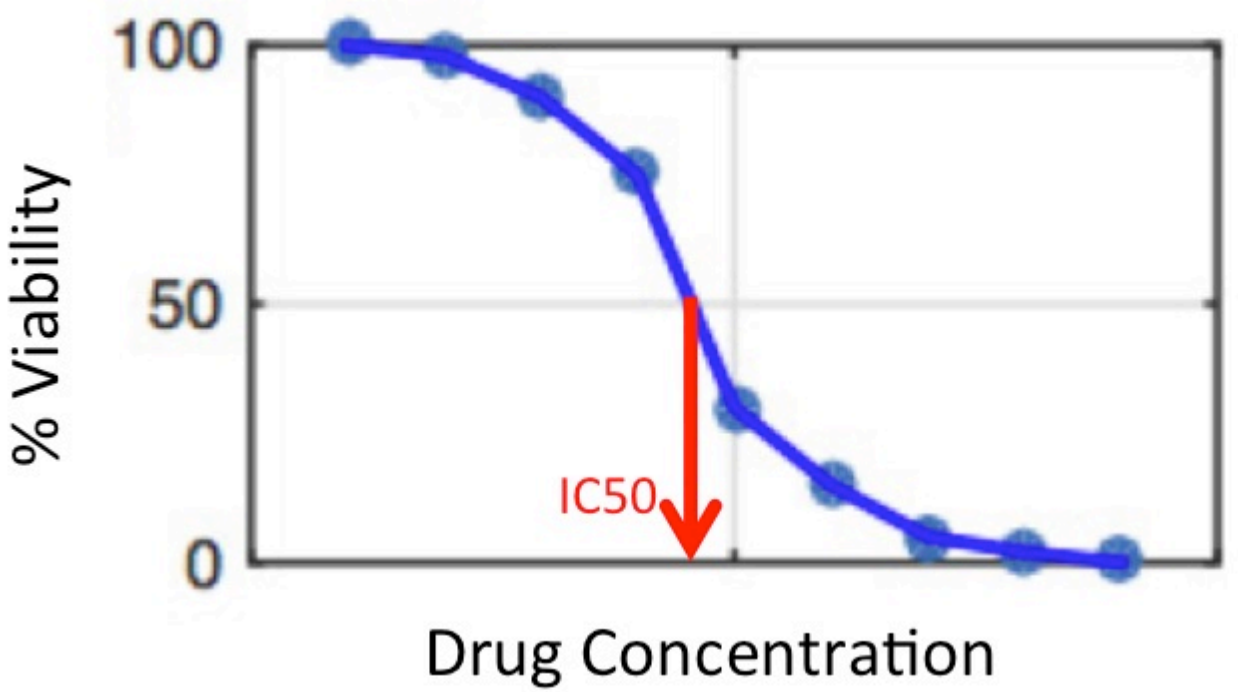
Summary Measures

In the summarized dataset, the cell line viability over all drug concentrations has been summarized into a single number for each cell line and drug combination.

This summary represents the *overall effect of the drug on the cell line*.

There are many different ways this could be done. Our data includes two summary measures that were used in the original studies.

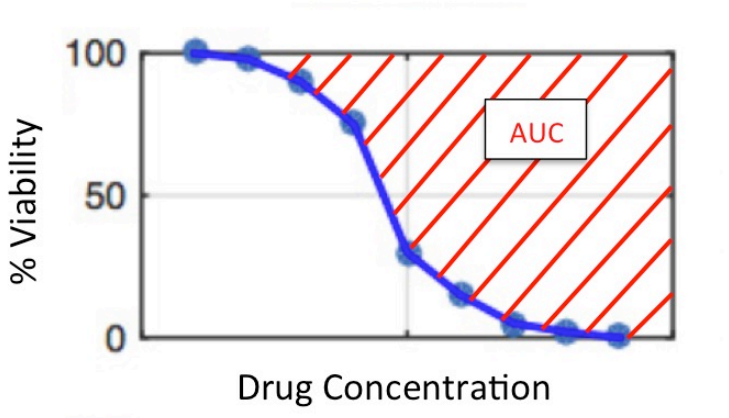
1. **IC50 (Half Maximal Inhibitory Concentration)**: the estimated concentration of the drug that will result in half (50%) of the cells surviving.



Are cell lines with higher IC50 values more or less susceptible? What about drugs with higher IC50 values - are they more or less toxic?

> Place your answer here

2. **AUC (Area Under the Curve)**: despite the name, this is actually the area *above* the curve estimated by the drug concentration and viability data. Note that the estimation of this curve is not a simple task in itself - in the next tutorial we will explore how to summarize the relationship between two variables. (**Tutorial 2b**).



Are cell lines with higher values of AUC more or less resistant?

> Place your answer here

Are drugs with higher AUC more or less toxic?

> Place your answer here

Setup Workspace

We start by loading the tidyverse family of packages and specifying a default plotting theme for our `ggplot` graphics.

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

Load Summarized Dataset

Let's start by loading the RDS file containing the summarized pharmacological data (including the IC50 and AUC values for each drug and cell line combination, as described above).

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

As we did with the raw data, we'll take a quick peek at this data before getting started.

```
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 ...
```

We can count the number of cell lines and drugs in the data.

```r
## with base R
length(unique(summarizedData$cellLine))
```

```
[1] 288
```

```r
length(unique(summarizedData$drug))
```

```
[1] 15
```

```r
## with the tidyverse
summarizedData |>
    summarize(nCellLines = n_distinct(cellLine),
              nDrugs     = n_distinct(drug))
```

```
  nCellLines nDrugs
1        288     15
```

Notice that there are 2557 rows - each row here corresponds to a cell line-drug combination. Making up these combinations are 288 unique cell lines, and 15 drugs.

Was every cell line in the dataset tested with every drug?

> Place your answer here

Comparing Studies using Plots

So we now have summary measures (IC50 and AUC) that indicate the responses of cell lines to drugs. However, each study measured these values separately. The goal of our analysis is to **investigate how well these two studies agree with each other**. In other words, do the drug response results in one study **replicate** in the other study?

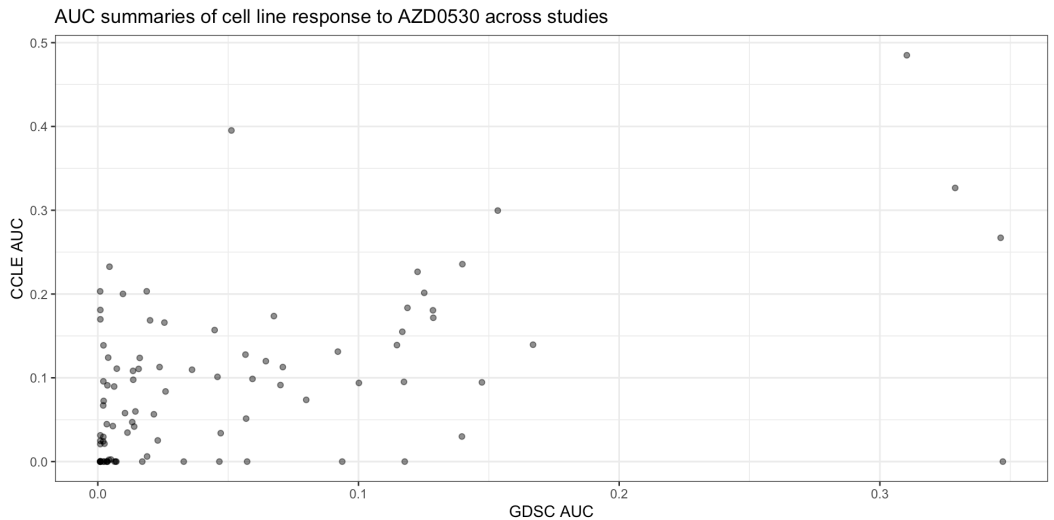
AZD0530

First, we'll examine this question for one of the drugs in particular: **AZD0530**. To make our code easier to read, let's create a separate object from for this subset of the data and call it `azdSummary`.

```
1  azdSummary <- subset(summarizedData, drug == "AZD0530")
```

We'll start out by visually exploring how the AUC values for AZD0530 compare in the two datasets using a scatterplot.

```r
ggplot(azdSummary, aes(x = auc_GDSC, y = auc_CCLE)) +
    geom_point(alpha = 1/2) +
    xlab("GDSC AUC") +
    ylab("CCLE AUC") +
    ggtitle("AUC summaries of cell line response to AZD0530 across studies")
```

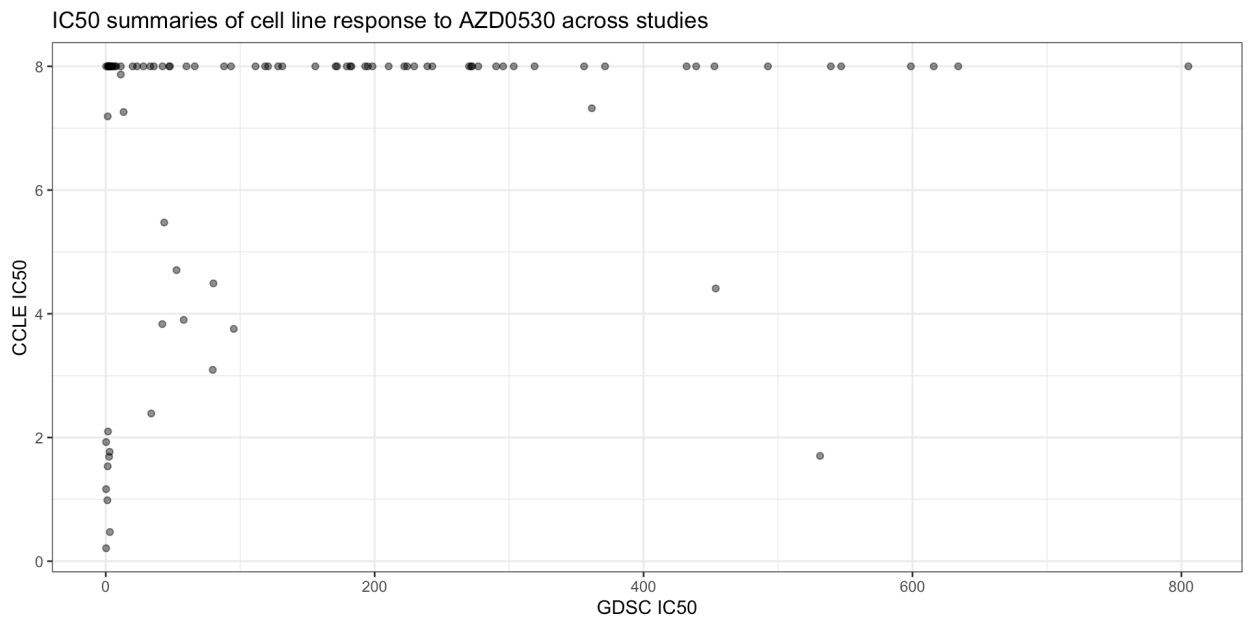


Perfect agreement between the GDSC and CCLE AUC values would mean all the points falling on a straight line. How would you describe the level of agreement between the AUC values for cell line response to AZD0530 in the CCLE and GDSC studies? Does it seem like higher values of AUC in the GDSC study correspond to higher values in the CCLE study?

> Place your answer here

Next, let's look at the same type of scatterplot for the IC50 values.

```
ggplot(azdSummary, aes(x = ic50_GDSC, y = ic50_CCLE)) +
    geom_point(alpha = 1/2) +
    xlab("GDSC IC50") +
    ylab("CCLE IC50") +
    ggtitle("IC50 summaries of cell line response to AZD0530 across studies")
```



What is different about this plot compared to the previous AUC plot?

> Place your answer here

First, you may notice that there are many points with the highest value of IC50 for CCLE. Recall from **Tutorial 1a** that this study measured a fixed set of doses, regardless of how the cells responded, whereas in the GDSC drug concentrations were increased if no response was observed.

What does this mean for the IC50 values in the CCLE? Are they more likely to be too high or too low? Why?

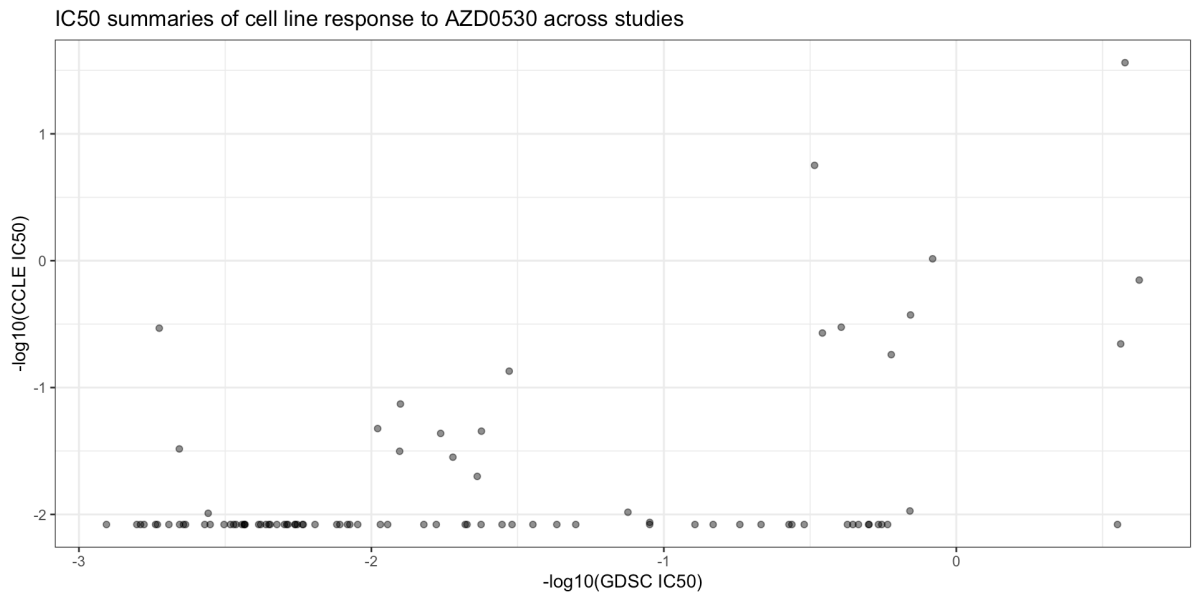
> Place your answer here

Now, let's take a closer look at the axes.

- There is a much larger range of IC50 values reported in the GDSC study, but most values are small.

This indicates a *skewed* distribution that may benefit from a log-transformation. Next, we'll plot log10-transformed values of IC50. We'll also plot *negative* values, so that increasing values of IC50 represent higher levels of drug toxicity (to make the plot more comparable to the AUC plot).

```
ggplot(azdSummary, aes(x = -log10(ic50_GDSC), y = -log(ic50_CCLE))) +
    geom_point(alpha = 1/2) +
    xlab("-log10(GDSC IC50)") +
    ylab("-log10(CCLE IC50)") +
    ggtitle("IC50 summaries of cell line response to AZD0530 across studies")
```



How would you describe the level of agreement between the IC50 values of cell line response to AZD0530 in the CCLE and GDSC studies? Does it seem like higher values of AUC in the GDSC correspond to higher values in the CCLE?

> Place your answer here

All Drugs

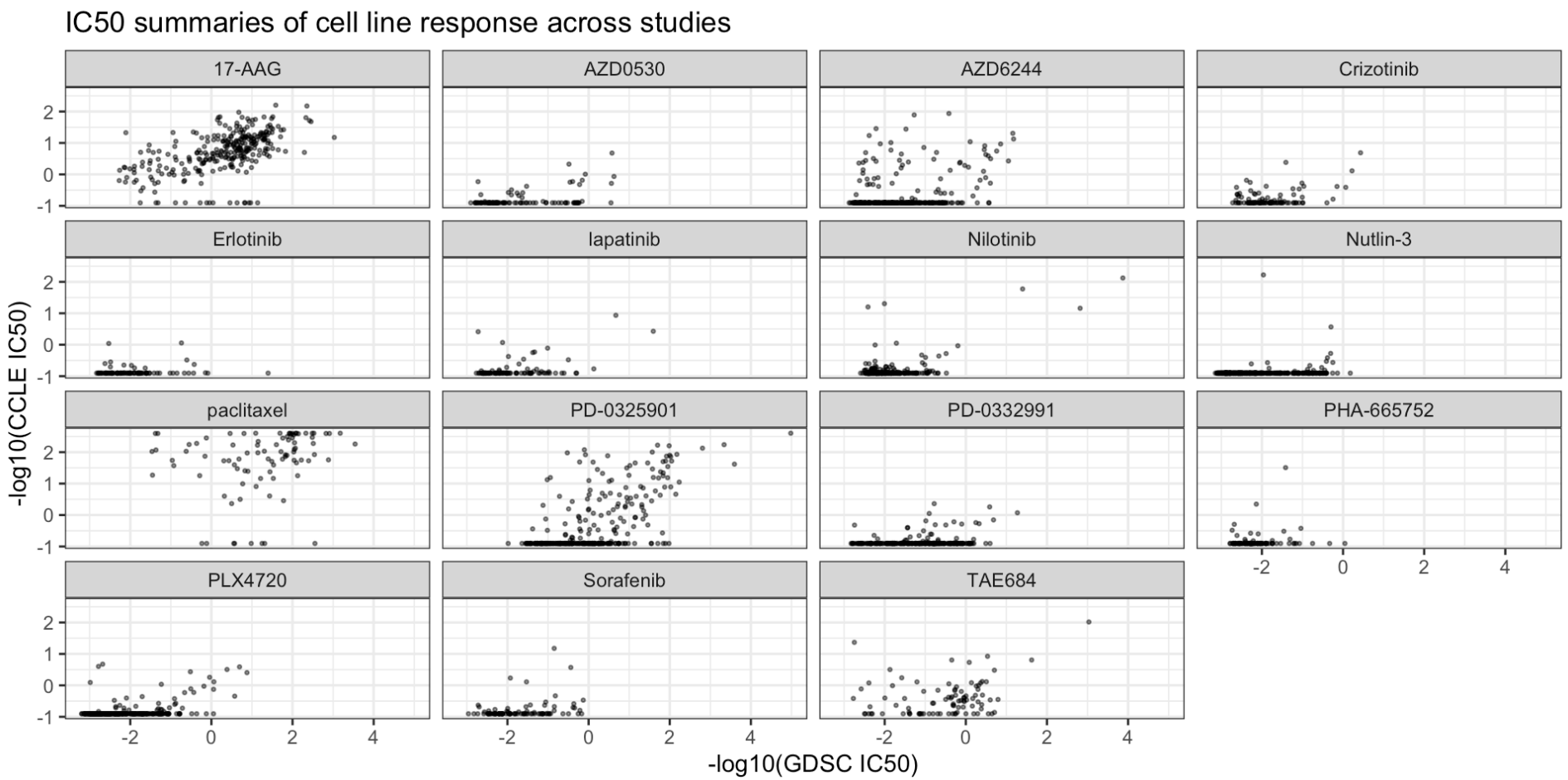
So far, we have only looked at how well the two studies agree for the response measurements of a single drug.

However, we need to look at the rest of the drugs to fully assess the level of replication.

First, we set out to reproduce Figure 2 in the Haibe-Kains paper, which displays scatter plots of IC50 values for the 15 drugs that were probed in both the CCLE and GDSC.

Essentially, we are going to make the plots from the previous section for all 15 drugs at once and compare them side-by-side.

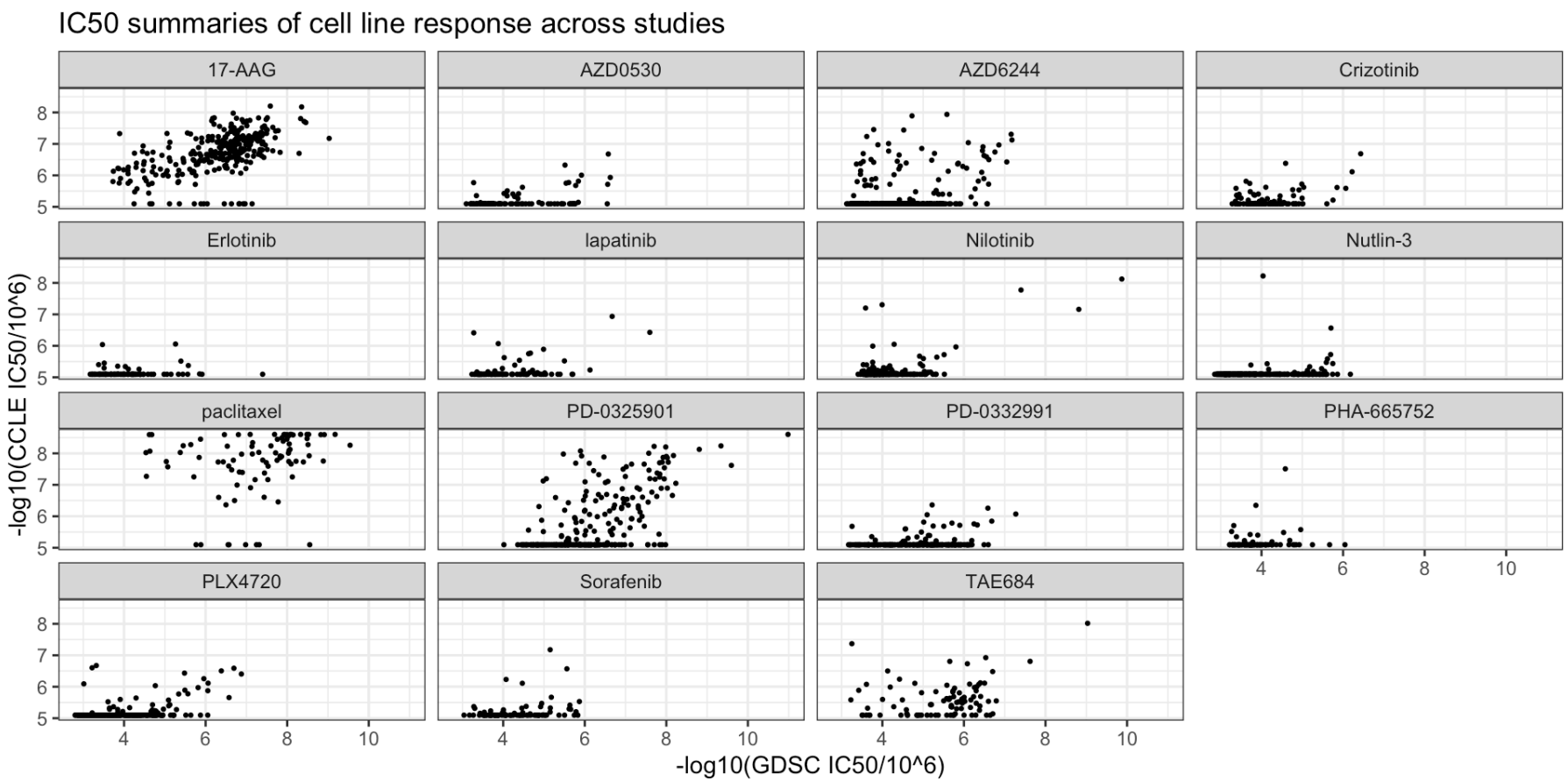
```r
summarizedData |>
    ggplot(aes(x = -log10(ic50_GDSC), y = -log10(ic50_CCLE))) +
    geom_point(alpha = 1/2, cex = 1/2) +
    facet_wrap(~ drug) +
    xlab("-log10(GDSC IC50)") +
    ylab("-log10(CCLE IC50)") +
    ggtitle("IC50 summaries of cell line response across studies")
```



Why do the ranges of the axes not exactly match the axes in Figure 2 of the published paper?

Note that the concentrations reported in the `summarizedData` dataset are given in units of mili-molar. However, the plots contain values calculated on the nano-molar scale (1 million times smaller). Thus, we can reproduce the original plot's axis labels by scaling the concentration values by 10^6.

```r
summarizedData |>
    ggplot(aes(x = -log10(ic50_GDSC / 10^6),
               y = -log10(ic50_CCLE / 10^6))) +
    geom_point(cex = 1/2) +
    facet_wrap(~ drug) +
    xlab("-log10(GDSC IC50/10^6)") +
    ylab("-log10(CCLE IC50/10^6)") +
    ggtitle("IC50 summaries of cell line response across studies")
```



Compare this plot to Figure 2 in the Haibe-Kains reanalysis paper. Does it seem to agree?

> Place your answer here

Looking at the IC50 values for both studies across all 15 drugs, would you say that they tend to agree? Why or why not?

> Place your answer here

Comparing Studies using Correlations

We have visually inspected the agreement of the drug response data between the two studies. Now, we'd like to go a step further and **quantify** the agreement with statistics. We'll do this using measures of correlation.

> If you aren't familiar with common measures of correlation (namely, Pearson and Spearman), don't worry! Check out the Supplementary Tutorial, **Supplement: Correlation Measures**, for a quick introduction!

Now we'll summarize different measures of correlation of the IC50 values to assess the level of replication between these two experiments.

First, we'll compute the two different measures for correlating continuous variables mentioned above (Pearson and Spearman correlation coefficients).

1. **Pearson's** correlation coefficient: measures the degree of *linearity* between continuous variables,
2. **Spearman's** correlation coefficient: measures the agreement of the *rankings* between variables.

```
## with the tidyverse
drugCorrs <- summarizedData |>
    group_by(drug) |>
    summarize(Pearson_ic50  = cor(-log10(ic50_GDSC / 10^6), -log10(ic50_CCLE / 10^6
              Spearman_ic50 = cor(-log10(ic50_GDSC / 10^6), -log10(ic50_CCLE / 10^6

drugCorrs
```

```
# A tibble: 15 × 3
   drug       Pearson_ic50 Spearman_ic50
   <chr>             <dbl>         <dbl>
 1 17-AAG           0.543         0.586
 2 AZD0530          0.455         0.360
 3 AZD6244          0.320         0.244
 4 Crizotinib       0.409         0.106
 5 Erlotinib        0.0812        0.0800
 6 Nilotinib        0.611         0.122
 7 Nutlin-3         0.143         0.306
 8 PD-0325901       0.625         0.580
 9 PD-0332991       0.240         0.141
10 PHA-665752       0.118         0.0554
```

11	PLX4720	0.456	0.358
12	Sorafenib	0.277	0.329
13	TAE684	0.287	0.268
14	lapatinib	0.427	0.289
15	paclitaxel	0.211	0.350

Next, we'll visualize the correlations the IC50 measurements in a grouped bar plot. To do this, we'll first have the reshape the data using the `pivot_longer` function from the tidyverse.

```
## with the tidyverse
drugCorrs <- drugCorrs |>
  pivot_longer(cols = c("Pearson_ic50", "Spearman_ic50"),
               names_to = "measure",
               values_to = "correlation")

drugCorrs
```

```
# A tibble: 30 × 3
   drug       measure       correlation
   <chr>      <chr>               <dbl>
 1 17-AAG     Pearson_ic50       0.543
 2 17-AAG     Spearman_ic50      0.586
 3 AZD0530    Pearson_ic50       0.455
 4 AZD0530    Spearman_ic50      0.360
 5 AZD6244    Pearson_ic50       0.320
 6 AZD6244    Spearman_ic50      0.244
 7 Crizotinib Pearson_ic50       0.409
 8 Crizotinib Spearman_ic50      0.106
 9 Erlotinib  Pearson_ic50       0.0812
10 Erlotinib  Spearman_ic50      0.0800
# ℹ 20 more rows
```

With this “tidy” data, we can now plot the two correlation measures across cell lines for each drug.

```
drugCorrs |>
    ggplot(aes(x = drug, y = correlation, fill = measure, group = measure)) +
    geom_bar(stat = "identity", position = position_dodge(), colour = "black") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_fill_grey() +
    ylim(0, 1) +
    ggtitle("Correlation of cell line IC50 summaries between studies for each drug"
```