BMI_stratification

Carlos Blázquez Bondia

2023-03-06

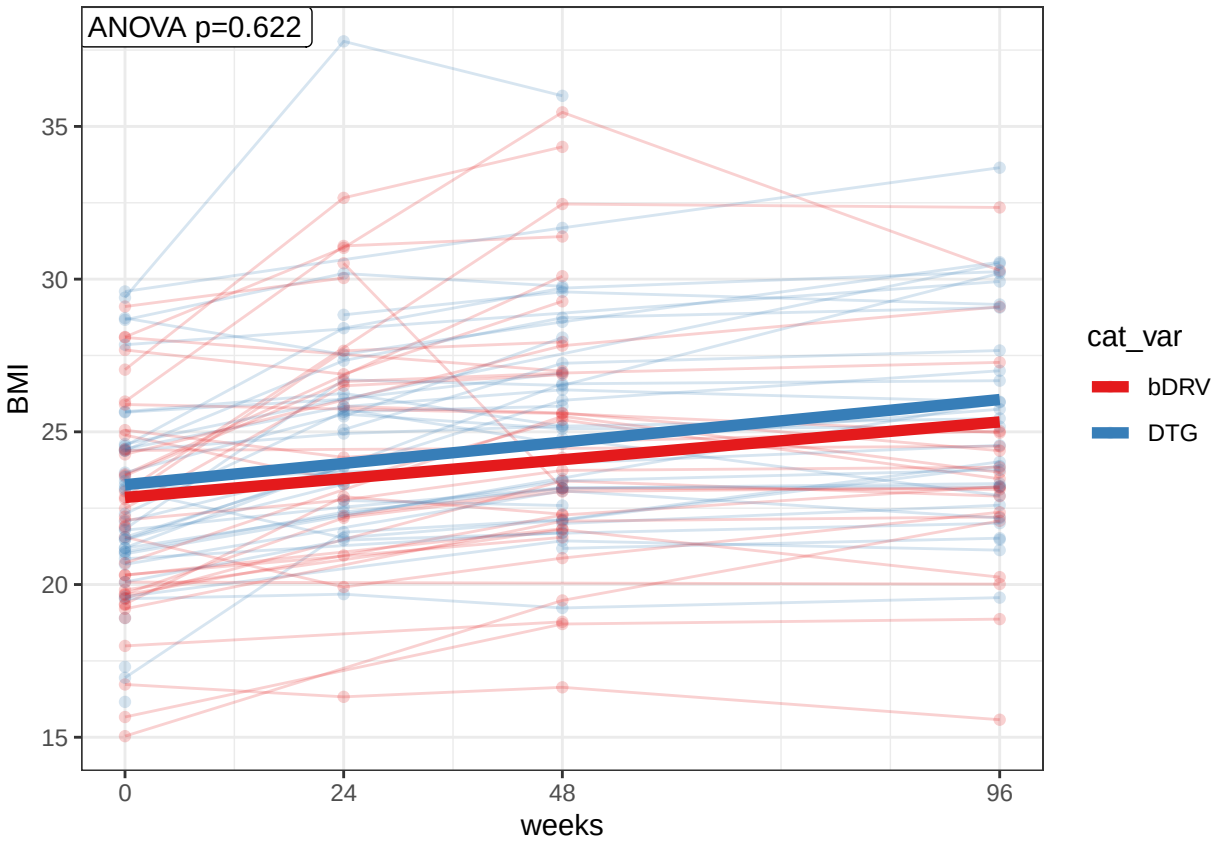
Stratification by BMI

This set of tests comes from recent studies which found that INSTIs are associated with greater BMI increases, respective to other treatments in ART-receiving patients. BMI is highly related to gut health, metabolism and microbiome, and may be a possible confussor regarding the interaction between treatment and microbiote, especially gene richness, as it may be masking the actual effect of treatment. The first is to assess how big of an impact our treatments have on the patients BMIs:

```
igc_df %>%
  dplyr::select(SampleID, link_var, cat_var, long_var, BMI) %>%
  get_lmm_effects(.,cat_vector = "cat_var", num_vector = "BMI", long_var = "long_var", link_var = "link
```

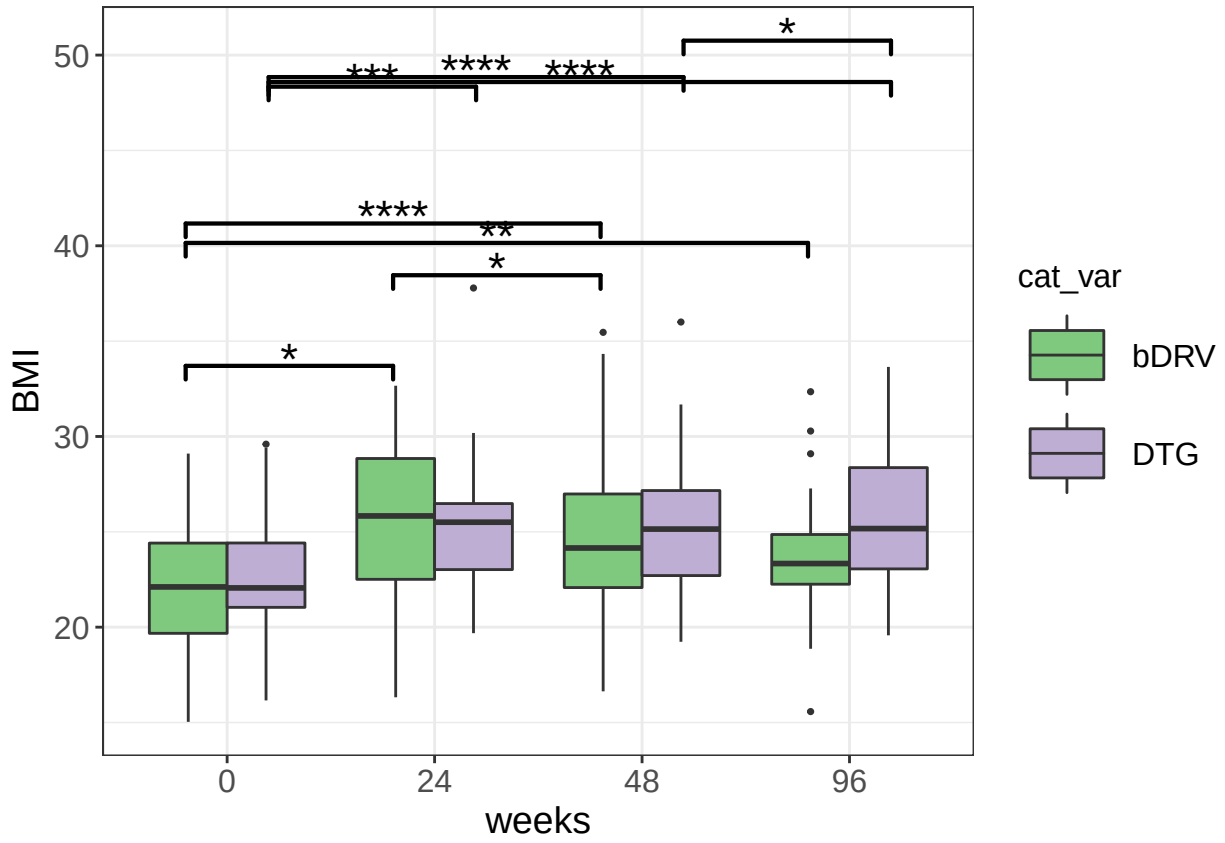
```
## [1] "cat_var"
## [1] "BMI"
```

```
## $cat_var
## $cat_var$BMI
## $cat_var$BMI$plot
```




```
##
## $cat_var$BMI$summary
##                          estimate      st.err       df       tval            p
## (Intercept)         22.843254724 0.571780322 101.7421 39.9511033 5.318804e-64
## long_var             0.025897438 0.004761211 161.2773  5.4392542 1.951105e-07
## cat_varDTG           0.432868903 0.792613839 102.4527  0.5461284 5.861650e-01
## long_var:cat_varDTG  0.003153891 0.006389171 161.2107  0.4936307 6.222388e-01
```

```
require(rstatix)
bmi_test <-
get_comp_boxplots(dat = igc_df, cat_var = "cat_var", num_var = "BMI", long_var = "long_var", link_var =


bmi_test$plot
```




```
bmi_test$test$cat %>%
  dplyr::select(long_var, contains ( "group"), p.adj.signif) %>%
  kableExtra::kable(format = "markdown")
```

long_var	group1	group2	groups	p.adj.signif
0	bDRV	DTG	bDRV, DTG	ns
24	bDRV	DTG	bDRV, DTG	ns
48	bDRV	DTG	bDRV, DTG	ns
96	bDRV	DTG	bDRV, DTG	ns

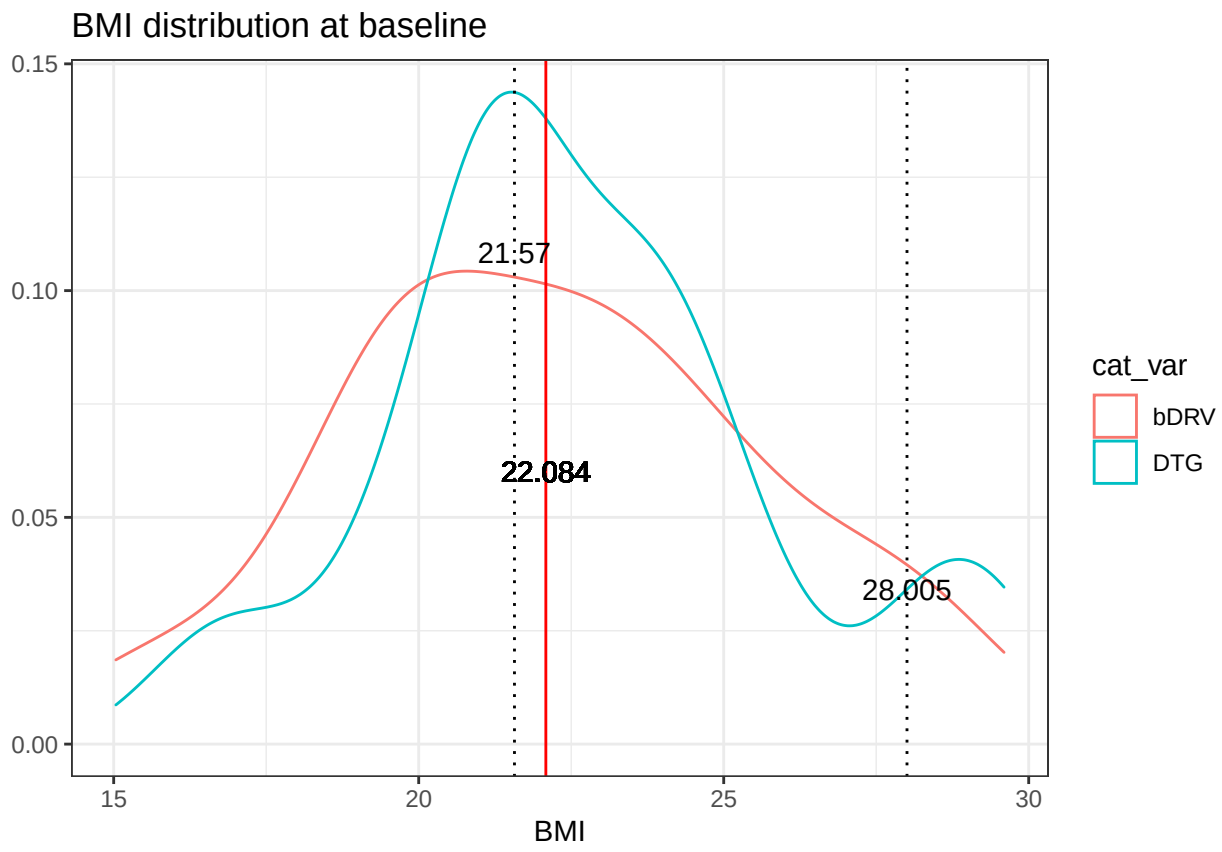
Turns out looking at BMI alone, no categorical differences can be found. While both treatment groups significantly increase in BMI, no diferences between DTG an DRV are found.

Anyhow, this may have something to do with BMI differences at BL. Let's check for its distribution:

```
peaks_bl <- igc_df %>%
  dplyr::filter(long_var == 0, !is.na(BMI)) %>%
  dplyr::pull(BMI) %>%
  density() %>%
  peak_finder(goal = "max")

median_bl <- igc_df %>%
  dplyr::filter(long_var == 0, !is.na(BMI)) %>%
  pull(BMI) %>%
```

```
  median()

igc_df %>%
  dplyr::filter(long_var == 0, !is.na(BMI)) %>%
  ggplot(.) +
  # geom_histogram(aes(x = BMI), fill = "grey", bins = 100) +
  geom_density(aes(x = BMI, color = cat_var),bins = 100) +
  theme_bw() +
  geom_vline(xintercept = peaks_bl$x, lty = 3) +
  geom_vline(xintercept = median_bl, col = "red") +
  geom_text(data = peaks_bl, aes(x = x, y = y*.9, label = round(x, 3))) +
  geom_text(aes(x = median_bl, y = max(peaks_bl$y)*.5, label = round(median_bl,3))) +
  labs(x = "BMI", y = "kernel density", title = "BMI distribution at baseline") +
  theme(axis.title.y = element_blank())
```

```
## Warning: Ignoring unknown parameters: bins
```




```
igc_df %>%
  dplyr::filter(long_var == 0, !is.na(BMI)) %>%
  dplyr::mutate(BMI_c = case_when(BMI >= 22 ~ "high",
                                  BMI < 22 ~ "low")) %>%
  group_by(BMI_c) %>%
  tally() %>%
  kableExtra::kable(format="markdown")
```

BMI_c	n
high	40
low	38

It seems there is a certain bimodality in the distribution of BMI at BL in the DTG. However, still seems to be close to normal, so the max peak of 21.57 is fairly close to the median of 22.084. From this it 22 seems a good cut-off point between high and low BMI.

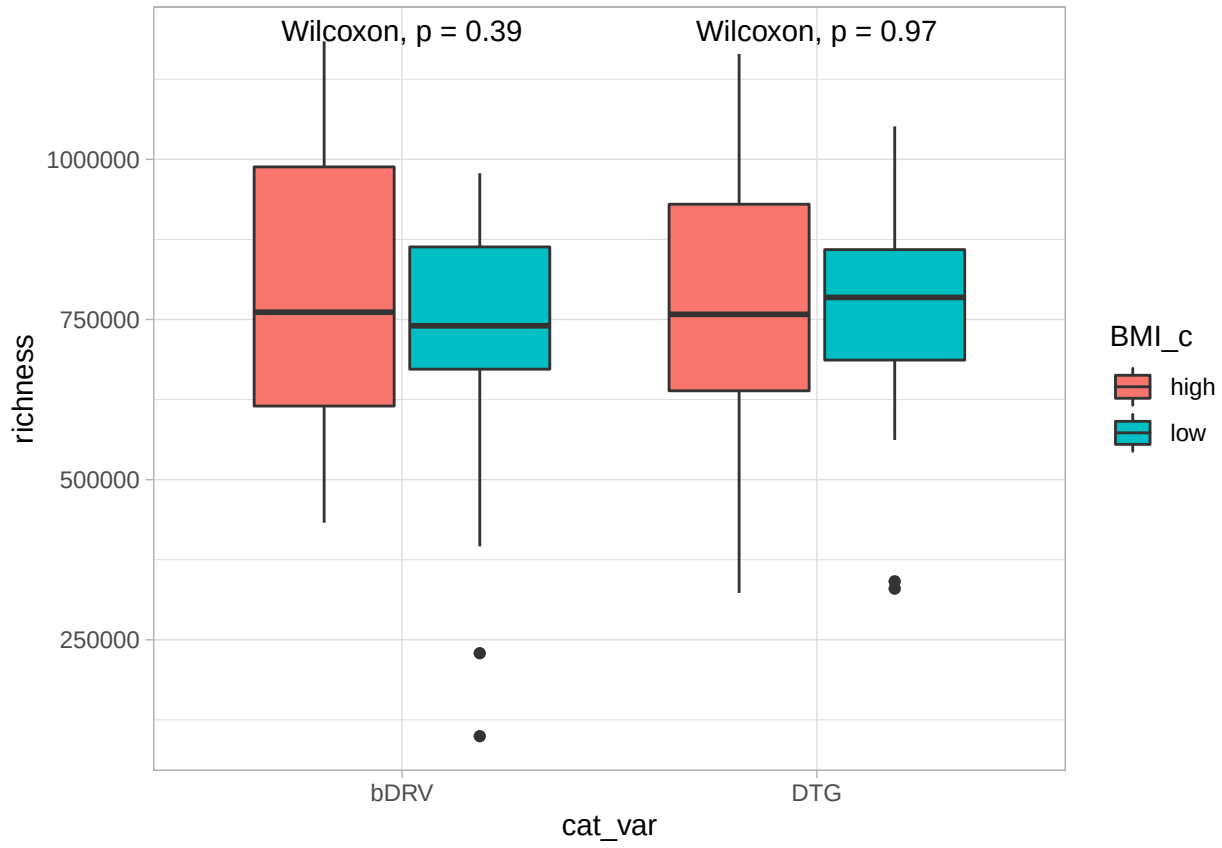
```
bmi_c <-
igc_df %>%
  dplyr::filter(long_var == 0, !is.na(BMI)) %>%
  dplyr::select(link_var, long_var, cat_var, BMI, richness) %>%
  mutate(cluster = kmeans(BMI, centers = 2)$cluster)
```

A kmeans clustering approach splits both groups at around BMI = 22.5. This should further reiforce the 22 threshold.

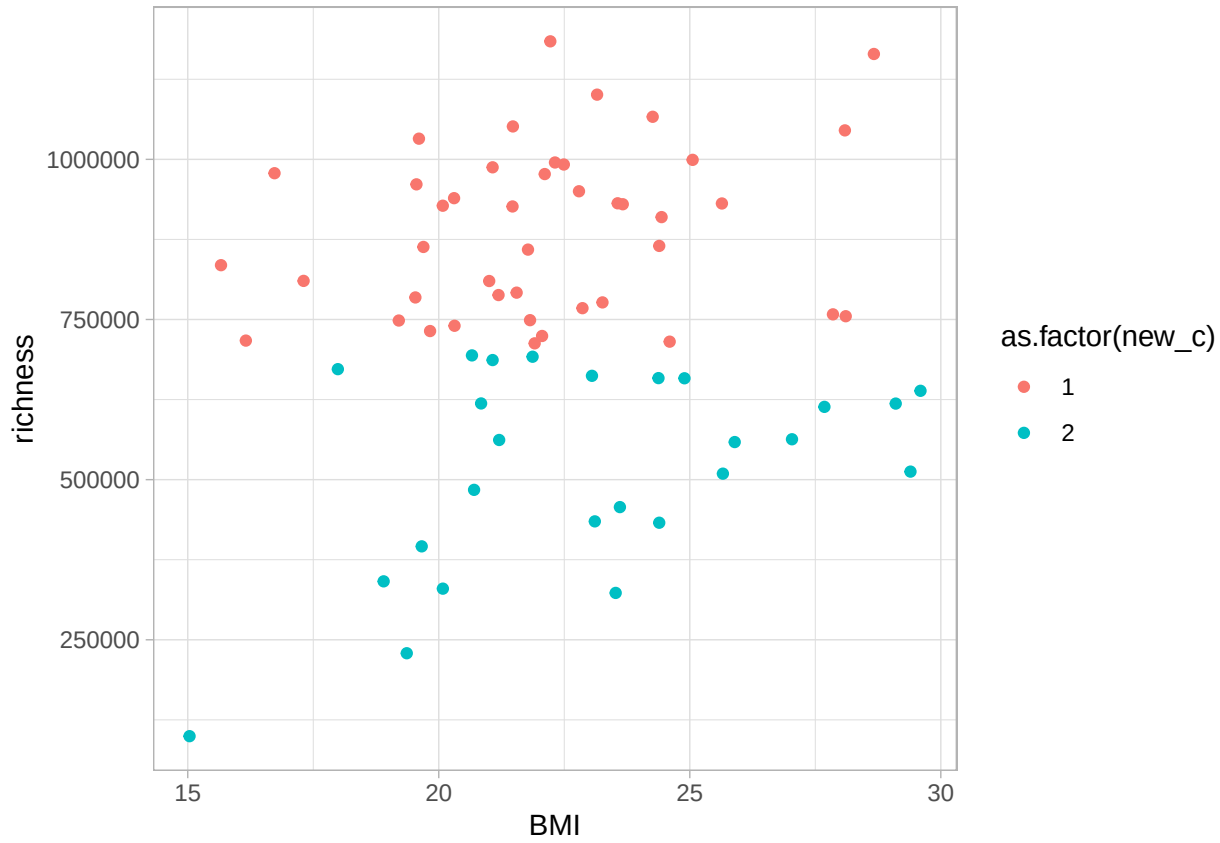
Stratification by BMI at baseline

Now we found a threshold appropiate to split by BMI, we can stratify the population by treatment and bmi, vs gene richness.

```
bmi_c %>%
  dplyr::mutate(BMI_c = case_when(
    BMI >= 22 ~ "high",
    BMI < 22 ~ "low"
  )) %>%

  # dplyr::mutate(group_bmi = paste(cat_var, BMI_c)) %>%
  ggplot(aes(x = cat_var, y = richness, fill = BMI_c)) +
  geom_boxplot() +
  ggpubr::stat_compare_means(method = "wilcox.test", hide.ns = F) +
  theme_light()
```




```
bmi_c %>%
  # dplyr::select(BMI, richness) %>%
  dplyr::filter(!is.na(richness)) %>%
  mutate(new_c = kmeans(.[,c("BMI","richness")], centers = 2)$cluster) %>%
  ggplot(aes( x = BMI, y = richness, color = as.factor(new_c))) +
  geom_point() +
  theme_light()
```



Doesn't seem to be any noticeable difference between groups or BMI types. No diferences richness between HBMI and LBMI could be found.

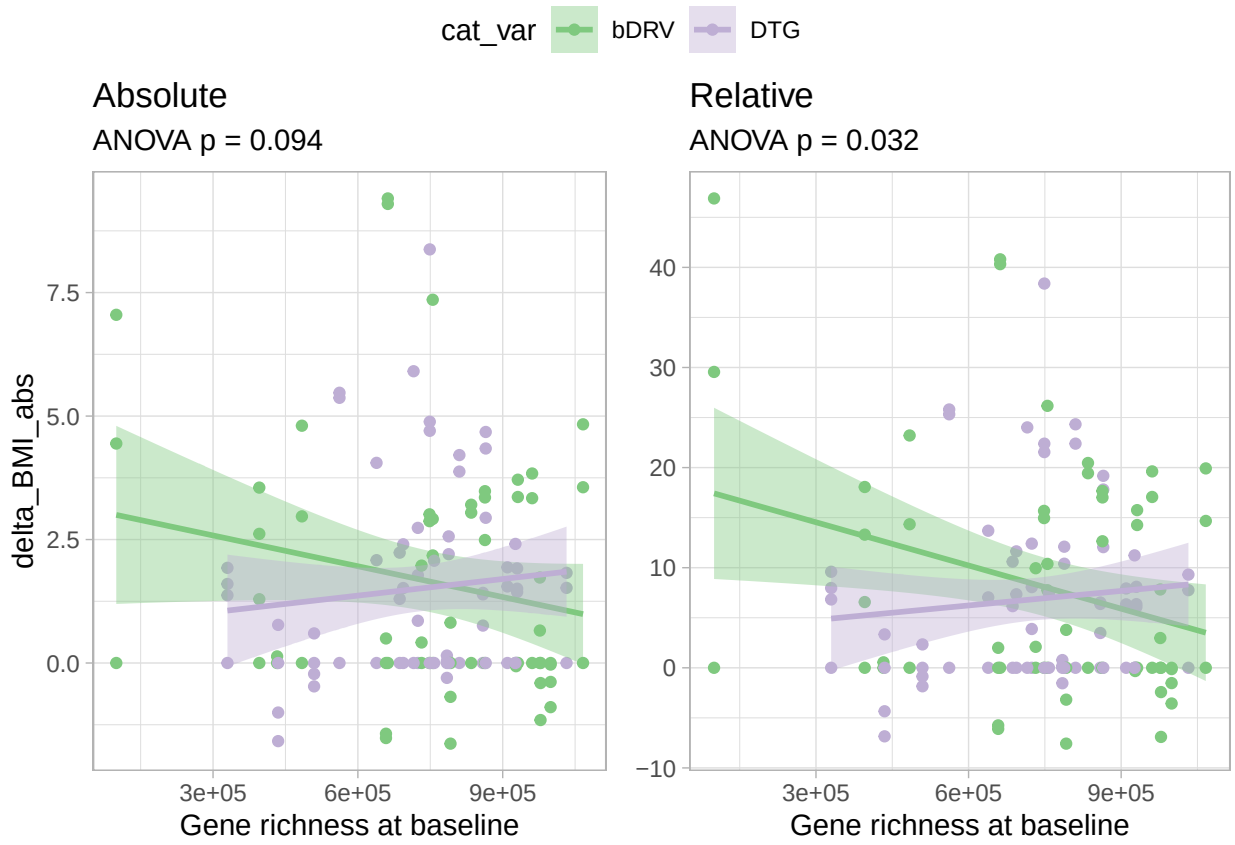
Stratification by BMI changes

Maybe the increase in gene richness may be associated with higher increase in BMI

vs GR BL

```
delta_bmi_df <-
    igc_df %>%
      dplyr::filter(!is.na(richness), !is.na(BMI)) %>%
      group_by(link_var) %>%
  dplyr::filter(any(long_var == 0),
                any(long_var == 96)) %>%
      mutate(bmi0 = BMI[long_var == 0],
          gr0 = richness[long_var == 0],
    delta_BMI_abs = BMI - BMI[long_var == 0],
          delta_BMI_rel = (BMI - BMI[long_var == 0])*100 / BMI[long_var == 0],
          delta_gr_abs = richness - richness[long_var == 0],
          delta_gr_rel = (richness - richness[long_var == 0]) / (richness + richness[long_var == 0])) %>%
  dplyr::mutate(bmi_c = case_when(
    delta_BMI_abs[long_var == 96] >= 2.5 ~ "H_BMI",
```

```
    delta_BMI_abs[long_var == 96]  < 2.5 ~ "L_BMI"
  )) %>%
  dplyr::mutate(bmi_c_g = paste(bmi_c, cat_var, sep = "_"))


test_gr0_abs <-
  delta_bmi_df %>%
  glm(delta_BMI_abs ~ gr0 * cat_var, data = .) %>%
  rstatix::anova_test() %>%
  as.tibble() %>%
  dplyr::slice(3L) %>%
  pull(p)

deltas_gr0_abs <-
  delta_bmi_df %>%
  ggplot(aes(x = gr0, y = delta_BMI_abs, color = cat_var)) +
  geom_point() +
  geom_smooth(method = "glm", aes(fill = cat_var)) +
  labs(title = "Absolute", subtitle = paste("ANOVA p =", test_gr0_abs),x = "Gene richness at baseline")
  theme_light() +
  scale_color_brewer(palette = "Accent") +
  scale_fill_brewer(palette = "Accent")


test_gr0_rel <-
  delta_bmi_df %>%
  glm(delta_BMI_rel ~ gr0 * cat_var, data = .) %>%
  rstatix::anova_test() %>%
  as.tibble() %>%
  dplyr::slice(3) %>%
  pull(p)

deltas_gr0_rel <-
  delta_bmi_df %>%
  ggplot(aes(x = gr0, y = delta_BMI_rel, color = cat_var)) +
  geom_point() +
  geom_smooth(method = "glm", aes(fill = cat_var)) +
  theme_light() +
  labs(title = "Relative", subtitle = paste("ANOVA p =", test_gr0_rel),x = "Gene richness at baseline")
  theme(axis.title.y = element_blank()) +
  scale_color_brewer(palette = "Accent") +
  scale_fill_brewer(palette = "Accent")



ggpubr::ggarrange(deltas_gr0_abs, deltas_gr0_rel, common.legend = T)
```

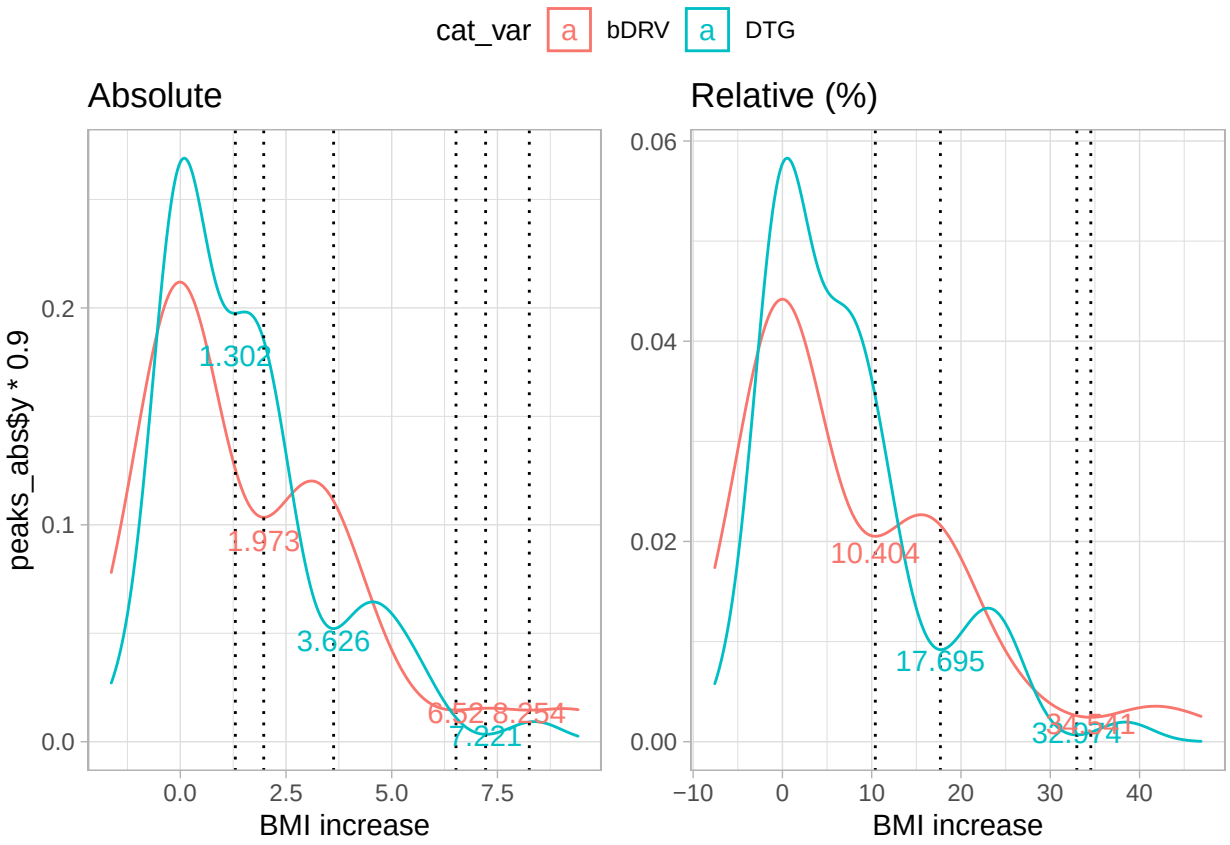


No correlation could be found relating

vs GR change

```
tp <- 96



peaks_abs <-
  unique(metadata$group) %>%
  purrr::set_names() %>%
  purrr::map_dfr(function(treat) {

    delta_bmi_df %>%
      dplyr::filter(cat_var == treat) %>%
      dplyr::pull(delta_BMI_abs) %>%
      density() %>%
      peak_finder(goal = "min") %>%
      round(., 3) %>%
      mutate(cat_var = treat)
  })


plot_delta_abs <-
```

```r
delta_bmi_df %>%
  ggplot(aes(x = delta_BMI_abs)) +
  geom_density(aes(color = cat_var)) +
  geom_vline(aes(color = cat_var), xintercept = peaks_abs$x, lty = 3) +
  geom_text(data = peaks_abs, aes(x = peaks_abs$x, y = peaks_abs$y*.9,label = peaks_abs$x, color = cat_
  theme_light() +
  labs(title = "Absolute", x = "BMI increase")




peaks_rel <-
  unique(metadata$group) %>%
  purrr::set_names() %>%
  purrr::map_dfr(function(treat) {

    delta_bmi_df %>%
      dplyr::filter(cat_var == treat) %>%
      dplyr::pull(delta_BMI_rel) %>%
      density() %>%
      peak_finder(goal = "min") %>%
      round(., 3) %>%
      mutate(cat_var = treat)
  })

plot_delta_rel <-
delta_bmi_df %>%
  ggplot(aes(x = delta_BMI_rel,color = cat_var)) +
  geom_density(aes(color = cat_var)) +
  geom_vline(data = peaks_rel, aes(color = cat_var), xintercept = peaks_rel$x, lty = 3) +
  geom_text(data = peaks_rel, aes(x = x, y = y*.9,label = x, color = cat_var)) +
  labs(title = "Relative (%)", x = "BMI increase") +
  theme_light()+
  theme(axis.title.y = element_blank())


# delta_bmi_df %>%



ggpubr::ggarrange(plot_delta_abs, plot_delta_rel, common.legend = T)
```

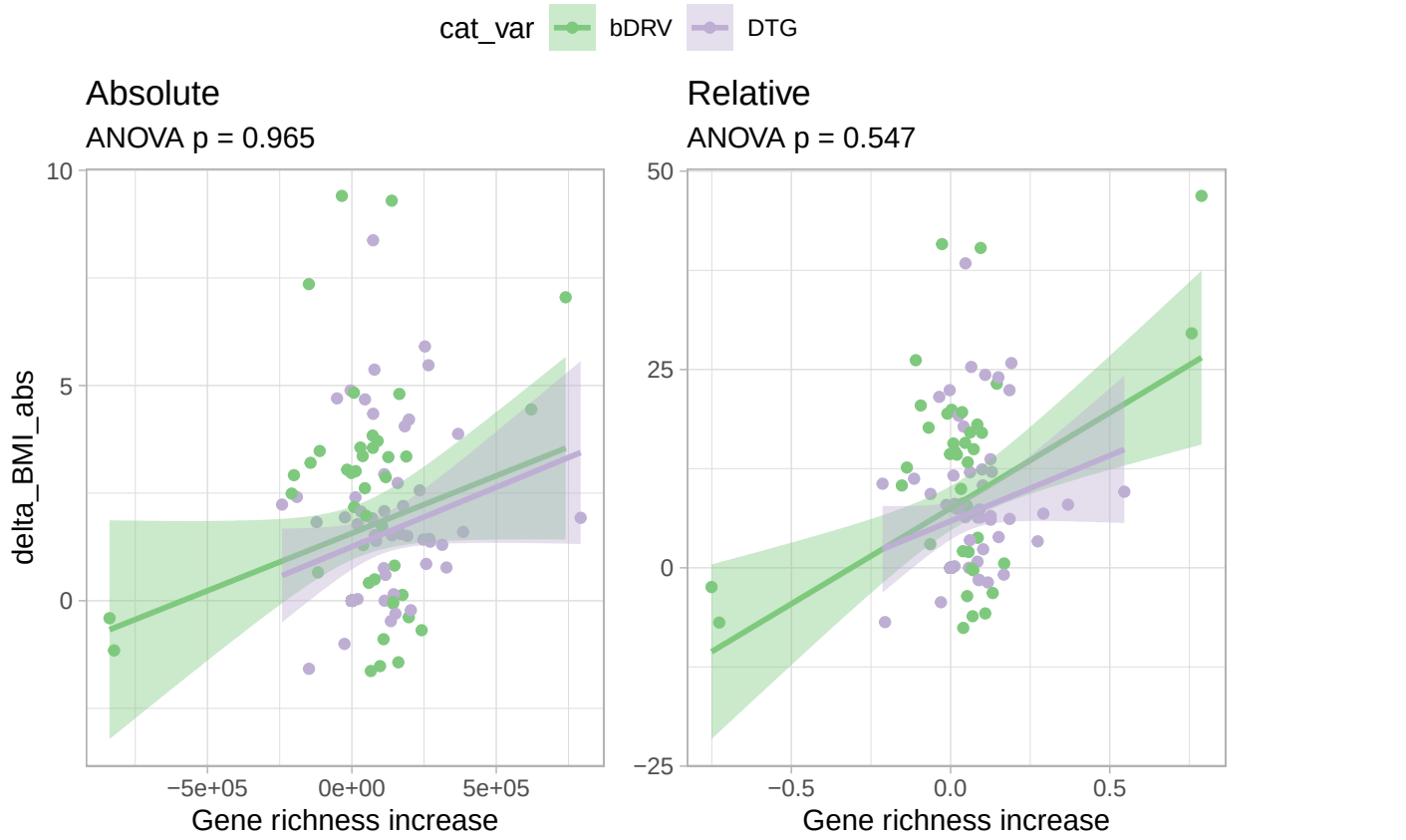


Both treatment groups appear to have a unimodal distribution regarding their BMI increase at week 96, both in absolute and relative terms. However there seems to be a shift between both, as the "valley" between both peaks in the DTG group seems to be shifted toward the left, relative to the DRV group.

A good threshold to separate two groups seems to be 2.5 increase of absolute BMI. It separates both peaks from each group, more or less equitatively.

```
test_gr_abs <-
  delta_bmi_df %>%
  glm(delta_BMI_abs ~ delta_gr_abs * cat_var, data = .) %>%
  rstatix::anova_test() %>%
  as.tibble() %>%
  dplyr::slice(3L) %>%
  pull(p)

deltas_gr_abs <-
  delta_bmi_df %>%
  ggplot(aes(x = delta_gr_abs, y = delta_BMI_abs, color = cat_var)) +
  geom_point() +
  geom_smooth(method = "glm", aes(fill = cat_var)) +
  labs(title = "Absolute", subtitle = paste("ANOVA p =", test_gr_abs), x = "Gene richness increase") +
  theme_light() +
  scale_color_brewer(palette = "Accent") +
  scale_fill_brewer(palette = "Accent")


test_gr_rel <-
```

```
  delta_bmi_df %>%
  glm(delta_BMI_rel ~ delta_gr_rel * cat_var, data = .) %>%
  rstatix::anova_test() %>%
  as.tibble() %>%
  dplyr::slice(3) %>%
  pull(p)

deltas_gr_rel <-
  delta_bmi_df %>%
  ggplot(aes(x = delta_gr_rel, y = delta_BMI_rel, color = cat_var)) +
  geom_point() +
  geom_smooth(method = "glm", aes(fill = cat_var)) +
  theme_light() +
  labs(title = "Relative", subtitle = paste("ANOVA p =", test_gr_rel),x = "Gene richness increase") +
  theme(axis.title.y = element_blank()) +
  scale_color_brewer(palette = "Accent") +
  scale_fill_brewer(palette = "Accent")



ggpubr::ggarrange(deltas_gr_abs, deltas_gr_rel, common.legend = T)
```



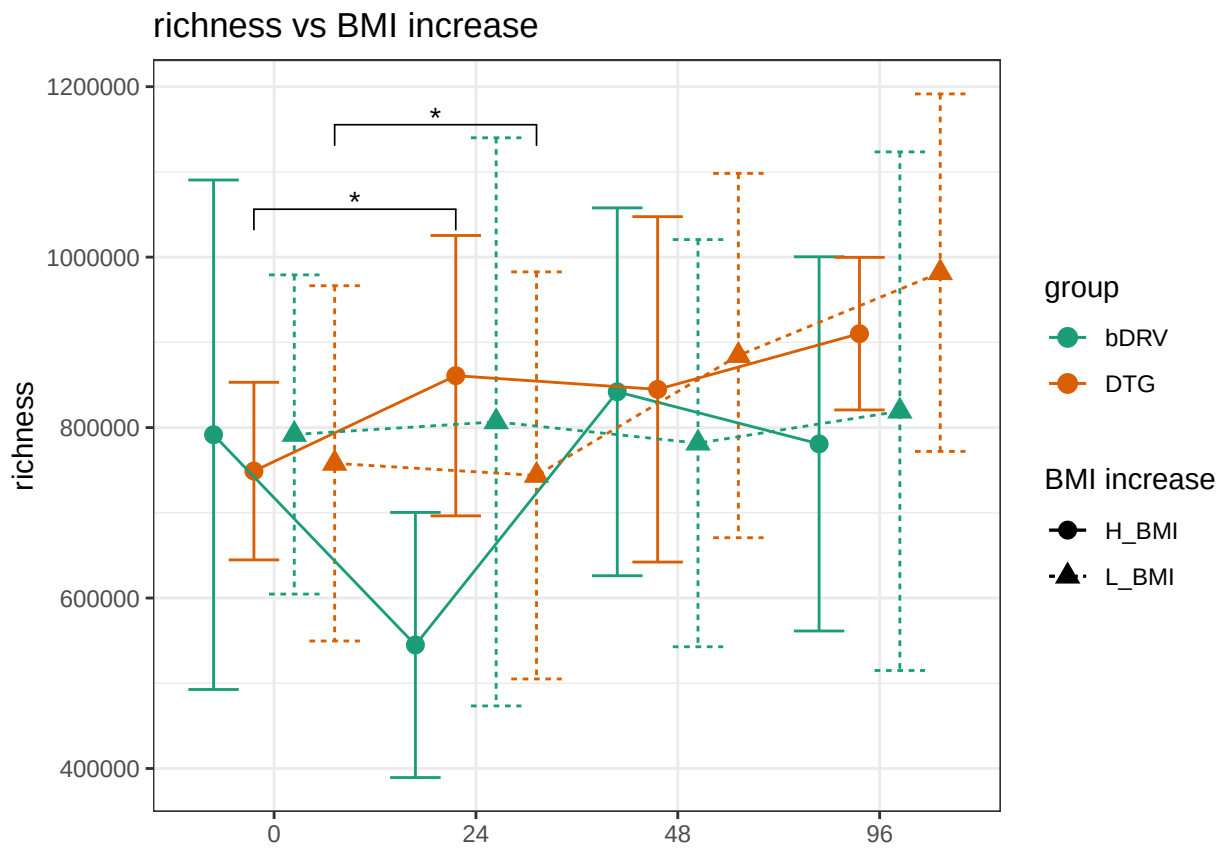
Stratifying population by BMI

We know from previous analysis that the biggest differences in GR beween grups happened at week 24: ## Week 24

```
cat_test <-
    obj <-
  delta_bmi_df %>%
  dplyr::filter(!is.na(richness)) %>%
  group_by(long_var) %>%
  get_group_comparisons(
    .,
    link_var = "link_var",
    long_var = "long_var",
    cat_vector = "bmi_c_g",
    num_vector = "richness",
    type = "categorical",
    comps = unique(delta_bmi_df$bmi_c_g),
    graph_coords = T
  )

objtest <- obj %>% pluck("bmi_c_g", "richness", "test")
objstats <- obj %>% pluck("bmi_c_g", "richness", "stats") %>%
  mutate(stat = paste(round(median, 2), " [", round(iqr, 2), "]", sep = ""))

cat_test_df <-
objtest %>%
  pivot_longer(cols = c("group1", "group2")) %>%
  dplyr::rename(cat_var = value) %>%
  dplyr::left_join(objstats[, c("cat_var", "long_var", "stat")], by = c("cat_var", "long_var")) %>%
  pivot_wider(names_from = "name",
              values_from = c("cat_var", "stat")) %>%
  dplyr::select(long_var,
                group1 = cat_var_group1,
                group2 = cat_var_group2,
                contains("stat"),
                y.position,
                xmin,xmax,
                p.adj,
                p.adj.signif)


long_test <-
combn(unique(delta_bmi_df$long_var), 2, simplify = F) %>%
  purrr::set_names() %>%
  purrr::map_dfr(function(comb){

    obj <-
    delta_bmi_df %>%
      dplyr::filter(!is.na(richness),
                    long_var %in% comb) %>%
      # group_by(bmi_c_g) %>%
      get_group_comparisons(.,link_var = "link_var", long_var = "long_var",cat_vector = "bmi_c_g",num_v
```

```
    obj %>%
      pluck("bmi_c_g","richness","test") %>%
      rstatix::adjust_pvalue( method = "BH", p.col = "p") %>%
      full_join(obj$bmi_c_g[["richness"]]$stats, by = "cat_var")%>%
      dplyr::mutate(.y. = "richness") %>%
      dplyr::select(cat_var, group1, group2,num_var = .y.,y.position, xmin, xmax, contains("median"), c
  })%>%
      rstatix::add_significance()



summ_df <-
delta_bmi_df %>%
  dplyr::group_by(long_var,bmi_c, cat_var, bmi_c_g) %>%
  summarise(median_gr = median(richness),
            sd_gr = sd(richness))




spaghetti_plot_list <-
    igc_df %>%
      # dplyr::filter(!is.na(!!sym(nv)), !is.na(bmi_c_g)) %>%
      mutate(long_var = as.factor(long_var)) %>%
      ggplot(aes(
        x = long_var,
        y = richness,
        color = bmi_c_g
      )) +
      geom_point(
        data = summ_df,
        aes(
          x = as.factor(long_var),
          y = median_gr,
          color = cat_var,
          shape = bmi_c,
          group = bmi_c_g
        ),
        size = 3,
        position = position_dodge(.8)
      ) +

      geom_errorbar(
        data = summ_df,
        aes(
          x = as.factor(long_var),
          ymin = median_gr - sd_gr,
          ymax = median_gr + sd_gr,
          color = cat_var,
          lty = bmi_c,
          group = bmi_c_g
        ),
        inherit.aes = F,
```

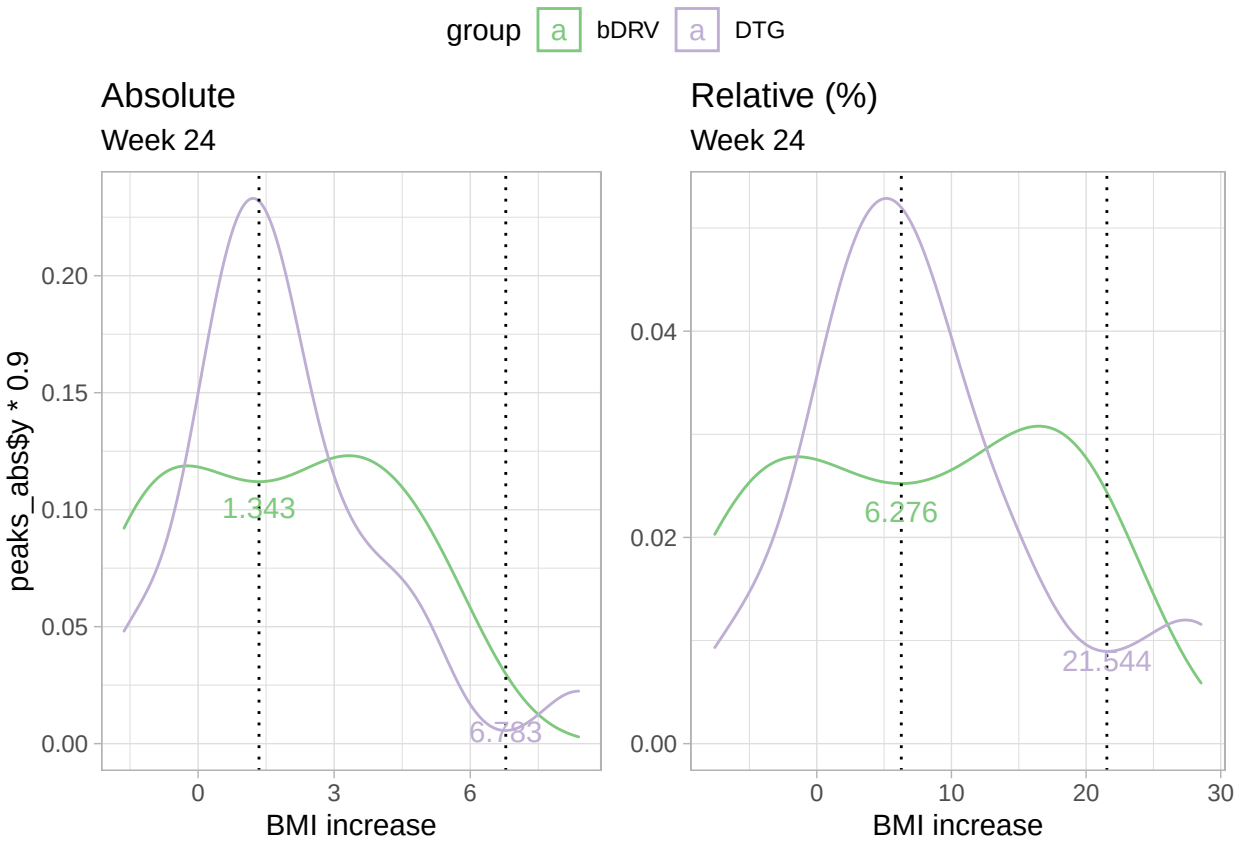
```
        width = 1,
        position = position_dodge(.8)
      ) +
      geom_line(
        data = summ_df,
        aes(
          x = as.factor(long_var),
          y = median_gr,
          color = cat_var,
          lty = bmi_c,
          group = bmi_c_g
        ),
        inherit.aes = F,
        position = position_dodge(.8)
      ) +
      # stat_smooth(method = "loess", alpha = .2, aes(fill = cat_var), position = position_dodge(.8)) +
      theme_bw() +
      labs(
        x = "weeks",
        y = "richness",
        color = "group",
        shape = "BMI increase",
        lty = "BMI increase",
        title = "richness vs BMI increase"
      ) +
      ggpubr::stat_pvalue_manual(cat_test_df, label = "p.adj.signif", hide.ns = T) +
      ggpubr::stat_pvalue_manual(long_test, label = "p.adj.signif", hide.ns = T) +
      scale_color_brewer(palette = "Dark2") +
      scale_fill_brewer(palette = "Dark2") +
      theme(axis.title.x = element_blank())
    # scale_x_continuous(breaks = c(0,48,96))

spaghetti_plot_list %>%
ggsave(plot= ., device = "svg",path = here::here("output","figures"), filename = "bmi_richness.svg", dp

spaghetti_plot_list
```



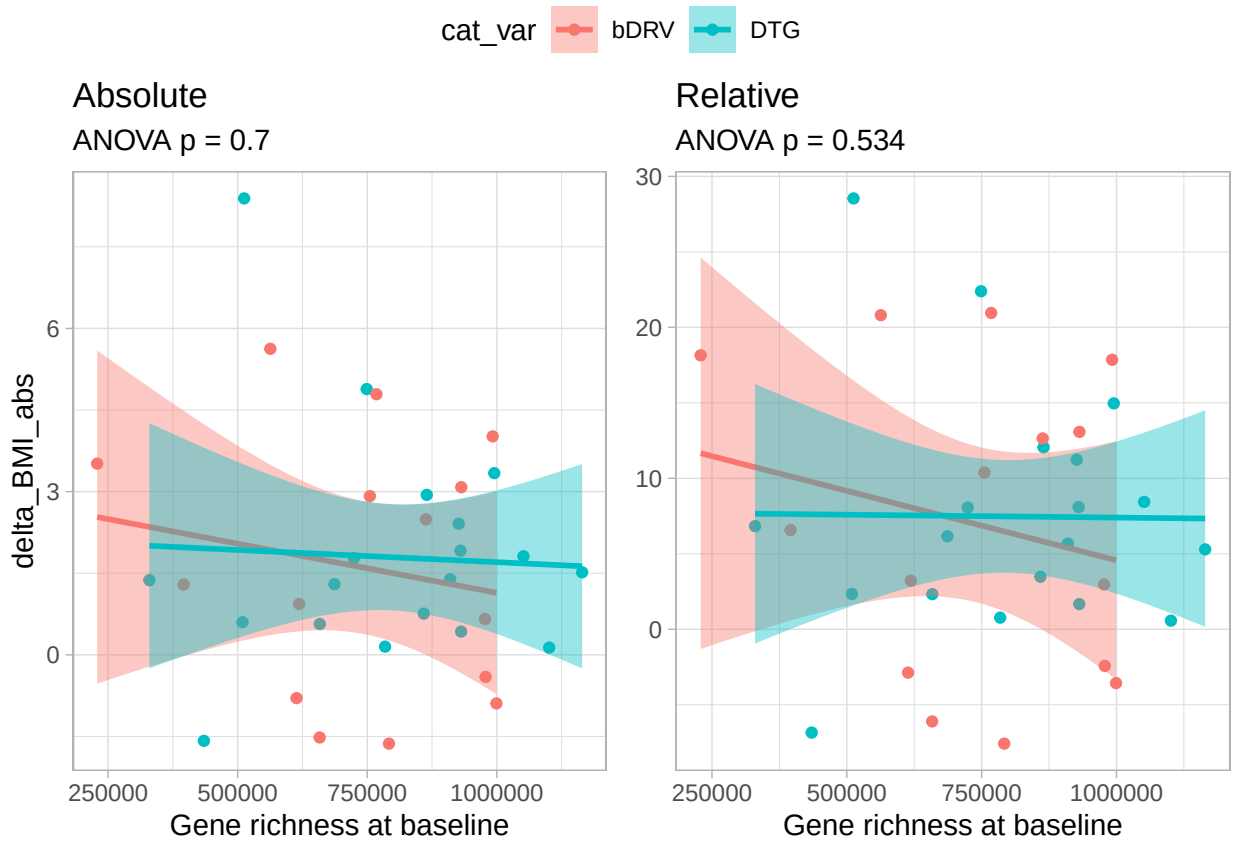
Apparently, the only significant differences are between both subgroups belonging to the DTG arm, this suggests the GR change relates more to group rather than BMI.

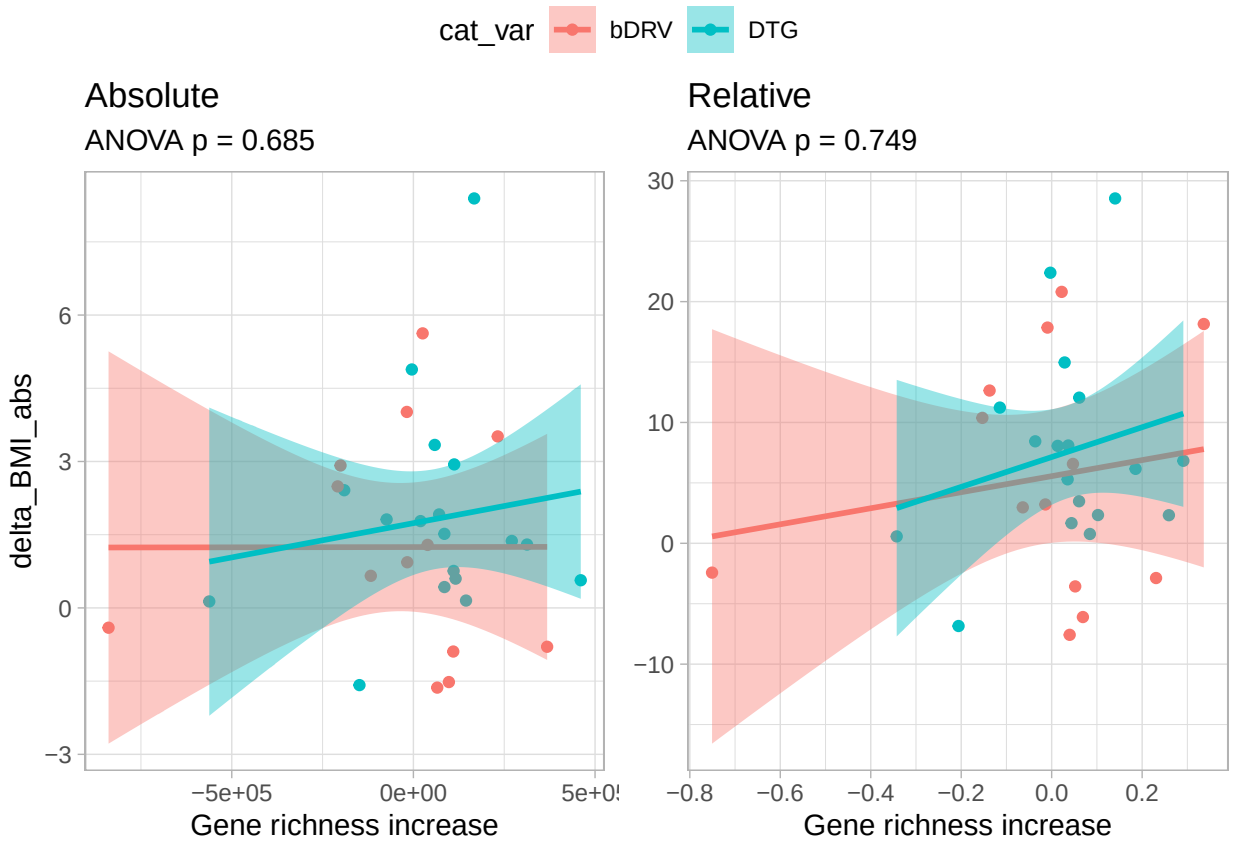
arm	group1	group2	median1	iqr1	median2	iqr2	medianChange	iqr_change	p.adj
H_BMI_DTG	0	96	748944	155107.8	910198	178215.3	197892	179453.0	0.0312
L_BMI_DTG	0	96	758006	287818.3	981756	327681.0	169678	172484.4	0.0312

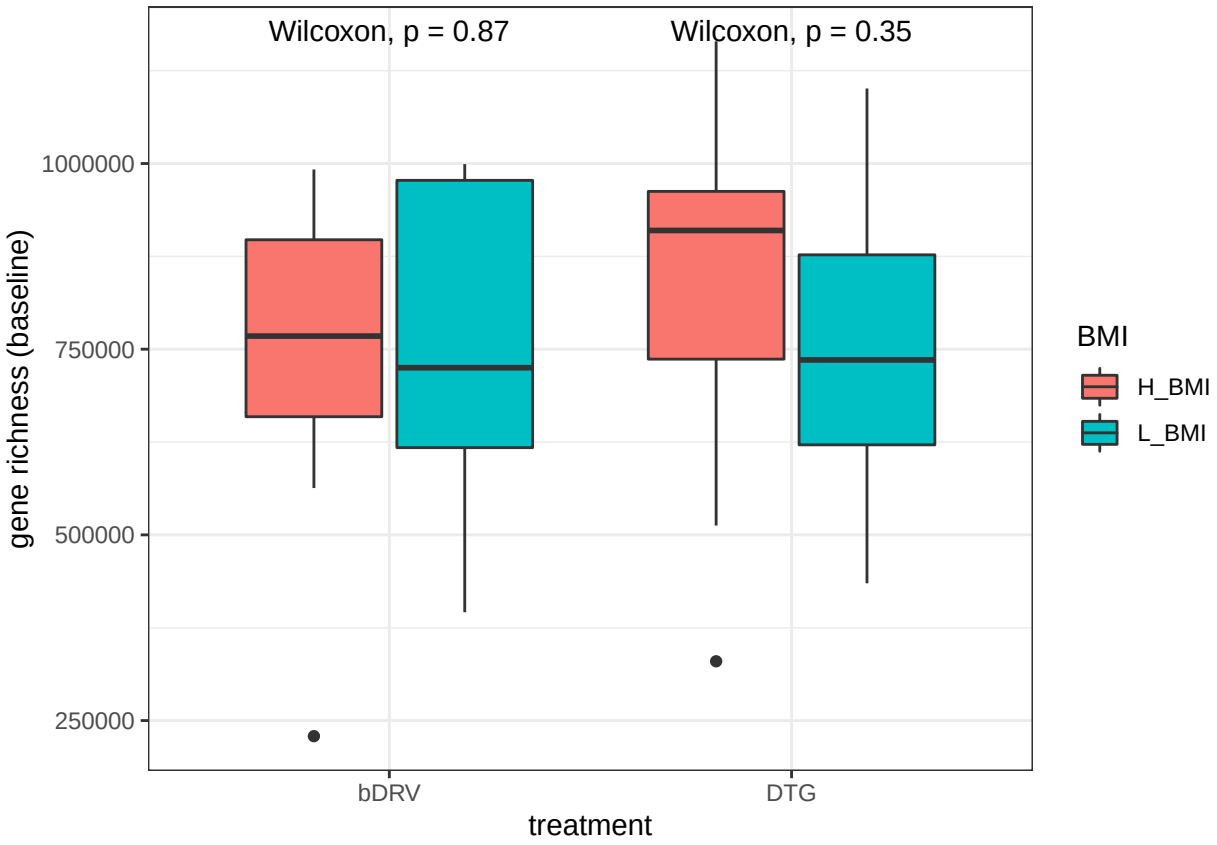
```
## # A tibble: 0 x 11
## # ... with 11 variables: long_var <int>, group1 <chr>, group2 <chr>, statistic <dbl>, stat_group1 <ch
## #   y.position <dbl>, xmin <dbl>, xmax <dbl>, p.adj <dbl>, p.adj.signif <chr>
```

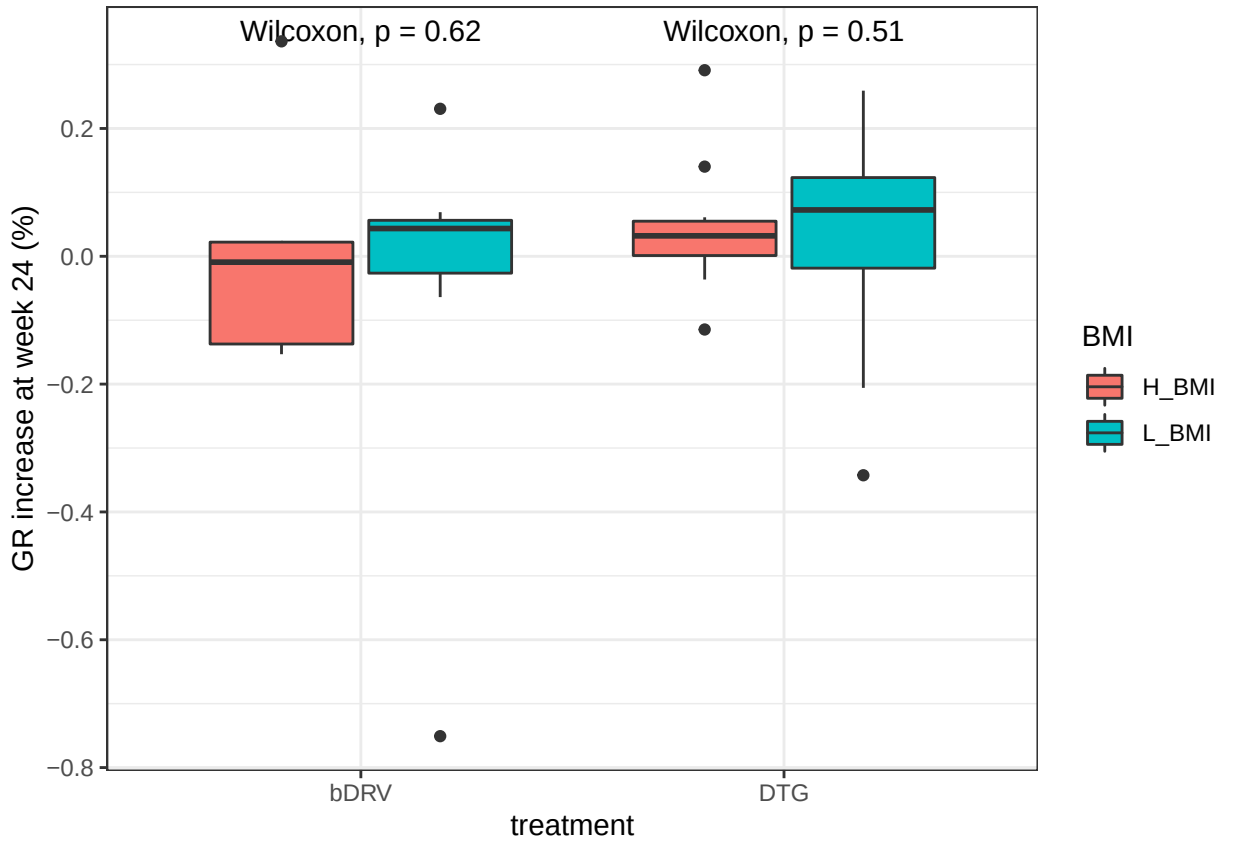


Here it appears the threshold should be un BMI increase of 1.3, let's do that:


cat_var
bDRV
DTG
Absolute
ANOVA p = 0.7
Relative
ANOVA p = 0.534
delta_BMI_abs
6
3
0
30
20
10
0
250000
500000
750000
1000000
Gene richness at baseline
Gene richness at baseline


cat_var
bDRV
DTG
Absolute
ANOVA p = 0.685
Relative
ANOVA p = 0.749
delta_BMI_abs
6
3
0
−3
−5e+05
0e+00
5e+05
Gene richness increase
30
20
10
0
−10
−0.8
−0.6
−0.4
−0.2
0.0
0.2
Gene richness increase


Wilcoxon, p = 0.87
Wilcoxon, p = 0.35
1000000
750000
500000
250000
gene richness (baseline)
bDRV
DTG
treatment
BMI
H_BMI
L_BMI

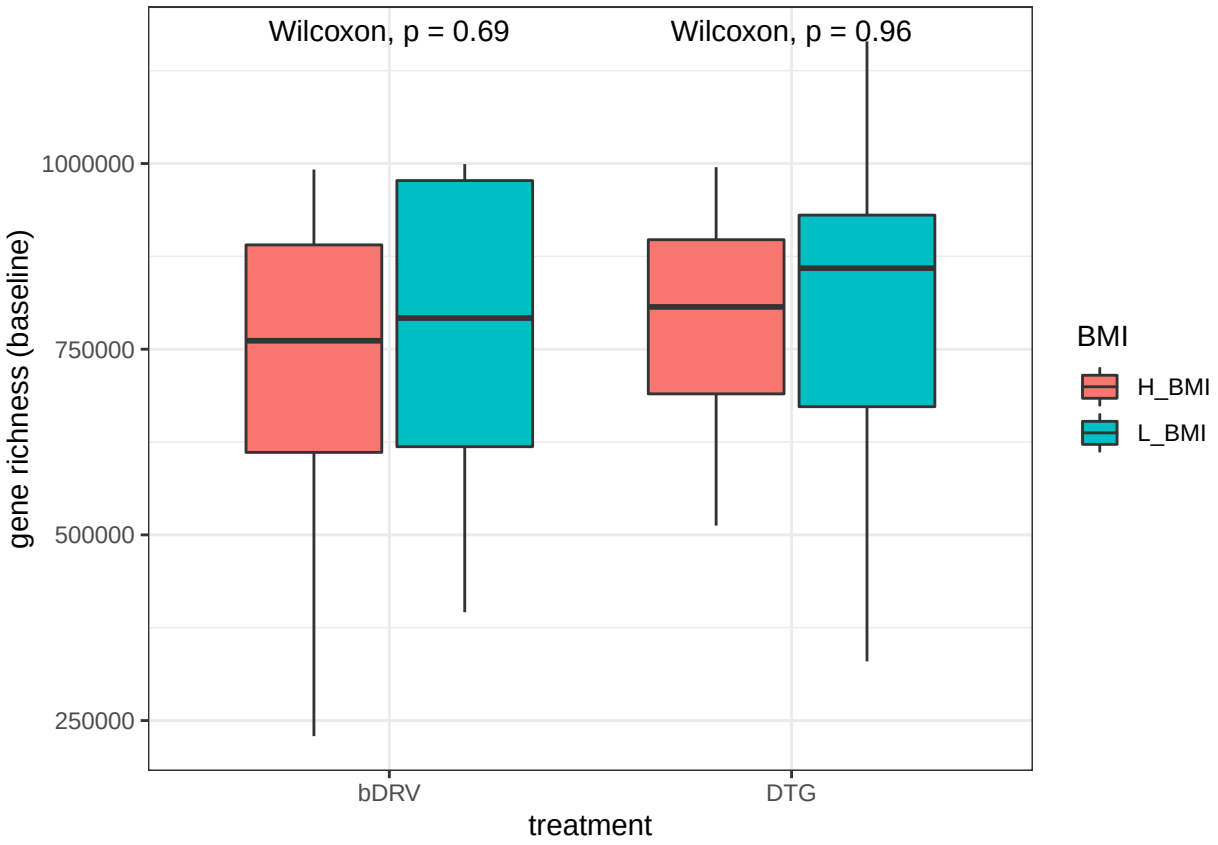


group_bmi	.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif

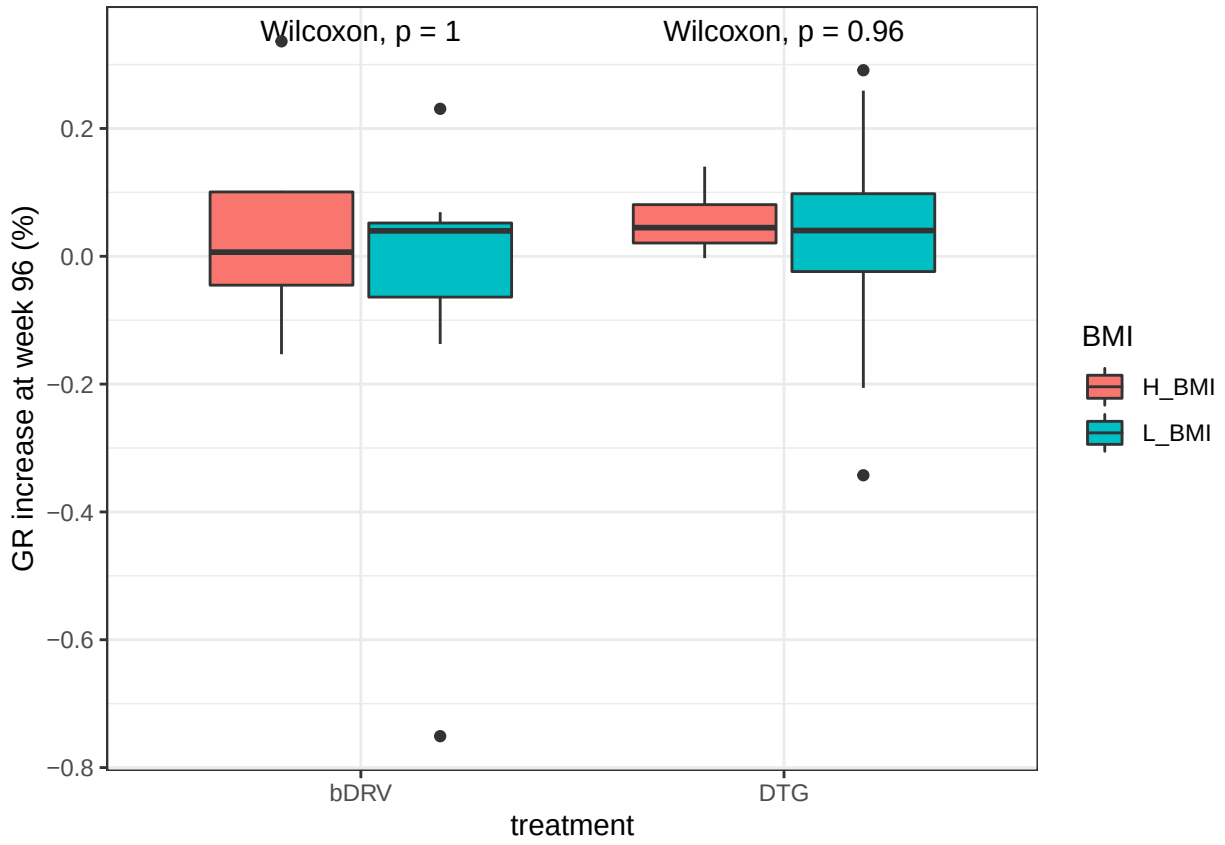
Week 96

Now that we saw the increase in gene richness follows a bimodal distribution in each group, it is possible that we didn't find an effect of gene richness per treatment group due to the confusor effect of BMI. We will split the population between high (BMI_H) and low BMI increase (BMI_L). The cutting point will be an absolute increase of 2.5 in BMI, as seen earlier in the delta distribution.

```
delta_bmi_df %>%
  dplyr::mutate(bmi_c = case_when(
    delta_BMI_abs >= 2.5 ~ "H_BMI",
    delta_BMI_abs < 2.5 ~ "L_BMI"
  )) %>%
  # dplyr::select(link_var = record_id, bmi_c) %>%
  # left_join(igc_df, by = "link_var") %>%
  # mutate(cat_var2 = paste(cat_var, bmi_c)) %>%
  dplyr::filter(!is.na(bmi_c)) %>%
  ggplot(aes(y = gr0, x = cat_var, fill = bmi_c)) +
  ggpubr::stat_compare_means(method = "wilcox.test") +
  geom_boxplot() +
  theme_bw() +
  labs(x = "treatment", y = "gene richness (baseline)", fill = "BMI")
```




```
delta_bmi_df %>%
  dplyr::mutate(bmi_c = case_when(
    delta_BMI_abs >= 2.5 ~ "H_BMI",
    delta_BMI_abs < 2.5 ~ "L_BMI"
  )) %>%
  dplyr::filter(!is.na(bmi_c)) %>%
  ggplot(aes(y = delta_gr_rel, x = cat_var, fill = bmi_c)) +
  ggpubr::stat_compare_means(method = "wilcox.test") +
  geom_boxplot() +
  theme_bw() +
  labs(x = "treatment", y = "GR increase at week 96 (%)", fill = "BMI")
```




```
delta_bmi_df %>%
  dplyr::mutate(bmi_c = case_when(
    delta_BMI_abs >= 2.5 ~ "H_BMI",
    delta_BMI_abs < 2.5 ~ "L_BMI"
  )) %>%
  dplyr::select(record_id, bmi_c) %>%
  dplyr::rename(link_var = record_id) %>%
  dplyr::right_join(igc_df, by = "link_var") %>%
  dplyr::filter(!is.na(bmi_c)) %>%
  dplyr::mutate(group_bmi = paste(cat_var, bmi_c, sep = "_")) %>%
  group_by(group_bmi) %>%
  rstatix::wilcox_test(richness ~ long_var)
```

```
## # A tibble: 24 x 10
##    group_bmi  .y.      group1 group2    n1    n2 statistic     p p.adj p.adj.signif
##  * <chr>      <chr>    <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <chr>
##  1 bDRV_H_BMI richness 0      24         6     5        19 0.537     1 ns
##  2 bDRV_H_BMI richness 0      48         6     6        17 0.937     1 ns
##  3 bDRV_H_BMI richness 0      96         6     6        15 0.699     1 ns
##  4 bDRV_H_BMI richness 24     48         5     6        13 0.792     1 ns
##  5 bDRV_H_BMI richness 24     96         5     6         9 0.329     1 ns
##  6 bDRV_H_BMI richness 48     96         6     6        17 0.937     1 ns
##  7 bDRV_L_BMI richness 0      24         9     9        45 0.73      1 ns
##  8 bDRV_L_BMI richness 0      48         9     6        21 0.529     1 ns
##  9 bDRV_L_BMI richness 0      96         9     8        37 0.963     1 ns
```

```
## 10 bDRV_L_BMI richness 24     48         9     6        21 0.529     1 ns
## # ... with 14 more rows
```

```
delta_bmi_df %>%
  dplyr::mutate(bmi_c = case_when(
    delta_BMI_abs >= 2.5 ~ "H_BMI",
    delta_BMI_abs < 2.5 ~ "L_BMI"
  )) %>%
  dplyr::select(record_id, bmi_c) %>%
  dplyr::rename(link_var = record_id) %>%
  dplyr::right_join(igc_df, by = "link_var") %>%
  dplyr::filter(!is.na(bmi_c)) %>%
  dplyr::mutate(group_bmi = paste(cat_var, bmi_c, sep = "_")) %>%
  ggplot(aes(y = richness, x = as.factor(long_var), color = group_bmi)) +
  geom_boxplot(aes(x = as.factor(long_var))) +
  ggpubr::stat_compare_means(method = "kruskal.test", hide.ns = F, label = "p.signif") +
  # geom_point() +
  geom_smooth(method = "glm", alpha = 0) +
  theme_bw() +
  labs(x = "week")
```