compare-expression-distributions-in-different-comparison cohorts 2023.11.21 11.24.15

hbeale

November 27, 2023

Contents

Version 2023.11.21_11.24.15 - focuses on genes found only relative to TCGA, not relative to any other cohort. Previous versions focused on outliers detected relative to TCGA and not treehouse, irrespective of whatever other cohorts they were outliers in

```
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
```

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.
```

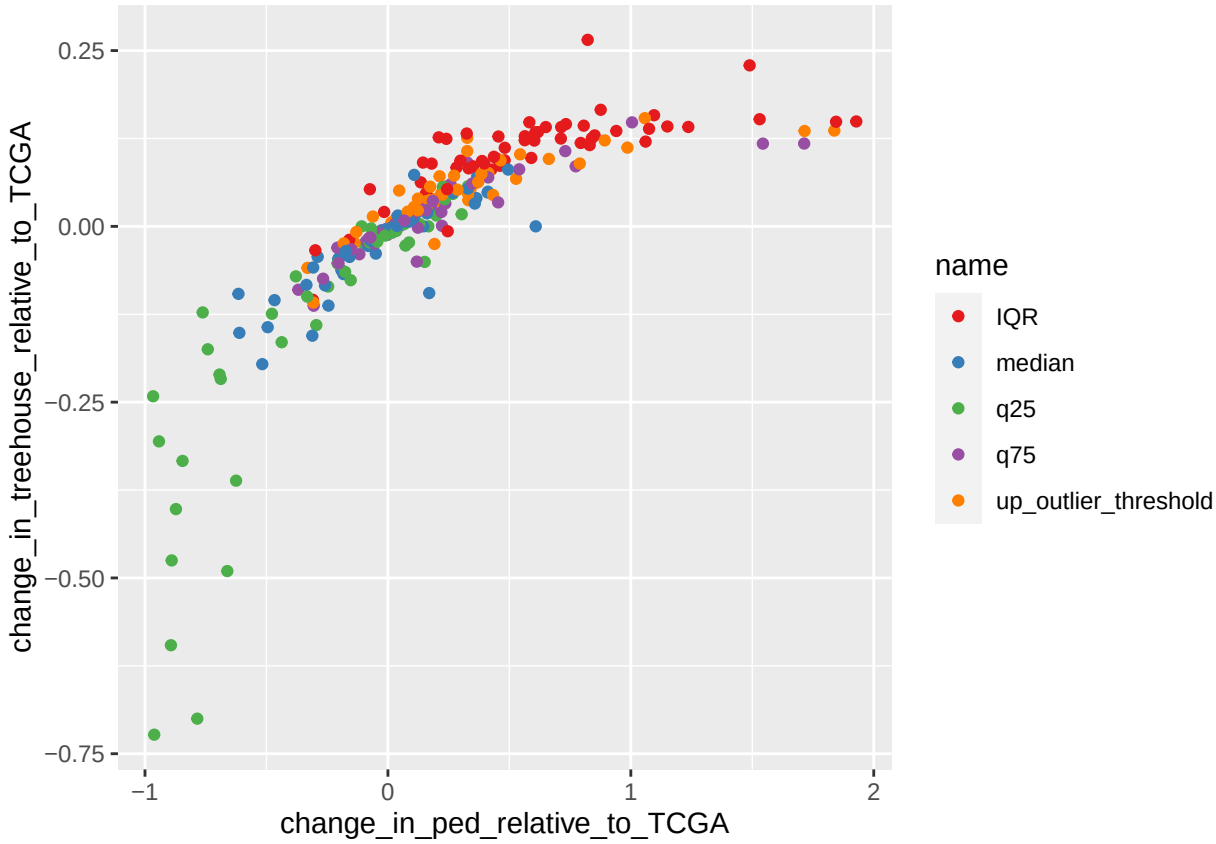
assess changes for all genes

```
cohort_thresholds <- cohort_thresholds_raw %>%
  pivot_longer(c(-Gene, -cohort)) %>%
```

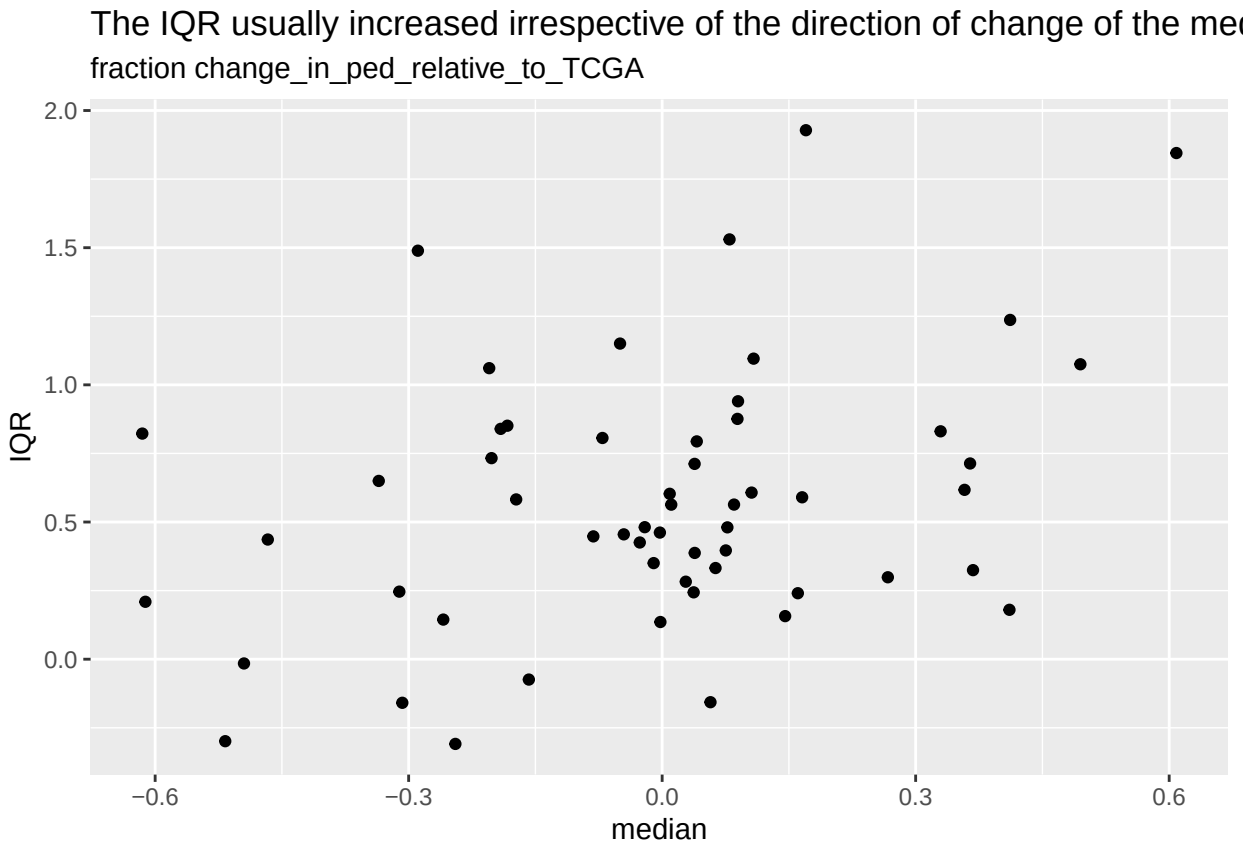
```
  pivot_wider(names_from = cohort, values_from = value) %>%
  mutate(change_in_ped_relative_to_TCGA =
              (PEDAYA - TCGA) / TCGA,
         change_in_treehouse_relative_to_TCGA =
            (Treehouse_pc - TCGA) / Treehouse_pc)
```

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")
```



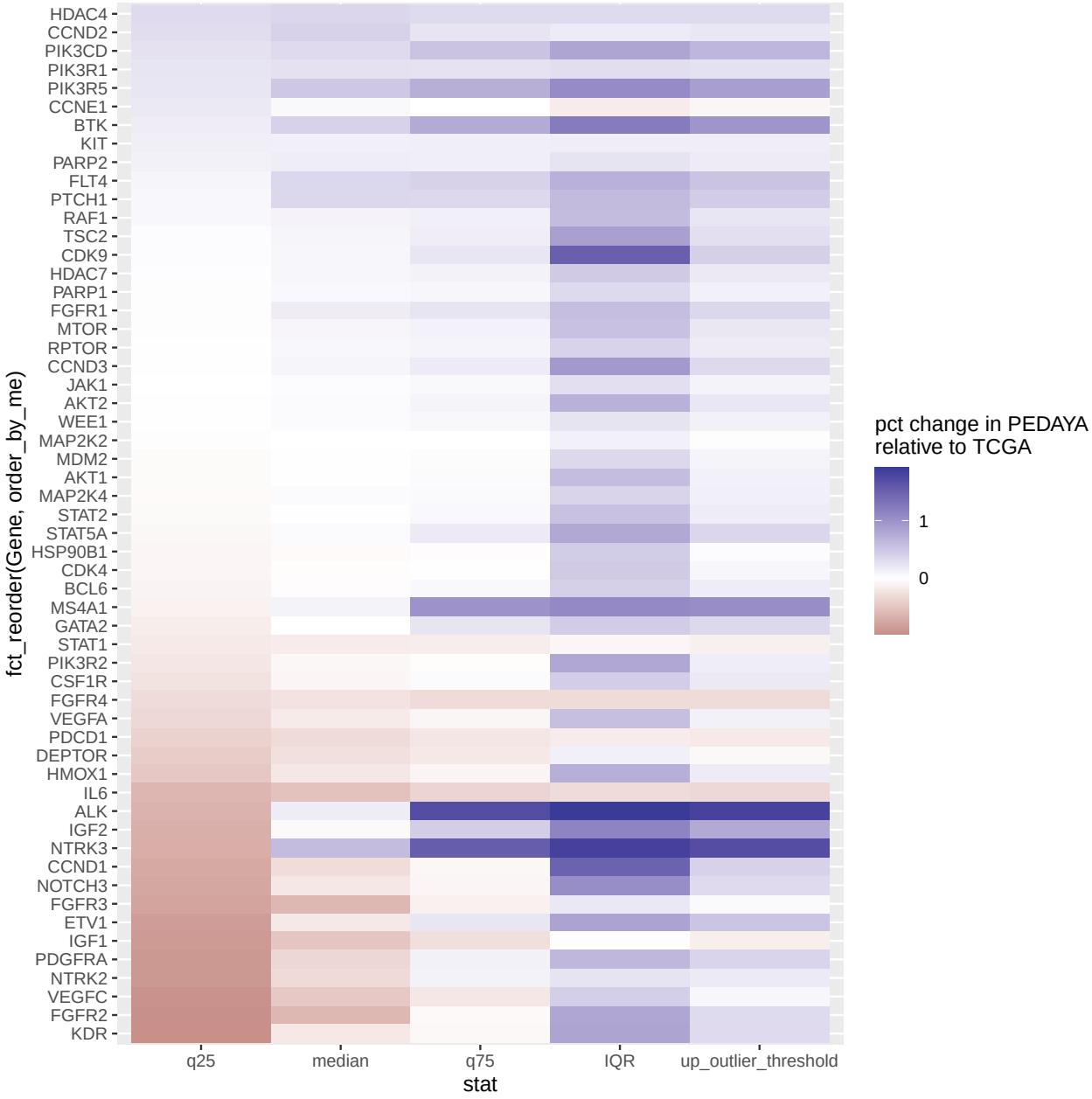

```
cohort_thresholds %>%
  filter(name %in% c("median", "IQR")) %>%
  select(Gene, name, change_in_ped_relative_to_TCGA) %>%
  pivot_wider(names_from = name,
              values_from = change_in_ped_relative_to_TCGA) %>%
  ggplot +
  geom_point(aes(x=median, y=IQR)) +
  ggtitle("The IQR usually increased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



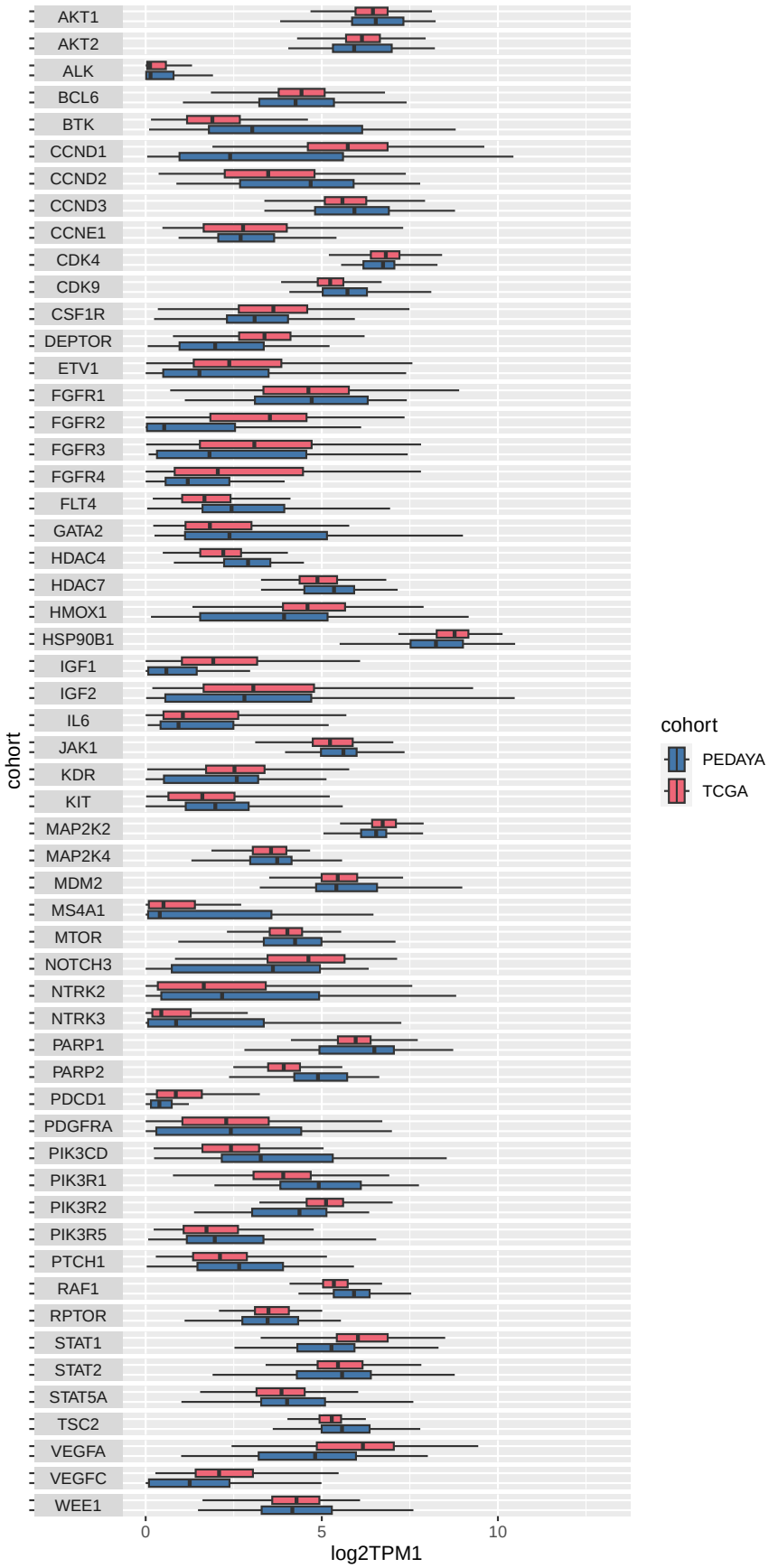

```
#geom_tile(aes(x=stat, y= fct_reorder(Gene, change_in_ped_relative_to_TCGA), fill = change_in_ped_relat
```

plot boxplots for TCGA and PEDAYA

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

```
TP_cohort_expr_subset <- TP_cohort_expr %>%
  slice_sample(n = 10000)
```

```r
ggplot(TP_cohort_expr_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
5
10
log2TPM1
cohort
PEDAYA
TCGA

```r
table(outliers$comparison_cohort)
```

```
##
##       PEDAYA         TCGA    TH03_TH34 Treehouse_pc Treehouse_pd
##           26           98           53           72           38
```

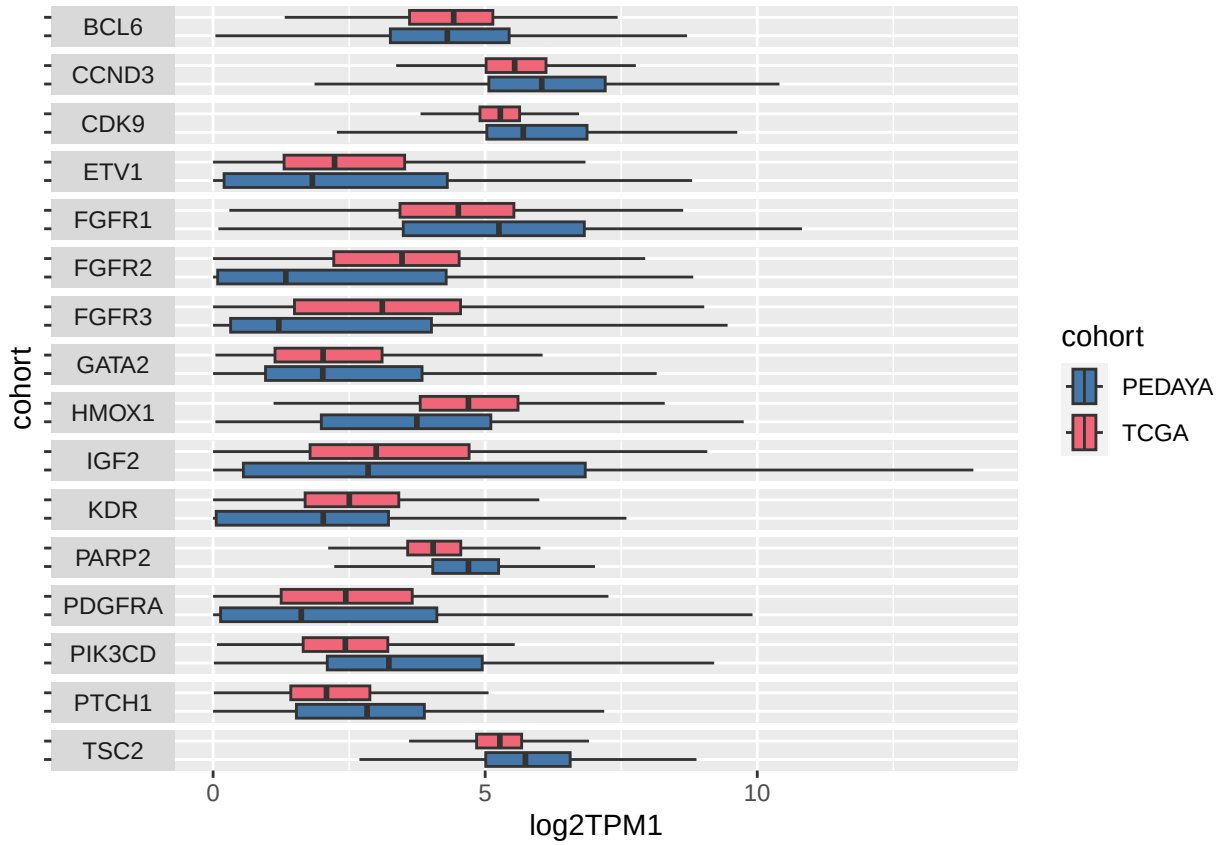
```r
TCGA_only_outliers <- outliers %>%
  group_by(gene, Sample_ID) %>%
  mutate(TCGA_only = "TCGA" %in% comparison_cohort &
           n() == 1) %>%
  filter(TCGA_only) %>%
  arrange(Sample_ID, gene)

n_distinct(TCGA_only_outliers$gene)
```

```
## [1] 16
```

```r
TP_cohort_expr_of_TCGA_only_outliers <- TP_cohort_expr %>%
         filter(Gene %in% TCGA_only_outliers$gene)

ggplot(TP_cohort_expr_of_TCGA_only_outliers) +
  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()
```



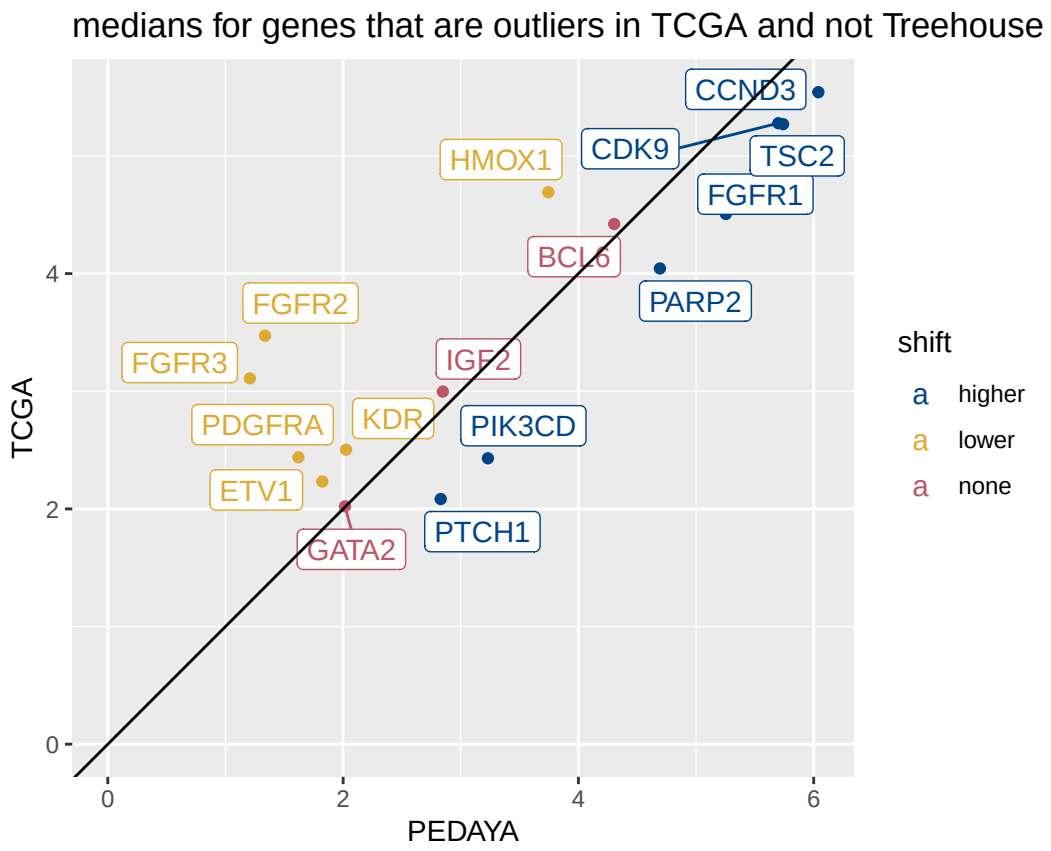

```
manual_annotation_of_pedaya_relative_to_TCGA <- tribble(
  ~Gene, ~IQR, ~shift,
  "BCL6", "wider", "none",
  "CCND1", "wider", "lower",
  "CCND3", "wider", "higher",
  "CDK9", "wider", "higher",
  "ETV1", "wider", "lower",
  "FGFR1", "wider", "higher",
  "FGFR2", "wider", "lower",
  "FGFR3", "wider", "lower",
  "GATA2", "wider", "none",
  "HMOX1", "wider", "lower",
  "IGF2", "wider", "none",
  "KDR", "wider", "lower",
  "MTOR", "similar", "higher",
  "PARP2", "similar", "higher",
  "PDGFRA", "wider", "lower",
  "PIK3CD", "wider", "higher",
  "PTCH1", "wider", "higher",
  "RPTOR", "similar", "higher",
  "TSC2", "wider", "higher",
  "VEGFA", "wider", "lower") %>%
  filter(Gene %in% TCGA_only_outliers$gene)

# cat(paste0("\"", unique(sort(TCGA_only_outliers$gene)), "\""), sep = ",\n")
```

```
anno_cohort_thresholds_of_TCGA_only_outliers <-
  left_join(cohort_thresholds_for_plot %>%
              filter(Gene %in% TCGA_only_outliers$gene),
            manual_annotation_of_pedaya_relative_to_TCGA)
```

```
## Joining, by = "Gene"
```

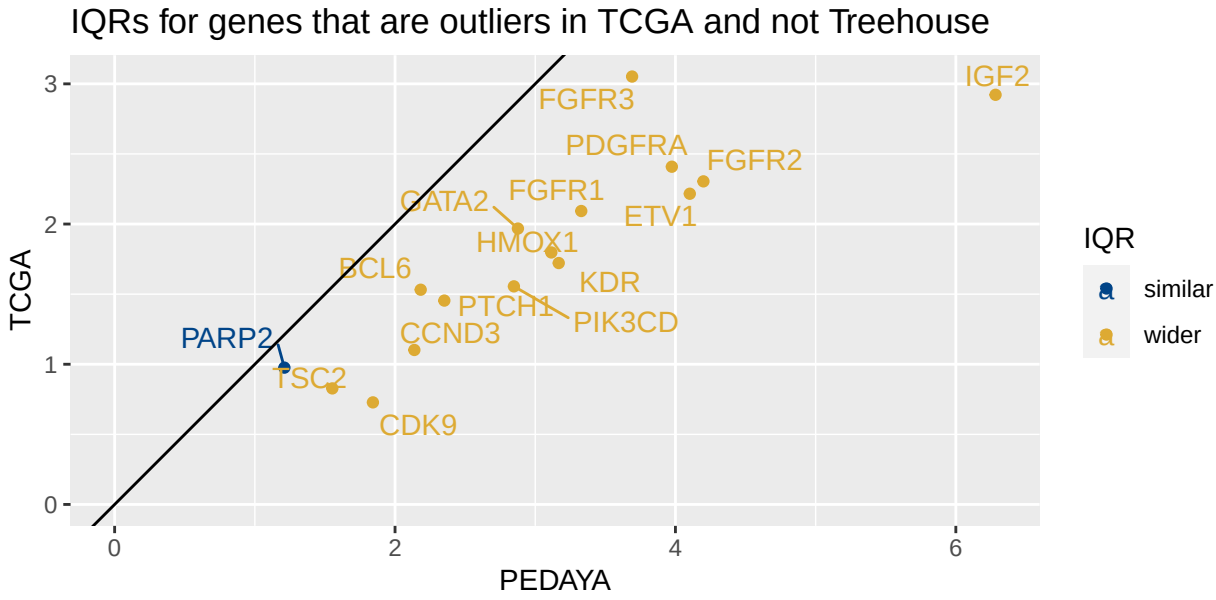
```
  ggplot(anno_cohort_thresholds_of_TCGA_only_outliers  %>%
              filter(stat == "median"),
         aes(x=PEDAYA, y=TCGA, color = shift)) +
  #geom_histogram(aes(x=change_in_ped_relative_to_TCGA))
  geom_point() +
  geom_label_repel(aes(label = Gene)) +
  geom_abline() +
  coord_equal() +
  #scale_y_continuous(breaks = c(0:3)) +
  expand_limits(y=0,x=0) +
  ggtitle("medians for genes that are outliers in TCGA and not Treehouse") +
  scale_color_highcontrast()
```




```
# cohort_thresholds_for_plot %>%
#   filter(Gene %in% TCGA_only_outliers$gene) %>%
#   filter(stat == "median") %>%
#   arrange(change_in_ped_relative_to_TCGA)
```

Compare IQRs

```
ggplot(anno_cohort_thresholds_of_TCGA_only_outliers  %>%
         filter(stat == "IQR"),
       aes(x=PEDAYA, y=TCGA, color = IQR)) +
  #geom_histogram(aes(x=change_in_ped_relative_to_TCGA))
  geom_point() +
  geom_text_repel(aes(label = Gene)) +
  geom_abline() +
  coord_equal() +
  scale_y_continuous(breaks = c(0:3)) +
  expand_limits(y=0,x=0) +
  ggtitle("IQRs for genes that are outliers in TCGA and not Treehouse") +
  scale_color_highcontrast()
```



Review data to identify a good quantitative IQR cutoff

```
anno_cohort_thresholds_of_TCGA_only_outliers %>%
  mutate(abs_change = PEDAYA-TCGA) %>%
  select(Gene, stat, PEDAYA, TCGA,
         pct_change = change_in_ped_relative_to_TCGA,
         abs_change,
         IQR, shift) %>%
  arrange(abs_change) %>% # change to pct_change to view alternative consideration
  adorn_pct_formatting(,,,pct_change) %>%
  filter(stat == "IQR") %>%
  kbl(digits = c(NA, NA, 1, 1, NA, 2, NA, NA)) %>%
  kable_styling(full_width = F)
```

```
# IQR change greater than 0.5 log2TPM1

# percent cutoff doesn't really work, because some relatively small pcts are large in log2tpm1 space,
```

Gene	stat	PEDAYA	TCGA	pct_change	abs_change	IQR	shift
PARP2	IQR	1.2	1.0	24.1%	0.23	similar	higher
FGFR3	IQR	3.7	3.1	20.9%	0.64	wider	lower
BCL6	IQR	2.2	1.5	42.5%	0.65	wider	none
TSC2	IQR	1.6	0.8	87.6%	0.73	wider	higher
PTCH1	IQR	2.4	1.5	61.7%	0.90	wider	higher
GATA2	IQR	2.9	2.0	46.2%	0.91	wider	none
CCND3	IQR	2.1	1.1	94.0%	1.04	wider	higher
CDK9	IQR	1.8	0.7	153.0%	1.11	wider	higher
FGFR1	IQR	3.3	2.1	59.0%	1.24	wider	higher
PIK3CD	IQR	2.8	1.6	83.1%	1.29	wider	higher
HMOX1	IQR	3.1	1.8	73.3%	1.32	wider	lower
KDR	IQR	3.2	1.7	84.0%	1.45	wider	lower
PDGFRA	IQR	4.0	2.4	65.0%	1.57	wider	lower
ETV1	IQR	4.1	2.2	85.1%	1.89	wider	lower
FGFR2	IQR	4.2	2.3	82.2%	1.90	wider	lower
IGF2	IQR	6.3	2.9	115.0%	3.36	wider	none

```
# e.g. FGFR3, BCL6, wider
```

Review data to see identify a good quantitative shift cutoff

```
anno_cohort_thresholds_of_TCGA_only_outliers %>%
  mutate(abs_change = PEDAYA-TCGA) %>%
  select(Gene, stat, PEDAYA, TCGA,
         pct_change = change_in_ped_relative_to_TCGA,
         abs_change,
         IQR, shift) %>%
  arrange(abs_change) %>%
  adorn_pct_formatting(,,,pct_change) %>%
  filter(stat == "median") %>%
  kbl(digits = c(NA, NA, 1, 1, NA, 2, NA, NA)) %>%
  kable_styling(full_width = F)
```

```
# if PEDAYA median is 0.25 higher or lower than TCGA median, the shift is higher or lower, respectively

# percent cutoff doesn't really work, because some relatively small pcts are large in log2tpm1 space,
```

Summary of differences

```
#manual_annotation_of_pedaya_relative_to_TCGA

tabyl(manual_annotation_of_pedaya_relative_to_TCGA, IQR)
```

```
##      IQR  n percent
##  similar  1  0.0625
##    wider 15  0.9375
```

Gene	stat	PEDAYA	TCGA	pct_change	abs_change	IQR	shift
FGFR2	median	1.3	3.5	-61.5%	-2.14	wider	lower
FGFR3	median	1.2	3.1	-61.2%	-1.90	wider	lower
HMOX1	median	3.7	4.7	-20.2%	-0.95	wider	lower
PDGFRA	median	1.6	2.4	-33.5%	-0.82	wider	lower
KDR	median	2.0	2.5	-19.1%	-0.48	wider	lower
ETV1	median	1.8	2.2	-18.3%	-0.41	wider	lower
IGF2	median	2.8	3.0	-5.0%	-0.15	wider	none
BCL6	median	4.3	4.4	-2.6%	-0.12	wider	none
GATA2	median	2.0	2.0	-0.3%	-0.01	wider	none
CDK9	median	5.7	5.3	8.0%	0.42	wider	higher
TSC2	median	5.7	5.3	8.9%	0.47	wider	higher
CCND3	median	6.0	5.5	9.0%	0.50	wider	higher
PARP2	median	4.7	4.0	16.1%	0.65	similar	higher
PTCH1	median	2.8	2.1	35.8%	0.75	wider	higher
FGFR1	median	5.3	4.5	16.6%	0.75	wider	higher
PIK3CD	median	3.2	2.4	33.0%	0.80	wider	higher

```
tabyl(manual_annotation_of_pedaya_relative_to_TCGA, shift)
```

```
##   shift n percent
##  higher 7  0.4375
##   lower 6  0.3750
##    none 3  0.1875
```

Text summary

Each sample in the study was compared to TCGA, PEDAYA, Stanford and the full Treehouse compendium. Often genes would be identified as outliers with respect to multiple cohorts. We investigated the exceptions, genes found as outliers with respect to only one cohort. The cohort with the largest number of uniquely detected outliers was TCGA, We analyzed the distribution of expression for the outlier genes that were identified in any samples by comparison with TCGA and no other cohorts. Because the addition of pediatric samples is the primary difference between TCGA and the other cohorts, we focussed on comparing the PEDAYA cohort and TCGA

Of those 16, 15 had wider distributions (>0.5log2(TPM+1) bigger) in PEDAYA than TCGA, while one was similar. For 7, the median was higher (by more than 0.25 log2TPM1) in PEDAYA than TCGA. For 6, the median was lower (by more than 0.25 log2TPM1) in PEDAYA than TCGA. For 3, the medians were similar in PEDAYA and TCGA.

The Treehouse compendium is 77% TCGA (which is 96% adult). The remaining 23% of the Treehouse compendium is 97% PEDAYA (<=30). In the genes we looked at, the changes in distribution between TCGA and Treehouse compendium were consistent with the effects of adding the distribution of PEDAYA samples. The differences in the treehouse compendium from TCGA is mostly due to the addition of PEDAYA samples.