compare-expression-distributions-in-different-comparison cohorts 2023.11.20 11.35.44

hbeale

November 20, 2023

Contents

```
outliers <- read_tsv("../input_data/druggable_outliers_from_treehouse_and_other_cohorts_2023_11_09-13_4
  mutate(high_level_cohort = ifelse(str_detect(comparison_cohort, "Treehouse"),
                                    "Treehouse",
                                    comparison_cohort))
```

```
## Rows: 287 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): Sample_ID, comparison_cohort, gene, donor_ID
## lgl (1): pathway_support
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

COMPARE DISTRIBUTIONS FOR FOR OUTLIERS ACROSS COHORTS

```
outlier_genes_detected <- unique(outliers$gene)

expr <- read_tsv("../input_data/druggable_TumorCompendium_v11_PolyA_hugo_log2tpm_58581genes_2020-04-09.
  rename(Sample_ID = TH_id) %>%
  filter(Gene %in% outlier_genes_detected)
```

```
## Rows: 1414917 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): Gene, TH_id
## dbl (1): log2TPM1
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
stanford_samples  <- read_tsv("../gather_input_data/comparison_to_non_CARE_cohorts/data/TH03_TH34_rollu
                              col_names = "Sample_ID") %>%
  mutate(cohort = "TH03_TH34")
```

```
## Rows: 110 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Sample_ID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
TCGA_samples  <- read_tsv("../gather_input_data/comparison_to_non_CARE_cohorts/data/TCGA_rollup.sample_
                              col_names = "Sample_ID") %>%
  mutate(cohort = "TCGA")
```

```
## Rows: 9806 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Sample_ID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
PEDAYA_samples  <- read_tsv("../gather_input_data/comparison_to_non_CARE_cohorts/data/PEDAYA_rollup.samp
                              col_names = "Sample_ID") %>%
  mutate(cohort = "PEDAYA")
```

```
## Rows: 2814 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Sample_ID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pan_cancer_samples <- expr %>%
  select(Sample_ID) %>%
  distinct() %>%
  mutate(cohort = "Treehouse_pc")


samples_in_cohorts <- bind_rows(
  stanford_samples,
  TCGA_samples,
  PEDAYA_samples,
  pan_cancer_samples)


tabyl(samples_in_cohorts,
      cohort)
```

```
##        cohort     n    percent
##        PEDAYA  2814 0.11045257
##          TCGA  9806 0.38489618
##     TH03_TH34   110 0.00431762
##  Treehouse_pc 12747 0.50033363
```

expression in samples not in the compendium

```
rsem_path <- "../input_data/non_compendium_expression"

gene_name_conversion <- read_tsv(file.path(rsem_path,
                                           "EnsGeneID_Hugo_Observed_Conversions.txt"))
```

```
## Rows: 60498 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): HugoID, EnsGeneID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
relevant_gene_name_conversion <- gene_name_conversion %>%
  filter(HugoID %in% outlier_genes_detected)

rsem_kitchen_sink_data <- tibble(file_name = list.files(
  path = rsem_path,
  pattern = "_rsem_genes.results")) %>%
  rowwise() %>%
  mutate(rsem_raw = list(read_tsv(file.path(rsem_path, file_name),
                                  show_col_types = FALSE
                                  ))) %>%
  unnest(rsem_raw) %>%
  filter(gene_id %in% relevant_gene_name_conversion$EnsGeneID) %>%
  mutate(Sample_ID = str_extract(file_name, "TH[R]?[0-9]{2}_[0-9]{4}_S[0-9]{2}")) %>%
  left_join(relevant_gene_name_conversion,
            by=c("gene_id"="EnsGeneID")) %>%
    group_by(Sample_ID, HugoID) %>%
    summarize(sum_TPM = sum(TPM),
              n=n()) %>%
    mutate(log2TPM1 = log2(sum_TPM +1))
```

```
## `summarise()` has grouped output by 'Sample_ID'. You can override using the
## `.groups` argument.
```

```
table(rsem_kitchen_sink_data$n)
```

```
##
```

```
##   1   2
## 275   5
```

```
patient_expression_from_rsem_files <- rsem_kitchen_sink_data %>%
  select(gene = HugoID,
         log2TPM1,
         Sample_ID)

patient_expression_from_compendia <- outliers %>%
  select(Sample_ID, gene) %>%
  distinct() %>%
  left_join(expr,
            by=c("Sample_ID", "gene"="Gene")) %>%
  na.omit() # excludes samples not in compendium

patient_expression <- bind_rows(
  patient_expression_from_rsem_files,
  patient_expression_from_compendia)

length(outlier_genes_detected)
```

```
## [1] 56
```

```
outliers$Sample_ID[ ! outliers$Sample_ID %in% expr$Sample_ID] %>% unique()
```

```
## [1] "TH34_1400_S01" "TH34_2292_S01" "TH34_2666_S01" "TH34_1445_S02"
## [5] "TH34_1456_S02"
```

How many colors do i need

```
outliers %>%
  group_by(gene) %>%
  summarize(n_samples = length(unique(Sample_ID))) %>%
  arrange(desc(n_samples))
```

```
## # A tibble: 56 x 2
##    gene  n_samples
##    <chr>     <int>
##  1 IGF2         18
##  2 HMOX1         8
##  3 NTRK2         7
##  4 FGFR4         5
##  5 ETV1          4
##  6 NTRK3         4
##  7 BTK           3
##  8 CDK9          3
##  9 FGFR1         3
## 10 FLT4          3
## # i 46 more rows
```

Calculate statistics for each cohort

```
cohort_thresholds_raw <- left_join(samples_in_cohorts,
                                   expr,
                                   by=c("Sample_ID")) %>%
  group_by(Gene, cohort) %>%
  summarize(q25 = quantile(log2TPM1, 0.25),
            median = median(log2TPM1),
            q75 = quantile(log2TPM1, 0.75),
            IQR = q75-q25,
            up_outlier_threshold = q75 + (1.5*IQR))
```

```
## `summarise()` has grouped output by 'Gene'. You can override using the
## `.groups` argument.
```

pediatric vs TCGA for one gene

```
this_gene <- "ETV1"

cohort_thresholds_raw %>%
  filter(Gene == this_gene) %>%
  group_by(Gene) %>%
  pivot_longer(c(-Gene, -cohort)) %>%
  pivot_wider(names_from = cohort, values_from = value) %>%
  mutate(change_in_ped_relative_to_TCGA =
              (PEDAYA - TCGA) / TCGA,
         change_in_treehouse_relative_to_TCGA =
            (Treehouse_pc - TCGA) / Treehouse_pc) %>%
  select(-TH03_TH34)
```

```
## # A tibble: 5 x 7
## # Groups:   Gene [1]
##   Gene  name                 PEDAYA  TCGA Treehouse_pc change_in_ped_relative_~1
##   <chr> <chr>                 <dbl> <dbl>        <dbl>                     <dbl>
## 1 ETV1  q25                   0.202  1.30        0.978                    -0.845
## 2 ETV1  median                1.82   2.23        2.09                     -0.183
## 3 ETV1  q75                   4.30   3.52        3.52                      0.222
## 4 ETV1  IQR                   4.10   2.22        2.55                      0.851
## 5 ETV1  up_outlier_threshold 10.5    6.84        7.34                      0.528
## # i abbreviated name: 1: change_in_ped_relative_to_TCGA
## # i 1 more variable: change_in_treehouse_relative_to_TCGA <dbl>
```

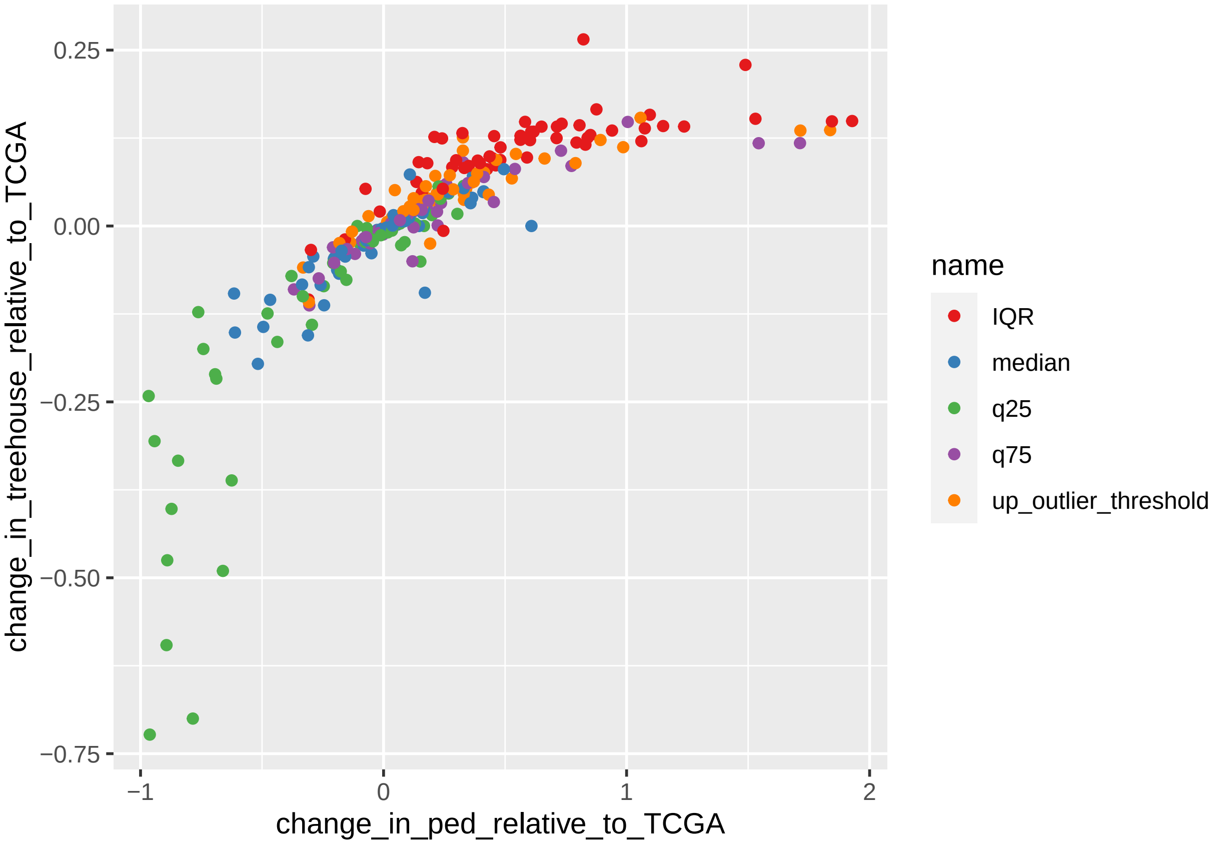
```
# # is the biggest change to the median or to the IQR?
#   summarize(q_25_change_in_ped_relative_to_TCGA =
#               (q25[cohort == "PEDAYA"] - q25[cohort == "TCGA"]) / q25[cohort == "TCGA"])
```

asess changes for all genes

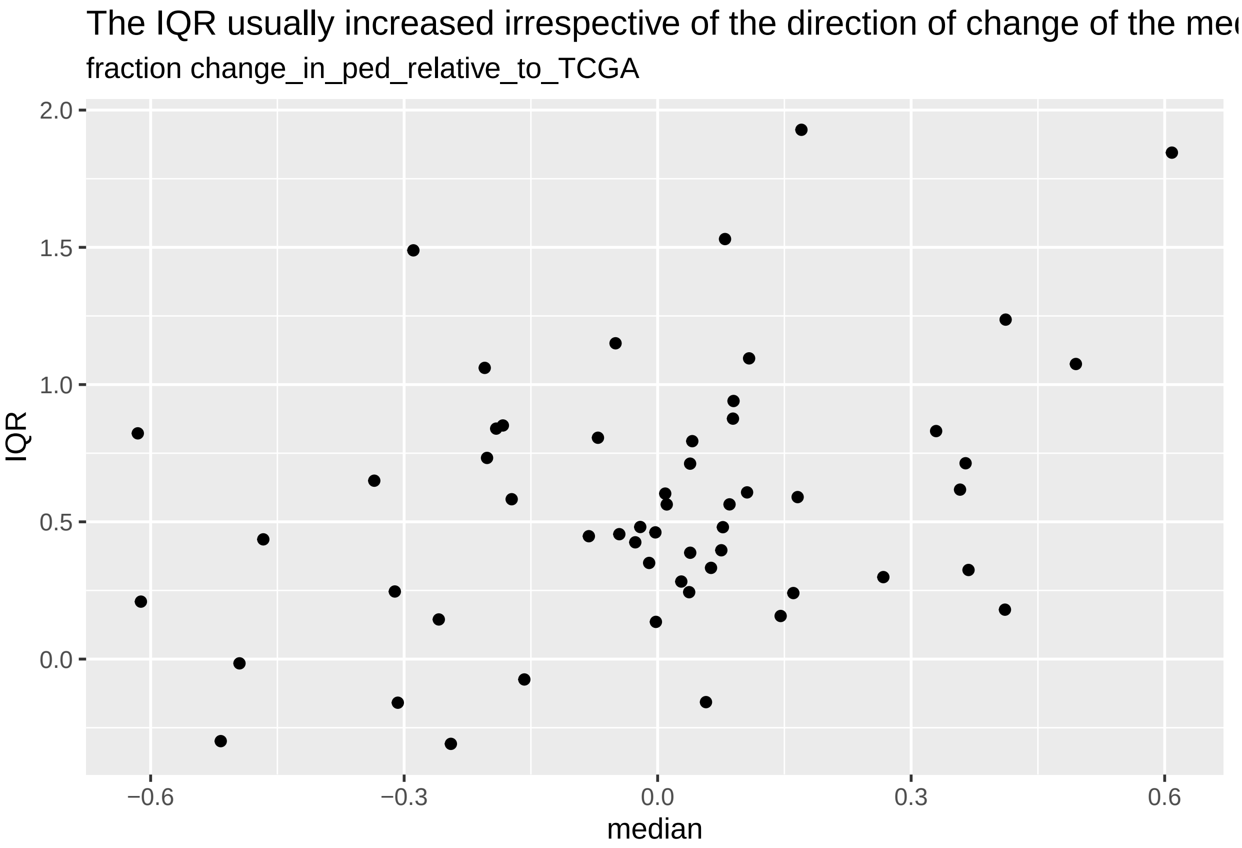
```
cohort_thresholds <- cohort_thresholds_raw %>%
  pivot_longer(c(-Gene, -cohort)) %>%
  pivot_wider(names_from = cohort, values_from = value) %>%
  mutate(change_in_ped_relative_to_TCGA =
              (PEDAYA - TCGA) / TCGA,
         change_in_treehouse_relative_to_TCGA =
            (Treehouse_pc - TCGA) / Treehouse_pc)
```

changes plotted

```
ggplot(cohort_thresholds) +
  geom_point(aes(x=change_in_ped_relative_to_TCGA,
                 y=change_in_treehouse_relative_to_TCGA,
                 color = name)) +
  scale_color_brewer(palette = "Set1")
```



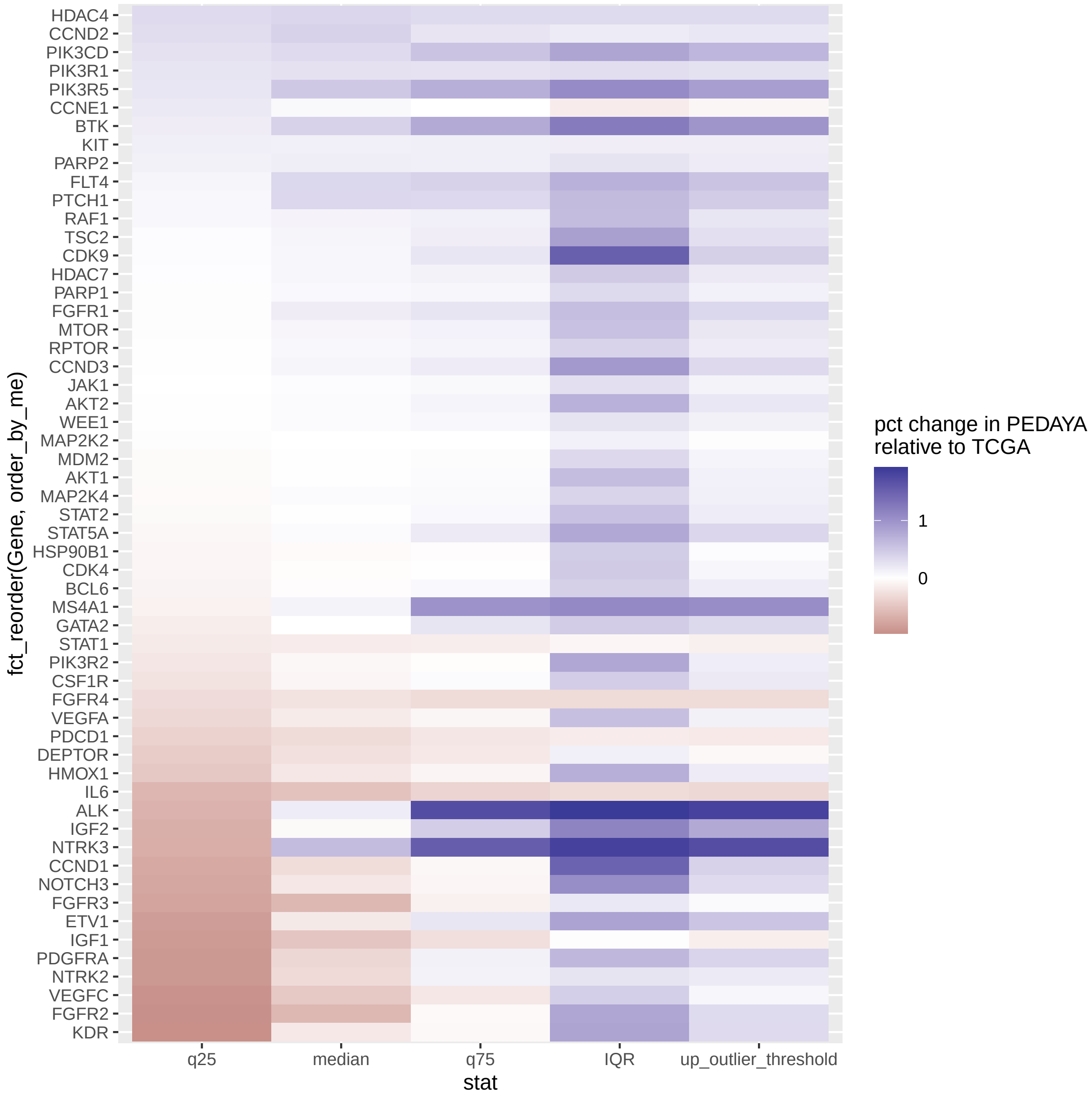
```
cohort_thresholds %>%
  filter(name %in% c("median", "IQR")) %>%
  select(Gene, name, change_in_ped_relative_to_TCGA) %>%
  pivot_wider(names_from = name,
              values_from = change_in_ped_relative_to_TCGA) %>%
  ggplot +
  geom_point(aes(x=median, y=IQR)) +
  ggtitle("The IQR usually increased irrespective of the direction of change of the median", "fraction
```



```
cohort_thresholds_for_plot <- cohort_thresholds %>%
        rename(stat = name) %>%
  mutate(stat = factor(stat, levels = c("q25", "median", "q75", "IQR", "up_outlier_threshold"))) %>%
  group_by(Gene) %>%
        mutate(order_by_me = change_in_ped_relative_to_TCGA[stat == "q25"])

# %>%
#   ungroup %>%
#   mutate(Gene = factor(Gene) %>% fct_reorder(Gene, order_by_me, .fun = min))
# levels(cohort_thresholds_for_plot$Gene)


ggplot(cohort_thresholds_for_plot) +
  #geom_tile(aes(x=stat, y= Gene, fill = change_in_ped_relative_to_TCGA)) +
  geom_tile(aes(x=stat, y= fct_reorder(Gene, order_by_me), fill = change_in_ped_relative_to_TCGA)) +
  scale_fill_gradient2("pct change in PEDAYA\nrelative to TCGA")
```



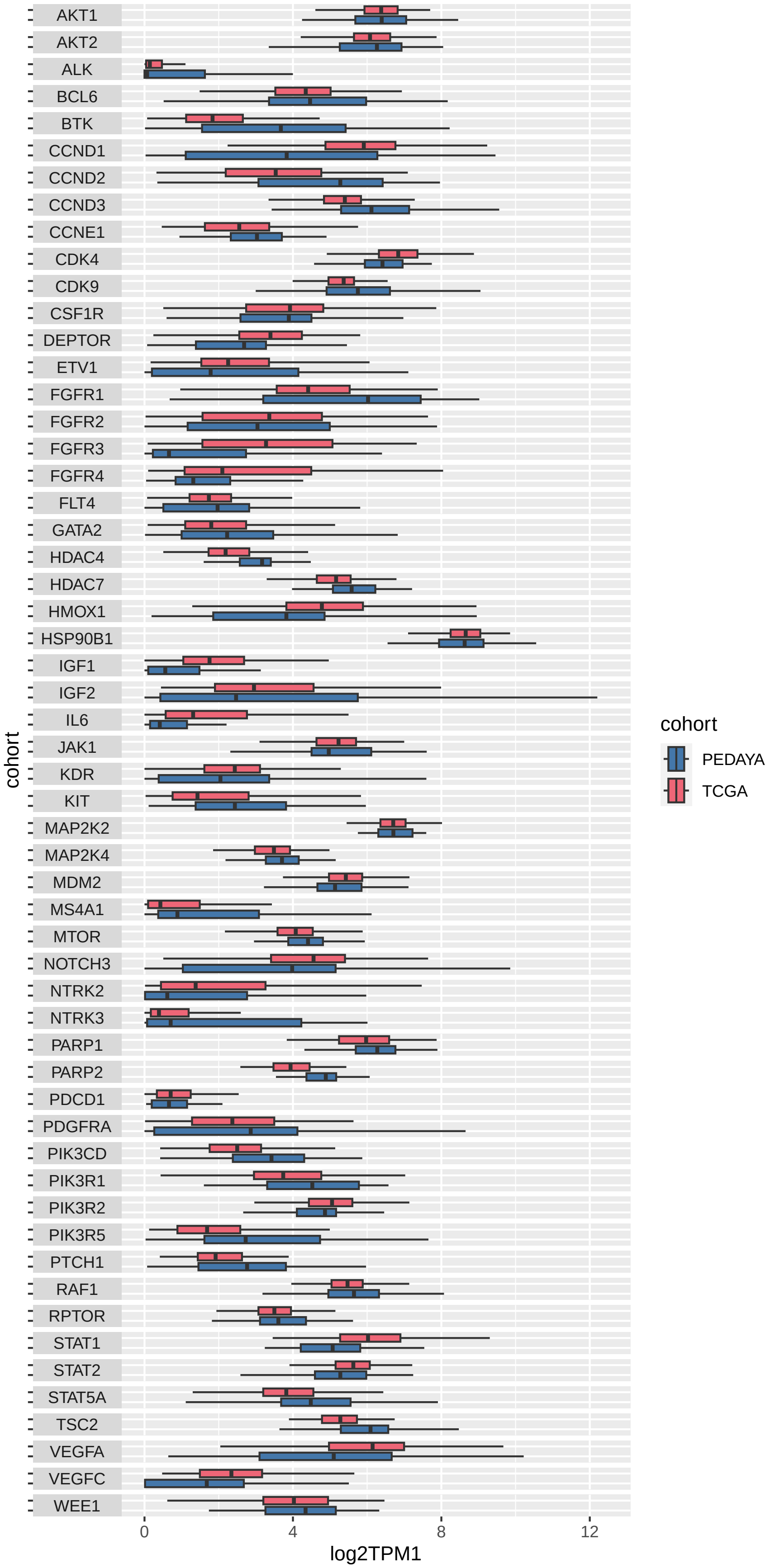
```
#geom_tile(aes(x=stat, y= fct_reorder(Gene, change_in_ped_relative_to_TCGA), fill = change_in_ped_relat
```

plot boxplots for TCGA and PEDAYA

```
TP_cohort_thresholds_raw <- left_join(samples_in_cohorts %>%
                                     filter(cohort %in% c("PEDAYA", "TCGA")),
                                   expr,
                                   by=c("Sample_ID"))
```

```
TP_cohort_thresholds_raw_subset <- TP_cohort_thresholds_raw %>%
  slice_sample(n = 10000)
```

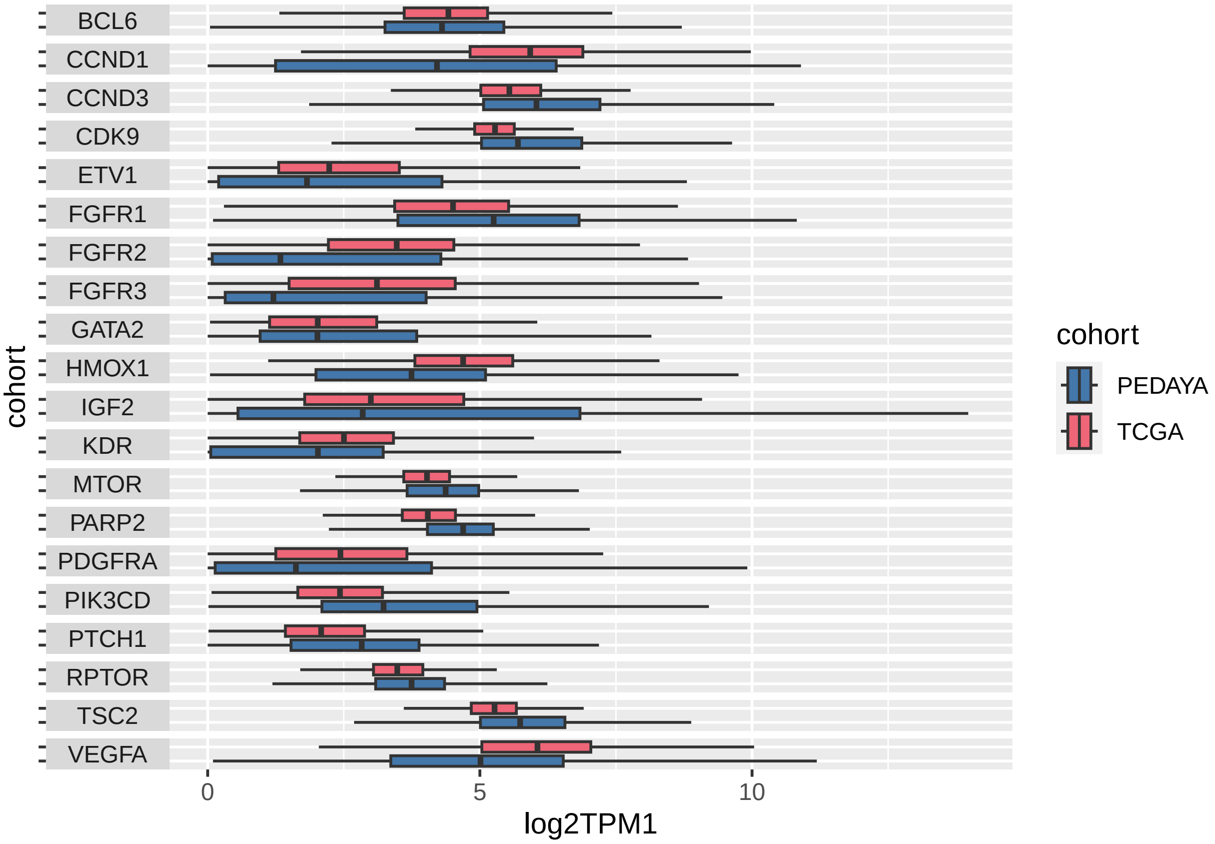
```r
ggplot(TP_cohort_thresholds_raw_subset) +
  geom_boxplot(aes(y=cohort, x=log2TPM1,
                   fill = cohort),
               outlier.shape = NA) +
  facet_wrap(~Gene, ncol = 1,
             strip.position = "left") +
 theme(strip.text.y.left = element_text(angle = 0),
       axis.text.y = element_blank(),
        panel.spacing = unit(0.2, "lines"))  +
  scale_fill_bright()
```


AKT1
AKT2
ALK
BCL6
BTK
CCND1
CCND2
CCND3
CCNE1
CDK4
CDK9
CSF1R
DEPTOR
ETV1
FGFR1
FGFR2
FGFR3
FGFR4
FLT4
GATA2
HDAC4
HDAC7
HMOX1
HSP90B1
IGF1
IGF2
IL6
JAK1
KDR
KIT
MAP2K2
MAP2K4
MDM2
MS4A1
MTOR
NOTCH3
NTRK2
NTRK3
PARP1
PARP2
PDCD1
PDGFRA
PIK3CD
PIK3R1
PIK3R2
PIK3R5
PTCH1
RAF1
RPTOR
STAT1
STAT2
STAT5A
TSC2
VEGFA
VEGFC
WEE1
cohort
0
4
8
12
log2TPM1
cohort
PEDAYA
TCGA

```
TCGA_not_Treehouse_pc_outliers <- outliers %>%
  group_by(gene, Sample_ID) %>%
  mutate(TCGA_not_Treehouse_pc = "TCGA" %in% comparison_cohort &
           ! "Treehouse_pc" %in% comparison_cohort) %>%
  filter(TCGA_not_Treehouse_pc) %>%
  arrange(Sample_ID, gene)
```

```
TP_cohort_thresholds_raw_subset <- TP_cohort_thresholds_raw %>%
  slice_sample(n = 10000)

ggplot(TP_cohort_thresholds_raw %>%
         filter(Gene %in% TCGA_not_Treehouse_pc_outliers$gene)) +
  geom_boxplot(aes(y=cohort, x=log2TPM1,
                   fill = cohort),
               outlier.shape = NA) +
  facet_wrap(~Gene, ncol = 1,
             strip.position = "left") +
 theme(strip.text.y.left = element_text(angle = 0),
      axis.text.y = element_blank(),
       panel.spacing = unit(0.2, "lines"))  +
  scale_fill_bright()
```


BCL6
CCND1
CCND3
CDK9
ETV1
FGFR1
FGFR2
FGFR3
GATA2
HMOX1
IGF2
KDR
MTOR
PARP2
PDGFRA
PIK3CD
PTCH1
RPTOR
TSC2
VEGFA
cohort
log2TPM1
cohort
PEDAYA
TCGA