Bioinformatics Scientist (4263) Programming test

January 28, 2023

1 1) Read in datasets

import libraries

[ ]:
```r
library(ggplot2)
library(readxl)
library(dplyr)
```

read in files as dataframes

[ ]:
```r
# read in files as dataframes (df)
olivieri <- read.csv("../data/Olivieri2020_drugz_subset.csv")

# add cell_line column to olivieri df (RPE1-hTERT)
olivieri$cell_line <- "RPE1-hTERT"

#reorder columns to GENE cell_line treatment normZ FDR
olivieri <- olivieri[c("GENE", "cell_line", "treatment", "normZ", "FDR")]

# rename columns to lowercase
colnames(olivieri) <- tolower(colnames(olivieri))

# read in xlsx, skip first row.
hustedt <- read_excel("../data/Hustedt et al. 2019 - results - rsob190156supp2.
  ↪xlsx", skip = 1)
```

```
New names:
• `` -> `…3`
• `` -> `…4`
• `` -> `…6`
• `` -> `…7`
• `` -> `…9`
• `` -> `…10`
• `` -> `…12`
• `` -> `…13`
```

preprocess hustedt df

[ ]:
```r
# make new df with columns 1-4
hustedt_subset_azd6738 <- hustedt[,c(1, 2, 3, 4)]
```

```
# make 2nd row the headers
colnames(hustedt_subset_azd6738) <- hustedt_subset_azd6738[1, ]

# remove 2nd row
hustedt_subset_azd6738 <- hustedt_subset_azd6738[-1, ]

# add cell_line column equal to RPE1-hTER
hustedt_subset_azd6738$cell_line <- "RPE1-hTERT"

# add treatment column equal to AZD6738
hustedt_subset_azd6738$treatment <- "AZD6738"

# rename column 1 to GENE
colnames(hustedt_subset_azd6738)[1] <- "GENE"

head(hustedt_subset_azd6738)
```

A tibble: 6 × 6

GENE	normZ	p-value	FDR	cell_line
<chr>	<chr>	<chr>	<chr>	<chr>
A1BG	-0.46	0.32400000000000001	1.1200000000000001	RPE1-hTERT
A1CF	0.54	0.70699999999999996	0.98399999999999999	RPE1-hTERT
A2M	0.13	0.55000000000000004	1.02	RPE1-hTERT
A2ML1	-0.25	0.40200000000000002	1.0900000000000001	RPE1-hTERT
A4GALT	-1.1299999999999999	0.129	1.27	RPE1-hTERT
A4GNT	1.35	0.91200000000000003	0.97099999999999997	RPE1-hTERT

preprocess 2nd subset from Hustedt et al. 2019

[ ]:

```
# make new df with columns 1, 5, 6 and 7
hustedt_subset_rpe1htert_ve821 <- hustedt[, c(1, 5, 6, 7)]

# make 2nd row the headers
colnames(hustedt_subset_rpe1htert_ve821) <- hustedt_subset_rpe1htert_ve821[1, ]

# remove 2nd row
hustedt_subset_rpe1htert_ve821 <- hustedt_subset_rpe1htert_ve821[-1, ]

# add cell_line column equal to RPE1-hTER
hustedt_subset_rpe1htert_ve821$cell_line <- "RPE1-hTERT"

# add treatment column equal to AZD6738
hustedt_subset_rpe1htert_ve821$treatment <- "VE821"

# rename column 1 to GENE
colnames(hustedt_subset_rpe1htert_ve821)[1] <- "GENE"

head(hustedt_subset_rpe1htert_ve821)
```

A tibble: 6 × 6

GENE <chr>	normZ <chr>	p-value <chr>	FDR <chr>	cell_line <chr>
A1BG	0.32	0.626	1	RPE1-hTERT
A1CF	-1.1000000000000001	0.13600000000000001	1.26	RPE1-hTERT
A2M	0.2	0.57899999999999996	1.02	RPE1-hTERT
A2ML1	-0.97	0.16500000000000001	1.26	RPE1-hTERT
A4GALT	-0.11	0.45500000000000002	1.07	RPE1-hTERT
A4GNT	0.39	0.65200000000000002	0.99299999999999999	RPE1-hTERT

preprocess 3rd subset Hustedt et al. 2019

[ ]:

```r
# make new df with columns 1, 8, 9 and 10
hustedt_subset_hela_ve821 <- hustedt[, c(1, 8, 9, 10)]

# make 2nd row the headers
colnames(hustedt_subset_hela_ve821) <- hustedt_subset_hela_ve821[1, ]

# remove 2nd row
hustedt_subset_hela_ve821 <- hustedt_subset_hela_ve821[-1, ]

# add cell_line column equal to RPE1-hTER
hustedt_subset_hela_ve821$cell_line <- "HeLa"

# add treatment column equal to AZD6738
hustedt_subset_hela_ve821$treatment <- "VE821"

# rename column 1 to GENE
colnames(hustedt_subset_hela_ve821)[1] <- "GENE"

head(hustedt_subset_hela_ve821)
```

A tibble: 6 × 6

GENE <chr>	normZ <chr>	p-value <chr>	FDR <chr>	cell_line <chr>	treatment <chr>
A1BG	0.03	0.51100000000000001	1.02	HeLa	VE821
A1CF	-0.64	0.26100000000000001	1.05	HeLa	VE821
A2M	1.05	0.85299999999999998	0.99299999999999999	HeLa	VE821
A2ML1	0.03	0.51100000000000001	1.02	HeLa	VE821
A4GALT	-1.9	2.8899999999999999E-2	0.83699999999999997	HeLa	VE821
A4GNT	-0.4	0.34300000000000003	1.04	HeLa	VE821

preprocess 4th subset Hustedt et al. 2019

[ ]:

```r
# make new df with columns 1, 11, 12 and 13
hustedt_subset_hct116_ve821 <- hustedt[, c(1, 11, 12, 13)]

# make 2nd row the headers
colnames(hustedt_subset_hct116_ve821) <- hustedt_subset_hct116_ve821[1, ]
# remove 2nd row
hustedt_subset_hct116_ve821 <- hustedt_subset_hct116_ve821[-1, ]
```

```
# add cell_line column equal to RPE1-hTER
hustedt_subset_hct116_ve821$cell_line <- "HCT116"

# add treatment column equal to AZD6738
hustedt_subset_hct116_ve821$treatment <- "VE821"

# rename column 1 to GENE
colnames(hustedt_subset_hct116_ve821)[1] <- "GENE"

head(hustedt_subset_hct116_ve821)
```

A tibble: 6 × 6

GENE <chr>	normZ <chr>	p-value <chr>	FDR <chr>	cell_line <chr>	treatment <chr>
A1BG	1.87	0.96899999999999997	0.998	HCT116	VE821
A1CF	-0.21	0.41499999999999998	1.02	HCT116	VE821
A2M	0.56999999999999995	0.71399999999999997	0.999	HCT116	VE821
A2ML1	-1.3	9.6000000000000002E-2	0.997	HCT116	VE821
A4GALT	0.56000000000000005	0.71099999999999997	0.999	HCT116	VE821
A4GNT	-0.09	0.46500000000000002	1.02	HCT116	VE821

merge the four subsets

[ ]:
```
# merge the four dataframes
hustedt <- rbind(hustedt_subset_azd6738, hustedt_subset_rpe1htert_ve821,␣
 ↪hustedt_subset_hela_ve821, hustedt_subset_hct116_ve821)

# make columns 2,3 and 4 numeric
hustedt[, 2:4] <- lapply(hustedt[, 2:4], as.numeric)

# reorder columns to GENE cell_line treatment normZ FDR
hustedt <- hustedt[c("GENE", "cell_line", "treatment", "normZ", "FDR",␣
 ↪"p-value")]

# rename columns to lowercase
colnames(hustedt) <- tolower(colnames(hustedt))

# remove p-value column
hustedt <- hustedt[, -6]

# remove NaNs
hustedt_nonan <- hustedt[complete.cases(hustedt), ]
olivieri_nonan <- olivieri[complete.cases(olivieri), ]
```

view files

[ ]:
```
head(olivieri_nonan)
```

A data.frame: 6 × 5

	gene	cell_line	treatment	normz	fdr
	<chr>	<chr>	<chr>	<dbl>	<dbl>
1	A1BG	RPE1-hTERT	Cisplatin1	-2.20	0.880
2	A1BG	RPE1-hTERT	IR	0.03	1.020
3	A1BG	RPE1-hTERT	UV	-1.14	1.320
4	A1BG	RPE1-hTERT	Olaparib	1.27	0.991
5	A1BG	RPE1-hTERT	AZD6738	-0.46	1.120
6	A1BG	RPE1-hTERT	Cisplatin2	0.83	0.922

[ ]:

```
head(hustedt_nonan)
```

A tibble: 6 × 5

gene	cell_line	treatment	normz	fdr
<chr>	<chr>	<chr>	<dbl>	<dbl>
A1BG	RPE1-hTERT	AZD6738	-0.46	1.120
A1CF	RPE1-hTERT	AZD6738	0.54	0.984
A2M	RPE1-hTERT	AZD6738	0.13	1.020
A2ML1	RPE1-hTERT	AZD6738	-0.25	1.090
A4GALT	RPE1-hTERT	AZD6738	-1.13	1.270
A4GNT	RPE1-hTERT	AZD6738	1.35	0.971

2 2) Explore both datasets

In the Olivieri dataset, not all genes are present in each treatment

[ ]:

```
# get number of unique genes in each treatment and cell line for each paper
olivieri_nonan %>% count(treatment, cell_line)
```

A data.frame: 9 × 3

treatment	cell_line	n
<chr>	<chr>	<int>
AZD6738	RPE1-hTERT	17272
Cisplatin1	RPE1-hTERT	17382
Cisplatin2	RPE1-hTERT	17249
Cisplatin3	RPE1-hTERT	17272
Formaldehyde	RPE1-hTERT	17293
IR	RPE1-hTERT	17315
KBrO3	RPE1-hTERT	17380
Olaparib	RPE1-hTERT	17277
UV	RPE1-hTERT	17361

In the Hustedt dataset, all treatments have an equal number (15910) of genes present. The are approximately 1400 fewer genes in the Hustedt dataset compared to the Olivieri dataset

[ ]:

```
# get number of unique genes in each treatment and cell line for each paper
hustedt_nonan %>% count(treatment, cell_line)
```

A tibble: 4 × 3

treatment	cell_line	n
<chr>	<chr>	<int>
AZD6738	RPE1-hTERT	15910
VE821	HCT116	15910
VE821	HeLa	15910
VE821	RPE1-hTERT	15910

2.1 2a)

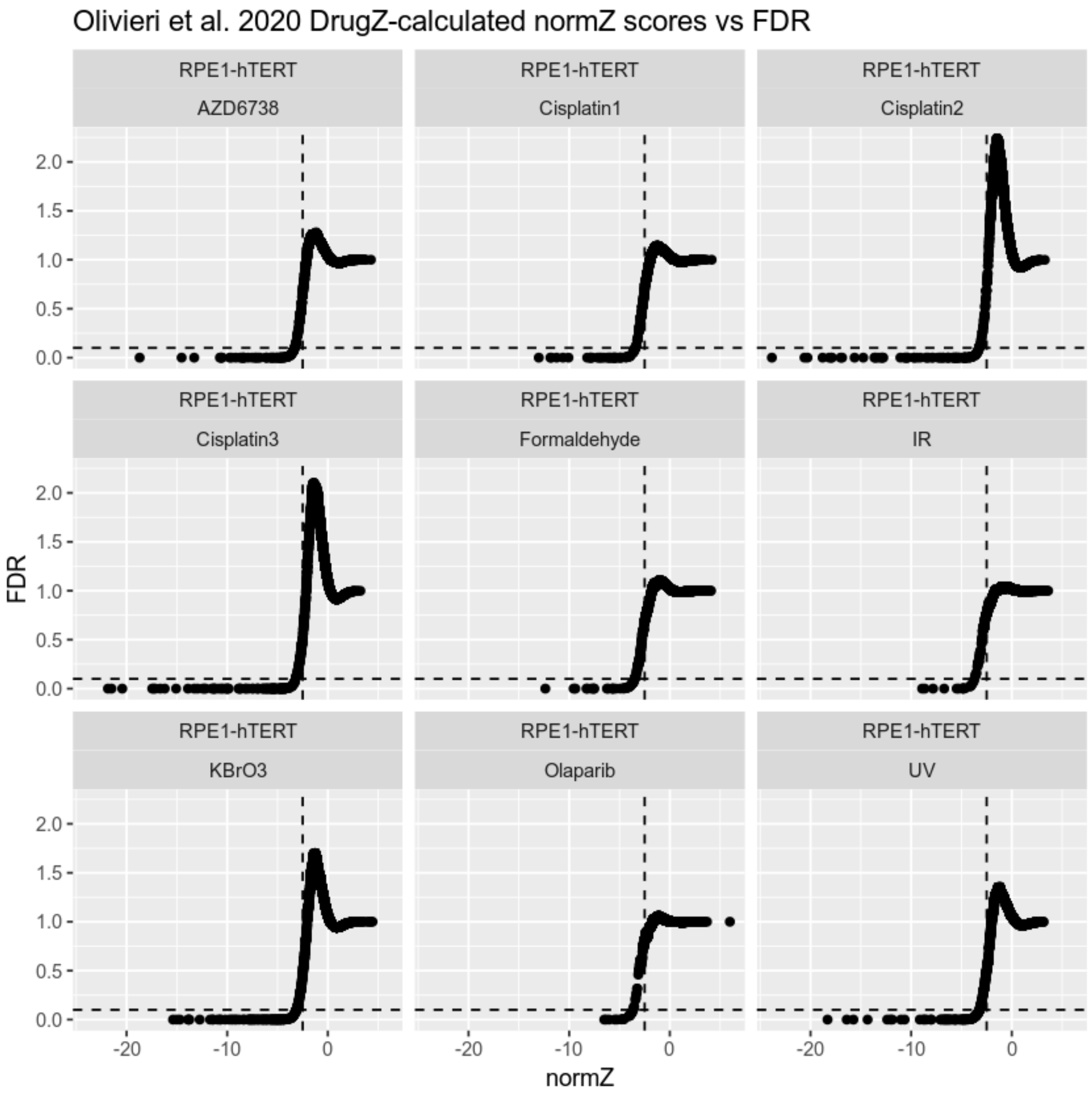
The shape of the plots of normZ values vs FDR in each cell line/treatment are roughly the same although Cisplatin2, Cisplatin3 and KBrO3 treatments have larger peaks with higher FDR scores between 0 and 2 normZ. AZD6738, Cisplatin2, Cisplatin3 and UV cause the most lethality, with many normZ scores under -15.

[ ]:

```
# create function for plotting the normZ values for each cell line and treatment
plotting <- function(df, cell_line, treatment, xlab, ylab, title) {
  ggplot(df, aes(x = normz, y = fdr)) + # Construct aesthetic mappings which␣
  ↪map
  # variables in the data to visual properties of the plot
    # Add points
    geom_point() +
    # Add y intercept at 0.1
    geom_hline(yintercept = 0.1, linetype = "dashed") +
    # Add x intercept at -2.5
    geom_vline(xintercept = -2.5, linetype = "dashed") +
    # Add a subplot titles
    facet_wrap(~ cell_line + treatment) +
    # Add axis labels and title
    labs(x = xlab, y = ylab, title = title)
}
```

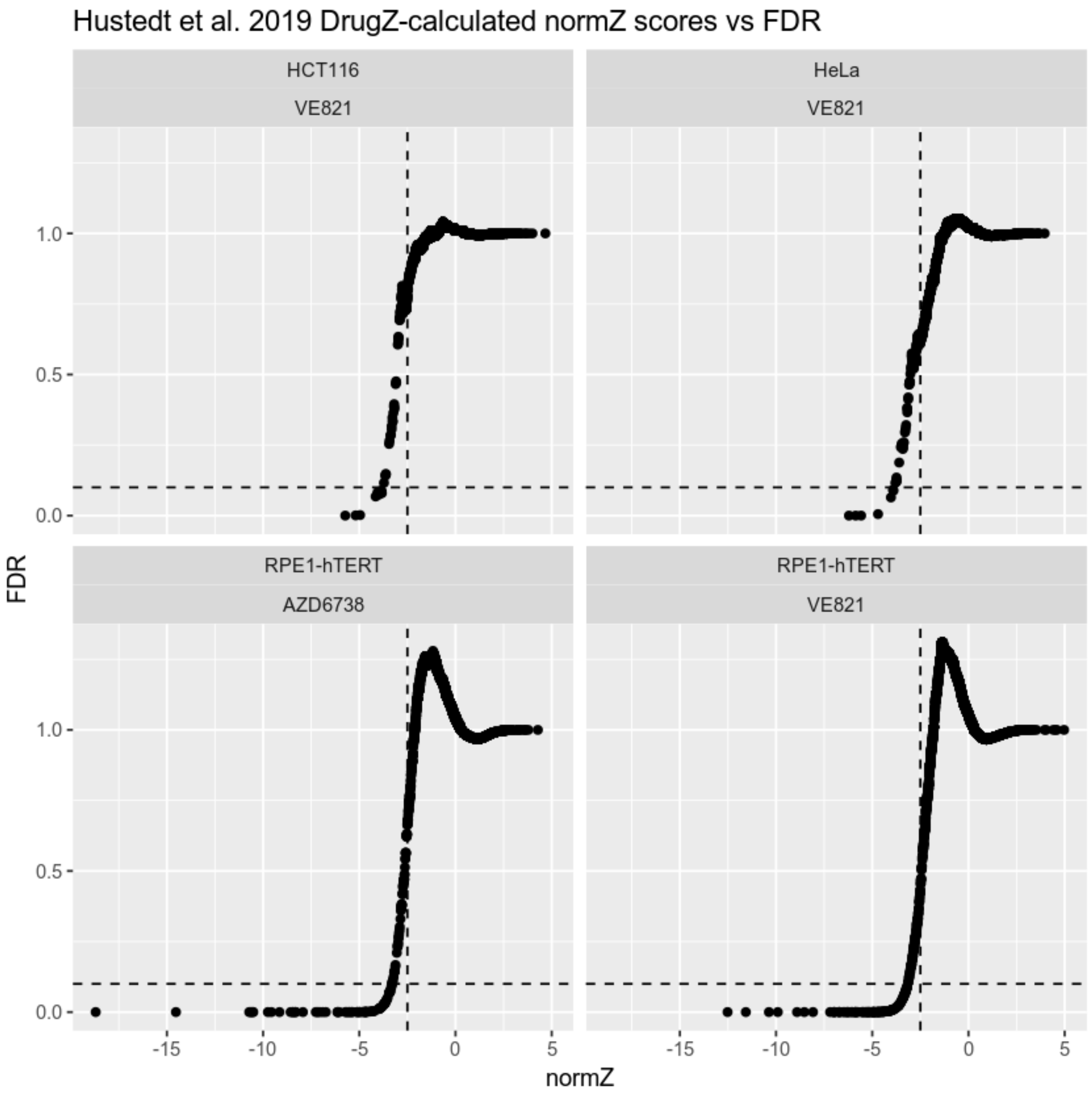
[ ]:

```
# plot the normZ values vs FDR for each cell line and treatment
plotting(olivieri_nonan, "cell_line", "treatment", "normZ", "FDR", "Olivieri et␣
  ↪al. 2020 DrugZ-calculated normZ scores vs FDR")
# save plot as pdf
ggsave("../data/Olivieri_normZ_FDR.pdf", width = 10, height = 10)
```



Again, the shape of the plots of normZ values vs FDR in each cell line/treatment in the Hustedt dataset are roughly the same although the RPE1-hTERT cell lines have larger peaks with higher FDR scores between 0 and 2 normZ. There are not many significantly lethal genes with normZ scores of -2.5 or under and FDR under 0.1 with the VE821 treatments

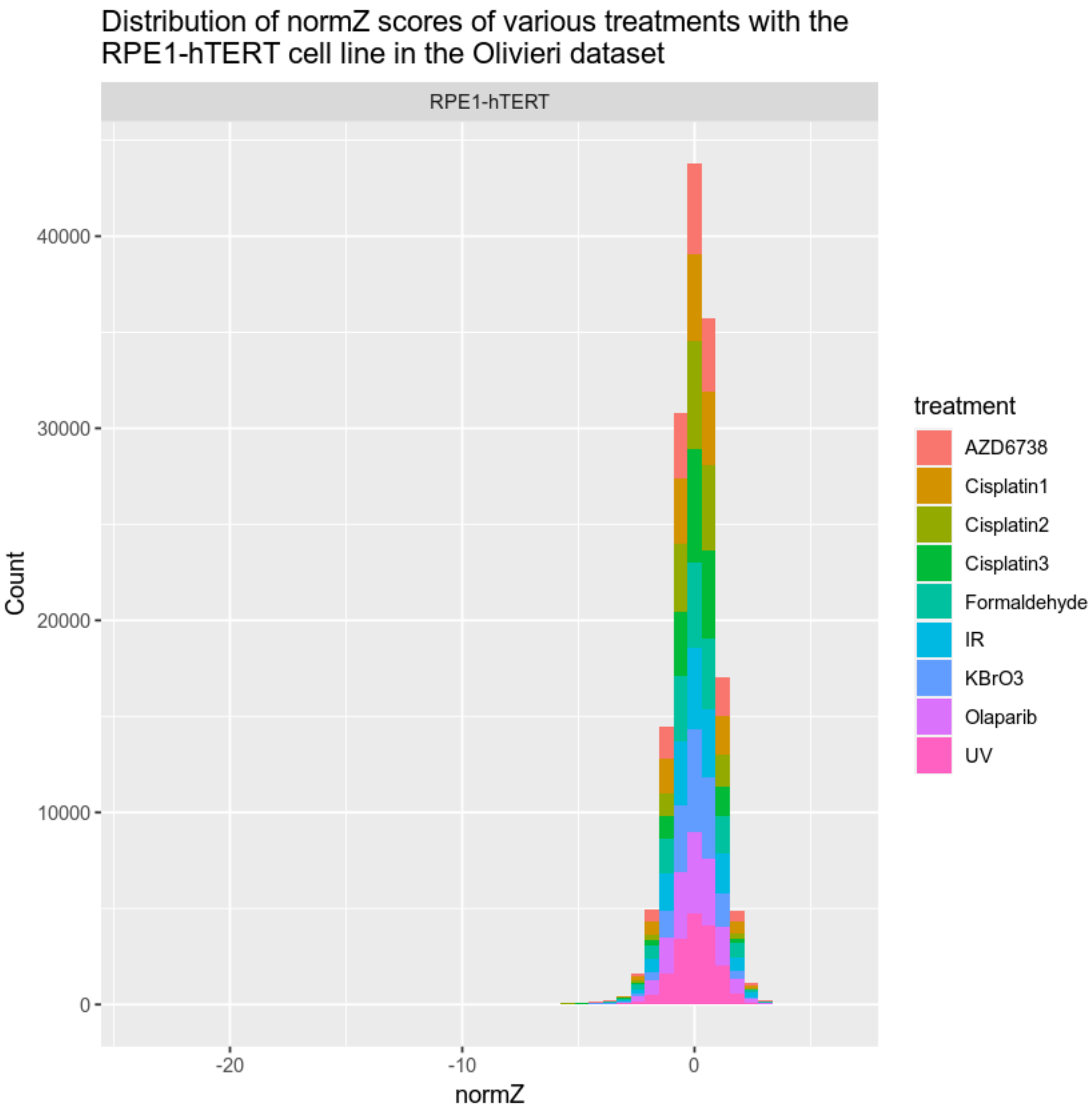
```
[ ]: # plot the normZ values vs FDR for each cell line and treatment, hustedt data
     plotting(hustedt_nonan, "cell_line", "treatment", "normZ", "FDR", "Hustedt et␣
      ↪al. 2019 DrugZ-calculated normZ scores vs FDR")
     # save plot as pdf
     ggsave("../data/Hustedt_normZ_FDR.pdf", width = 10, height = 10)
```



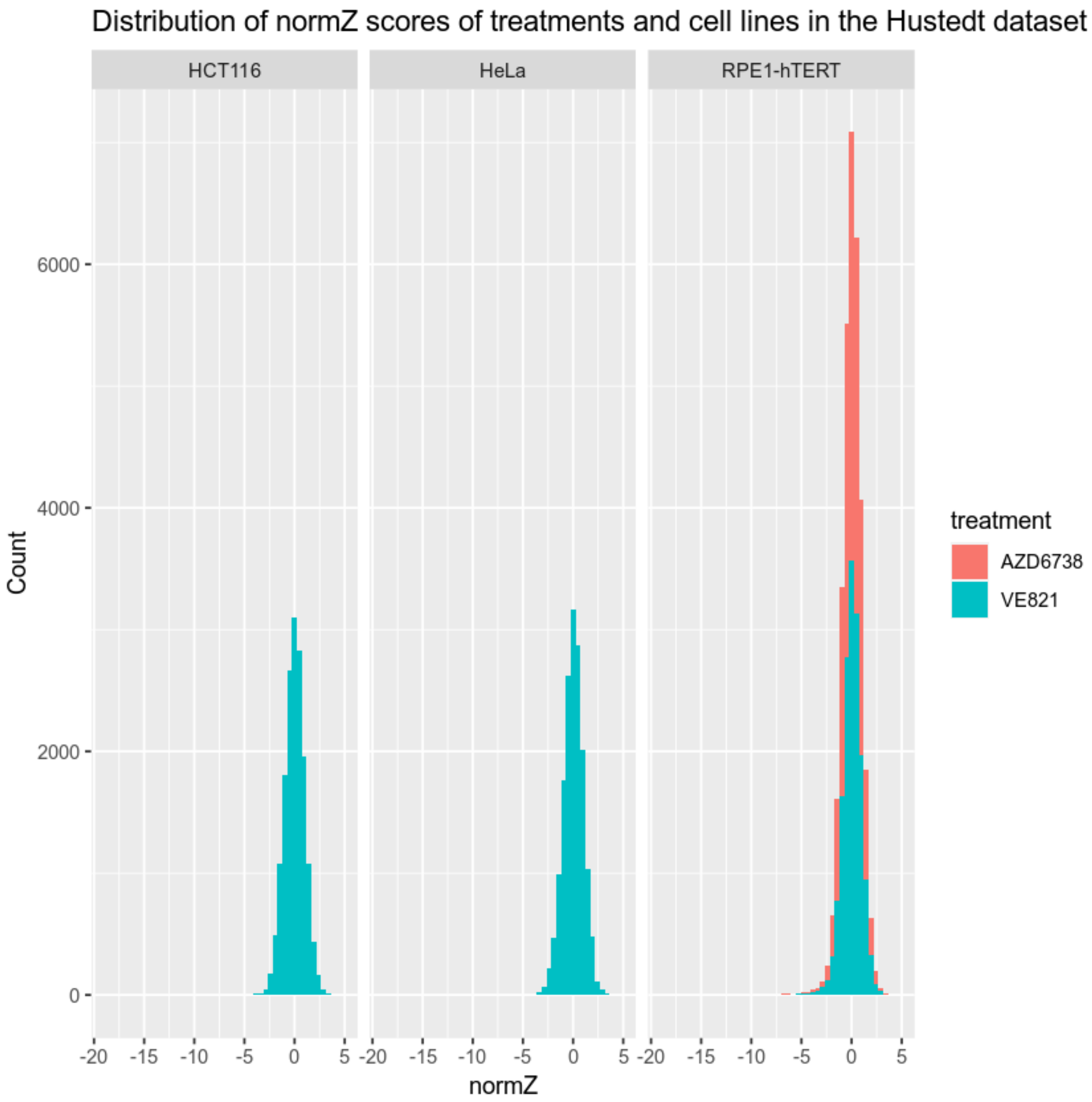
There is a slight negative skew, with a wider range of negative normZ scores than positive

```
[ ]: # plot the distribution of the normZ values for each cell line and treatment
ggplot(data = olivieri_nonan, aes(x = normz, fill = treatment)) +
  geom_histogram(bins = 50) +

  facet_wrap(~cell_line) +
  labs(x = "normZ", y = "Count", title = "Distribution of normZ scores of␣
↪various treatments with the \nRPE1-hTERT cell line in the Olivieri dataset")
```



```
[ ]: # plot the distribution of the normZ values for each cell line and treatment
     ggplot(data = hustedt_nonan, aes(x = normz, fill = treatment)) + # Construct␣
      ↪aesthetic mappings of variables to plot
       # display histogram
       geom_histogram(bins = 50) +
       # add subplot titles
       facet_wrap(~cell_line) +
       # add axis labels and title
       labs(x = "normZ", y = "Count", title = "Distribution of normZ scores of␣
      ↪treatments and cell lines in the Hustedt dataset")
```



plot RPE1-hTERT cell_line, AZD6738 treatment from olivieri_nonan with AZD6738 treatment from hustedt_nonan

[ ]:
```r
# merge the dfs, creating a column for the paper
olivieri_nonan$paper <- "Olivieri"
hustedt_nonan$paper <- "Hustedt"
merged <- merge(olivieri_nonan, hustedt_nonan, by = c("gene", "cell_line", "treatment", "normz", "fdr", "paper"), all = TRUE)

# remove duplicates from merged based on gene, cell_line, treatment, normz, fdr
merged_nodups <- merged[!duplicated(merged[, 1:5]), ]
merged_dups = merged[duplicated(merged[, 1:5]), ]
```

```
# compare RPE1-hTERT cell_line and AZD6738 treatment
merged_rpe1htert_azd6738 <- merged_nodups[merged_nodups$cell_line ==␣
 ↪"RPE1-hTERT" & merged_nodups$treatment == "AZD6738", ]
```

There are 15725 duplicate values between the two datasets!

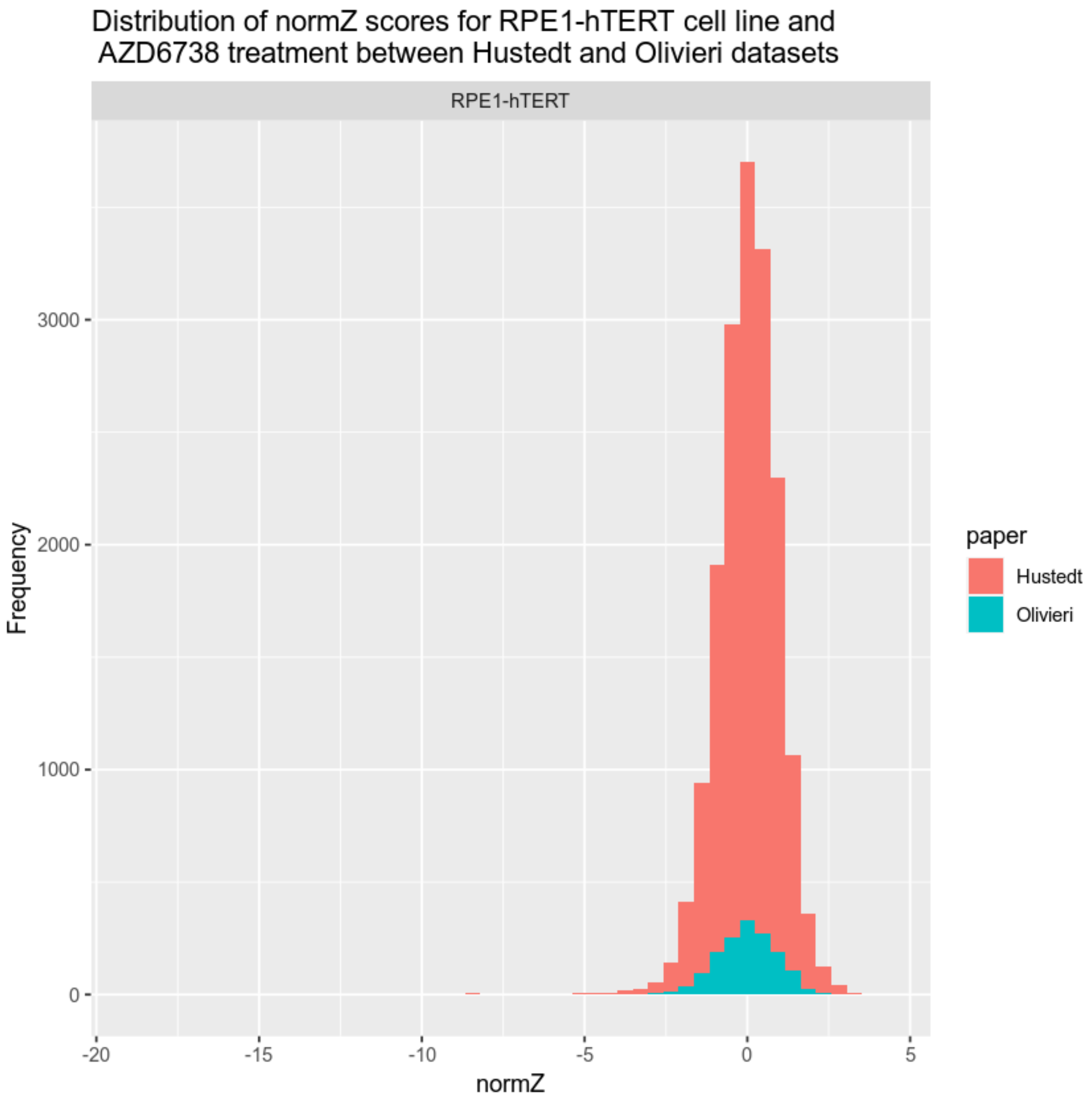
[ ]:
```
# number of genes in merged_dups
nrow(merged_dups)
```

15725

Using only non-duplicated values, the distribution of normZ scores between the two papers is similar, although the Hustedt dataset has a larger peak around 0. Ignoring the slight negative skew, the data look normally distributed

[ ]:
```
# plot the distribution of the normZ values for each paper
ggplot(data = merged_rpe1htert_azd6738, aes(x = normz, fill = paper), ) + #␣
 ↪Construct aesthetic mappings of variables to plot
  # display histogram
  geom_histogram(bins = 50) +
  # add subplot titles
  facet_wrap(~cell_line) +
  # add axis labels and title
  labs(x = "normZ", y = "Frequency", title = "Distribution of normZ scores for␣
 ↪RPE1-hTERT cell line and \n AZD6738 treatment between Hustedt and Olivieri␣
 ↪datasets")

# save plot as pdf
ggsave("../data/compare_datasets_RPE1hTERT_AZD6738.pdf", width = 10, height =␣
 ↪10)
```



```
[ ]: merged_rpe1htert_azd6738 %>% count(treatment, cell_line, paper)
```

A data.frame: 2 × 4

treatment <chr>	cell_line <chr>	paper <chr>	n <int>
AZD6738	RPE1-hTERT	Hustedt	15910
AZD6738	RPE1-hTERT	Olivieri	1547

2.2 2b) the datasets will be comparable if the intersect is taken based on genes present (so both datasets contain the same genes) since they are both normalised using DrugZ. This is assuming similar experimental conditions were used. The RPE1-hTERT cell lines are more comparable than the other cell lines, since they are present in both datasets. Note - some data points are present in both datasets so duplicates should be removed.

3 3)

3.1 3a) there are 3 cell lines in the two datasets (RPE1-hTERT, HeLa, HCT116)

3.2 3b) There are 10 treatments. i) AZD6738 and VE821 are most similar based on their distributions

3.3 3c) See the table below for the number of hits for each treatment/cell line per dataset

[ ]:

```r
# filter merged to include  normZ values <= -2.5 and FDR < 0.1
merged_filtered <- merged[merged$normz <= -2.5 & merged$fdr < 0.1, ]

# Count the number of genes in each treatment and cell line for each paper
merged_filtered %>% count(treatment, cell_line, paper)
```

A data.frame: 13 × 4

treatment	cell_line	paper	n
<chr>	<chr>	<chr>	<int>
AZD6738	RPE1-hTERT	Hustedt	76
AZD6738	RPE1-hTERT	Olivieri	84
Cisplatin1	RPE1-hTERT	Olivieri	52
Cisplatin2	RPE1-hTERT	Olivieri	104
Cisplatin3	RPE1-hTERT	Olivieri	124
Formaldehyde	RPE1-hTERT	Olivieri	39
IR	RPE1-hTERT	Olivieri	18
KBrO3	RPE1-hTERT	Olivieri	162
Olaparib	RPE1-hTERT	Olivieri	21
UV	RPE1-hTERT	Olivieri	114
VE821	HCT116	Hustedt	13
VE821	HeLa	Hustedt	7
VE821	RPE1-hTERT	Hustedt	125

3.4 3d) I would prioritise genes which have lower normZ scores and which are present in the most treatments. I would also prioritise those genes with low normZ scores which were similar between cell lines.

3.5 3e) I would visualise top synthetic lethal genes using a heatmap of gene vs treatment coloured by normZ score

4 4) You could explore the genes with the highest normZ scores (resistant genes). You could test which genes were most stably synthetic lethal or resistant across different treatments, and look at which genes were synthetic lethal or resistant with only certain treatments. You could also check whether the same genes were found to be most synthetic lethal/resistance between the two datasets.

5 5)

5.1 5a) no the ATRi treatments are not more similar to each other compared to other non-ATRi treatments (looking at the distributions). Differences between cell lines are larger.

5.2 5b) top 5 scoring genes are POLE3/4, RAD1, ATG9A and LCMT1

```
[ ]: # filter merged_filtered to include only AZD6738 and VE821 treatment
     merged_filtered_atri <- merged_filtered[merged_filtered$treatment == "AZD6738"␣
       ↪| merged_filtered$treatment == "VE821", ]



     # rank by normZ, ascending
     merged_filtered_atri_ranked <-␣
       ↪merged_filtered_atri[order(merged_filtered_atri$normz), ]

     # remove duplicates, ignoring paper
     merged_filtered_atri_nodups <- merged_filtered_atri_ranked[!
       ↪duplicated(merged_filtered_atri_ranked[, 1:5]), ]

     head(merged_filtered_atri_nodups)
```

A data.frame: 6 × 6

	gene <chr>	cell_line <chr>	treatment <chr>	normz <dbl>	fdr <dbl>	paper <chr>
145163	POLE4	RPE1-hTERT	AZD6738	-18.70	4.48e-74	Hustedt
145150	POLE3	RPE1-hTERT	AZD6738	-14.53	3.64e-44	Hustedt
154350	RAD1	RPE1-hTERT	AZD6738	-13.28	8.40e-37	Olivieri
14036	ATG9A	RPE1-hTERT	VE821	-12.52	4.57e-32	Hustedt
145160	POLE3	RPE1-hTERT	VE821	-11.57	2.30e-27	Hustedt
102051	LCMT1	RPE1-hTERT	AZD6738	-10.72	1.81e-23	Hustedt

```
[ ]: #get unique cell lines merged_filtered_atri
     unique(merged_filtered_atri$cell_line)
```

1. 'RPE1-hTERT' 2. 'HCT116' 3. 'HeLa'

5.3 5c) consensus genes would be those that are the top ranked in both ATRi treatments i) I would expect cell line specific effects too, which might be due to batch or biological effects specific to each cell line

5.4 5d) You could check whether genes were enriched in a certain molecular pathway using a Gene Ontology / pathway enrichment analysis