Differential expression (F2-73, Elite and PI127826 2020 plants)

Marc Galland

10/17/2021

Contents

1 Data import

1.1 Import scaled counts

```
scaled_counts <- read.delim("../Supplemental_data_RNA-seq/scaled_counts.tsv",
                            check.names = F,
                            stringsAsFactors = F)
```

```
scaled_counts <- read.delim("../Supplemental_data_RNA-seq/scaled_counts.tsv",
                            check.names = F,
                            stringsAsFactors = F) %>%
  mutate(gene = gsub(pattern = "mRNA:", replacement = "", x = gene)) %>%
  dplyr::select("gene", "F2.73", "Elite_2020", "PI127826_2020")
head(scaled_counts)
```

```
##                  gene       F2.73 Elite_2020 PI127826_2020
## 1 Solyc00g005000.2.1    3.494954     0.0000      0.000000
## 2 Solyc00g005020.1.1    0.000000     0.0000      2.650106
## 3 Solyc00g005040.2.1   55.919260    49.2653     13.250530
## 4 Solyc00g005050.2.1 1382.254220  1383.7123   1810.022376
## 5 Solyc00g005060.1.1    0.000000     0.0000      0.000000
## 6 Solyc00g005070.1.1    0.000000     0.0000      0.000000
```

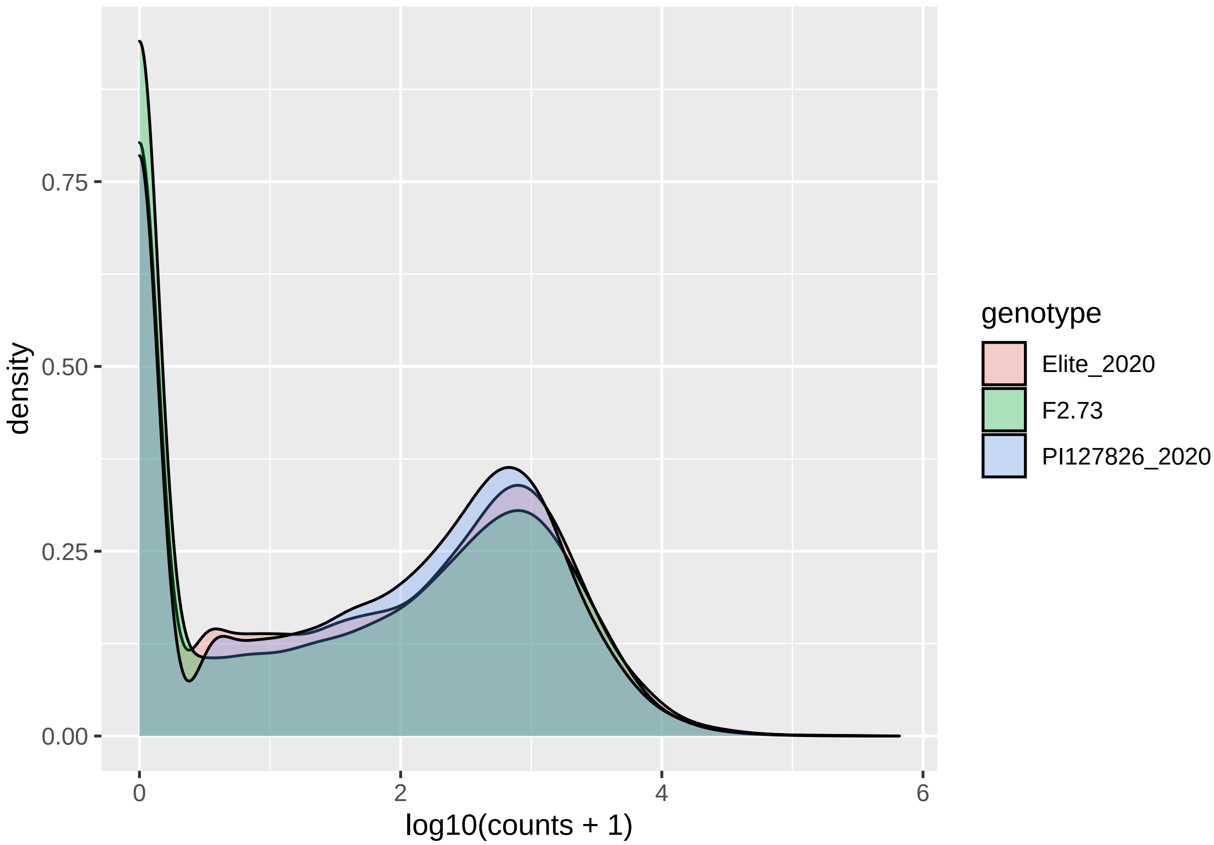
1.2 Mapping summary

```
# read.csv("../Supplemental_data_RNA-seq/mapping_summary.csv",
#          stringsAsFactors = F,
#          check.names = F) %>%
#   knitr::kable()
```

2 QC plots

2.1 Plot density counts

```
scaled_counts %>%
  pivot_longer(- gene, names_to = "genotype", values_to = "counts") %>%
  ggplot(aes(x = log10(counts + 1), fill = genotype)) +
  geom_density(alpha = 0.3)
```



The density of genes with low count values (log10 = 0.3) is lower for F2-73 but seems to be comparable for other genes.

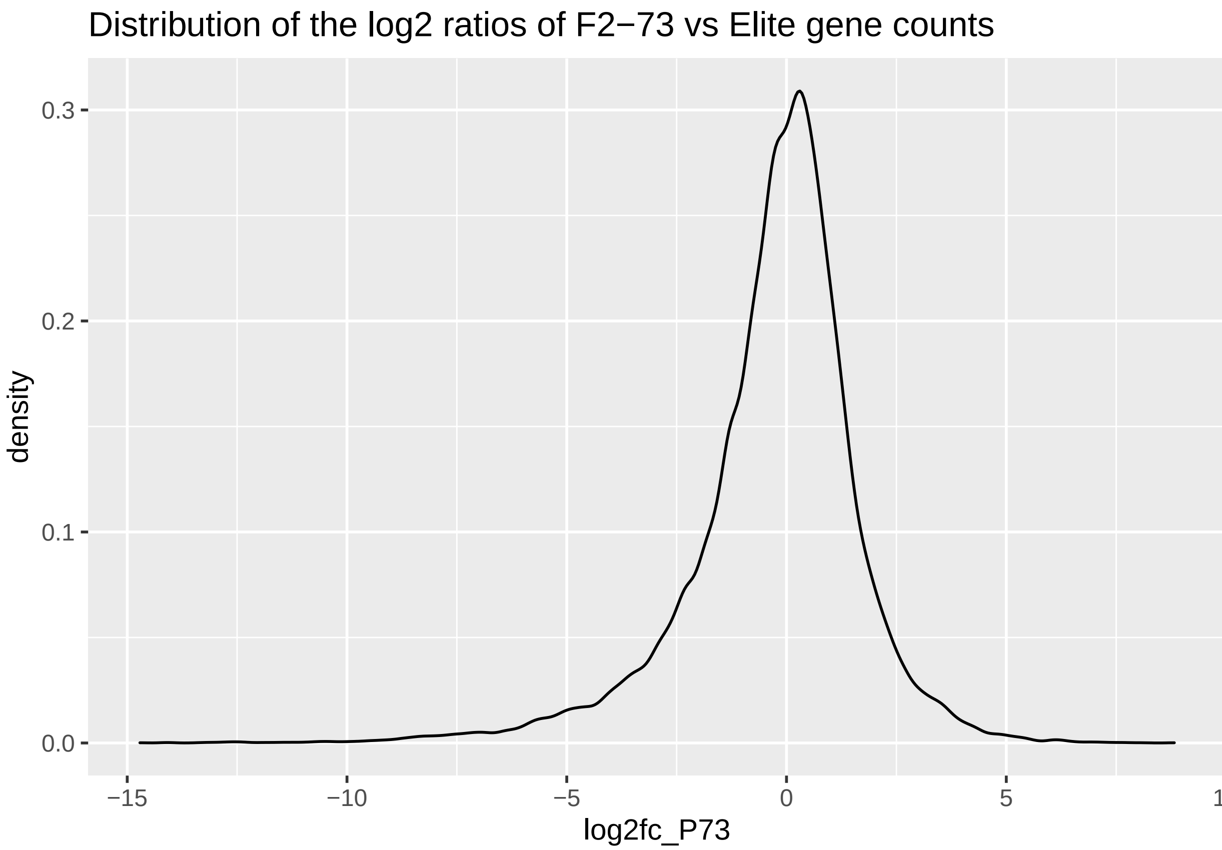
2.2 Plot log2ratio F2-73 vs Elite

First, let's extract genes with counts > 0

```
genes_sums <- scaled_counts %>% column_to_rownames("gene") %>% rowSums()
genes_non_null <- genes_sums[genes_sums > 0]
genes_non_null <- names(genes_non_null)
```

```
scaled_counts %>%
  filter(gene %in% genes_non_null) %>%
  mutate(log2fc_P73 = log2(`F2.73`/Elite_2020)) %>%
  ggplot(aes(x = log2fc_P73)) +
  geom_density() +
  ggtitle("Distribution of the log2 ratios of F2-73 vs Elite gene counts")
```

```
## Warning: Removed 5110 rows containing non-finite values (stat_density).
```

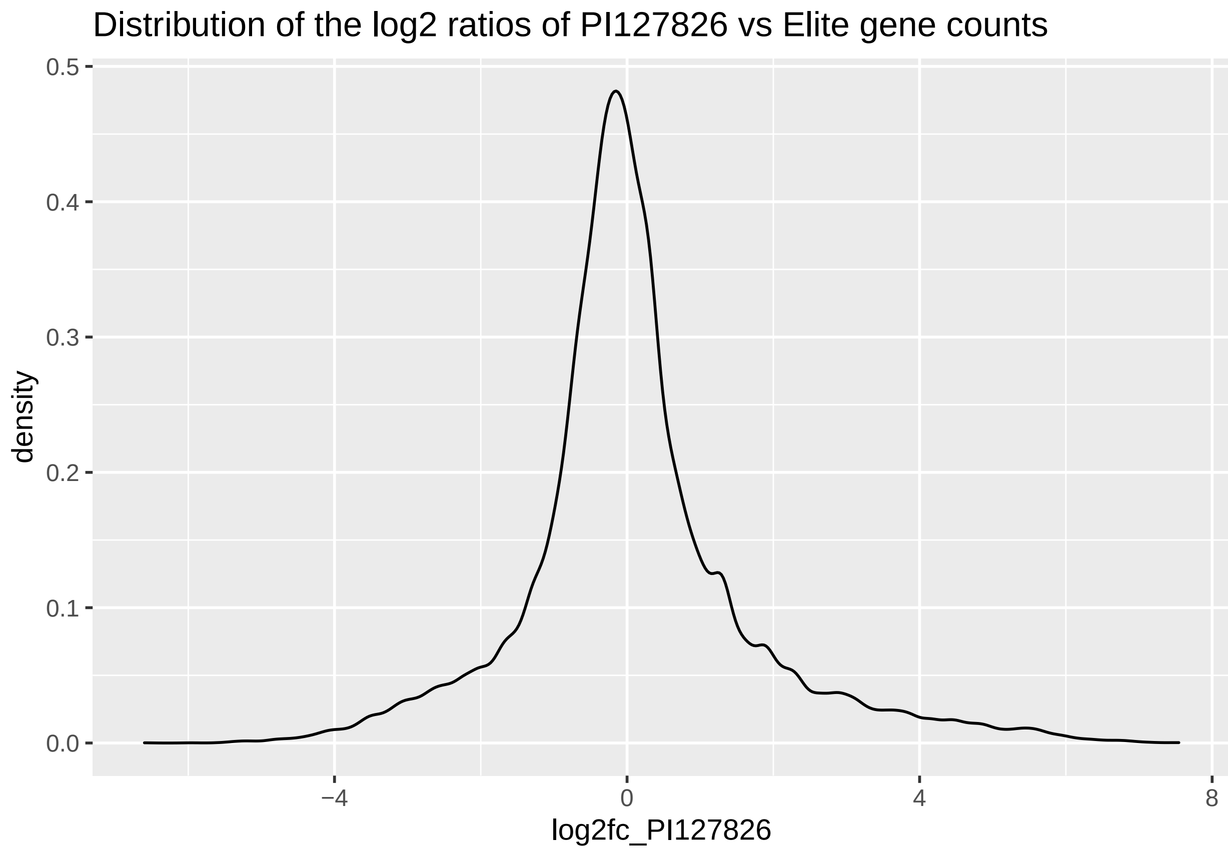


The log2ratio distribution looks OK. Compared to to the log2 ratios of F2-28, it is more skewed towards negative log2ratios.

2.3 Plot log2ratio PI127826 vs Elite

```
scaled_counts %>%
  filter(gene %in% genes_non_null) %>%
  mutate(log2fc_PI127826 = log2(PI127826_2020/Elite_2020)) %>%
  ggplot(aes(x = log2fc_PI127826)) +
  geom_density() +
  ggtitle("Distribution of the log2 ratios of PI127826 vs Elite gene counts")
```

```
## Warning: Removed 3633 rows containing non-finite values (stat_density).
```



Distribution of the log2 ratios of PI127826 vs Elite gene counts
0.5
0.4
0.3
0.2
0.1
0.0
density
−4
0
4
8
log2fc_PI127826

3 Compute DE genes based on log2ratio Z-score

3.1 Calculate log2ratios

A positive log2ratio for F2-73 and PI127826 means that the gene is more expressed in F2-28 and PI127826 (relative to the Elite line).

Let's calculate the log2ratio and remove the "Infinite" values.

```
log2ratio <-
  scaled_counts %>%
  filter(gene %in% genes_non_null) %>%
  mutate(log2ratio_P73 = log2(`F2.73`/Elite_2020)) %>%
  mutate(log2ratio_PI127826 = log2(PI127826_2020/Elite_2020)) %>%
  select(gene, log2ratio_P73, log2ratio_PI127826) %>%
  filter(!grepl(pattern = "Inf", x = log2ratio_P73)) %>%
  filter(!grepl(pattern = "Inf", x = log2ratio_PI127826))

head(log2ratio)
```

```
##                  gene log2ratio_P73 log2ratio_PI127826
## 1 Solyc00g005040.2.1   0.182773531         -1.8945217
## 2 Solyc00g005050.2.1  -0.001520983          0.3874636
## 3 Solyc00g005840.2.1  -0.957797227          0.6485176
## 4 Solyc00g005860.1.1  -2.615592607         -3.0148159
## 5 Solyc00g006470.1.1 -10.568964562          1.7963727
## 6 Solyc00g006490.2.1  -0.683533784          0.1157466
```

A total of **21054** have a finite log2ratio in both F2-73 vs Elite and PI127826 vs Elite.

3.2 Calculate Z-scores and associated p-values

Let's calculate the Z-score of the log2ratio + its associated p-value

```
log2ratio_zscores_pvals <-
  log2ratio %>%
  mutate(zscore_P73 = scale(log2ratio_P73, center = T, scale = T)) %>%
  mutate(zscore_PI127826 = scale(log2ratio_PI127826, center = T, scale = T)) %>%
  mutate(pval_P73 = pnorm(q = abs(zscore_P73), mean = 0, sd=1, log.p = FALSE, lower.tail=FALSE)) %>%
  mutate(pval_PI127826 = pnorm(q = abs(zscore_PI127826), mean = 0, sd=1, log.p = FALSE, lower.tail=FALSE
  arrange(desc(log2ratio_P73)) %>%
  as_tibble()

head(log2ratio_zscores_pvals)
```

```
## # A tibble: 6 x 7
##   gene  log2ratio_P73 log2ratio_PI127~ zscore_P73[,1] zscore_PI127826~
##   <chr>         <dbl>            <dbl>          <dbl>            <dbl>
## 1 Soly~          8.82             3.63           4.67             2.41
## 2 Soly~          8.10             6.21           4.29             4.17
## 3 Soly~          7.74             1.63           4.11             1.05
## 4 Soly~          7.57             5.70           4.03             3.82
## 5 Soly~          7.36             7.17           3.92             4.82
## 6 Soly~          7.17             3.63           3.82             2.41
## # ... with 2 more variables: pval_P73[,1] <dbl>, pval_PI127826[,1] <dbl>
```

3.3 Add original counts and annotations

Add back the scaled counts.

```
log2ratio_zscores_pvals_with_counts <- inner_join(scaled_counts, log2ratio_zscores_pvals, by = "gene")
```

Add descriptions

```
annots <- read.csv("info/ITAG2.4_loci_gene_descriptions.csv", stringsAsFactors = F)

final <-
  log2ratio_zscores_pvals_with_counts %>%
  mutate(locus = substr(gene, start = 1, stop = 14)) %>%
  inner_join(x = ., y = annots, by = "locus") %>%
  as_tibble()

dim(final)
```

```
## [1] 21054    12
```

3.4 Write to CSV file

```
dir.create(path = "./tables_F2-73/", showWarnings = F, recursive = T)
write.csv(final,
          file = "tables_F2-73/diff_res_F2-73_or_PI127826_vs_Elite.csv",
          row.names = F,
          quote = F)
```

```
final %>% filter(pval_P73 < 0.01) %>% dim()
```

```
## [1] 791  12
```

4 MEP and MVA pathway gene analysis

4.1 Import MEP and MVA gene identifiers

```
mep_mva_gene_ids <- read.csv("info/mep_mva_terpene_gene_ids.csv",
                             stringsAsFactors = F)
```

4.2 Filter for significant DE genes

Should be significant ($p < 0.05$) in F2-73 vs Elite.

```
signif_genes <- filter(final, pval_P73 < 0.05) %>% pull(gene)
```

4.3 Keep only MEP and MVA genes significant

```
mep_mva_genes <- inner_join(final, mep_mva_gene_ids)
```

```
## Joining, by = "locus"
```

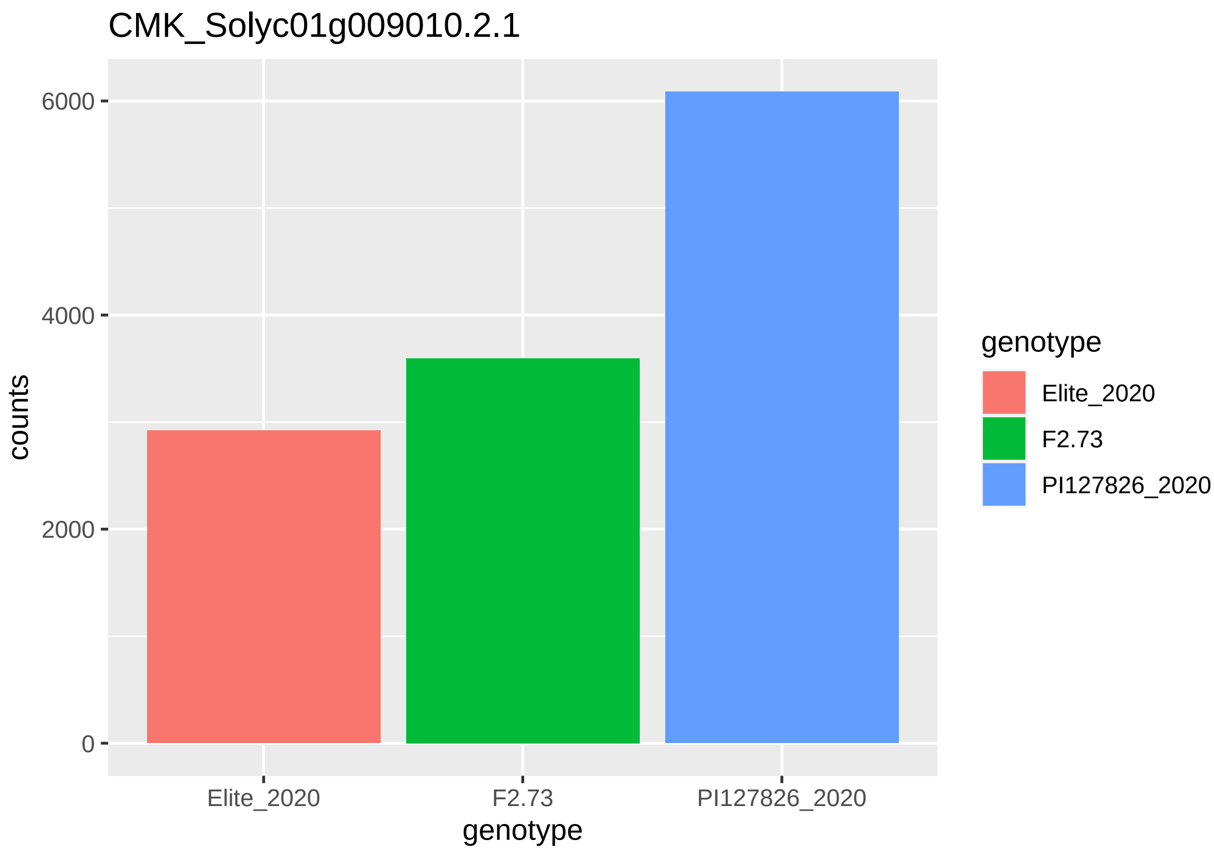
```
mep_mva_gene_signif <-
  mep_mva_genes %>%
  filter(gene %in% signif_genes)

# show table
mep_mva_gene_signif %>%
  select(name, gene, pathway, log2ratio_P73, log2ratio_PI127826, pval_P73, pval_PI127826) %>%
  knitr::kable()
```

name	gene	pathway	log2ratio_P73	log2ratio_PI127826	pval_P73	pval_PI127826
HMGR	Solyc02g038740.2.1	MVA	5.535523	3.671742	0.001424666	0.007318178
HMGR	Solyc03g032010.2.1	MVA	3.914613	2.923354	0.015629531	0.026619455
HMGR	Solyc03g032020.2.1	MVA	3.315479	2.032462	0.032370075	0.092102361
pMVK	Solyc06g066310.2.1	MVA	2.927697	3.474898	0.049621651	0.010511663
AACT	Solyc07g045350.2.1	MVA	3.651830	2.152733	0.021731436	0.079328502

```
write.csv(mep_mva_gene_signif,
          file = "tables_F2-73/mep_mva_gene_signif.csv",
          row.names = F,
          quote = F)
```

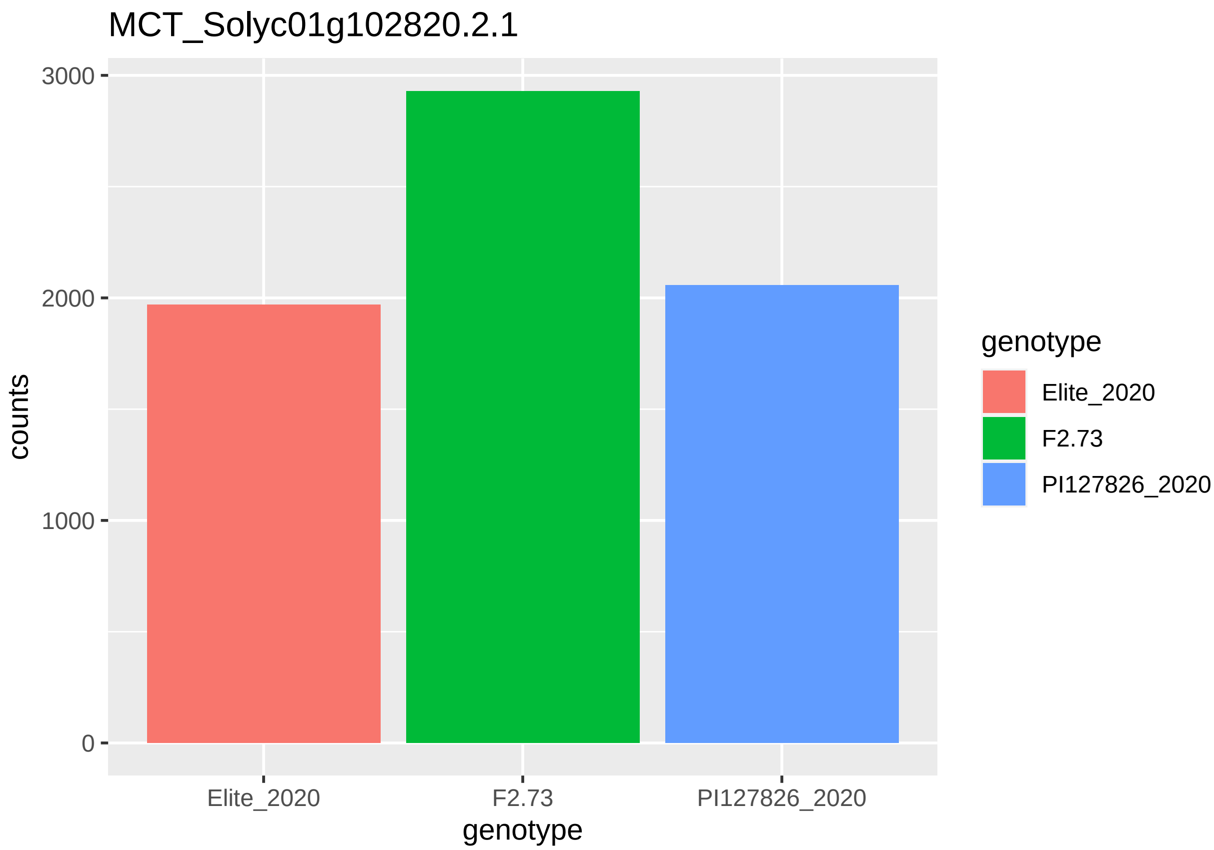
4.4 Plot all MEP and MVA genes

```
for (i in seq_along(mep_mva_genes$gene)){
  tmp_df <- mep_mva_genes[i,]
  tmp_df$title4plot <- paste(tmp_df$name, tmp_df$gene, sep = "_")

  p <-
    tmp_df %>%
#     mutate(plot_title = paste(name, gene, sep = "_")) %>%
    select(title4plot, `F2.73`, Elite_2020, PI127826_2020) %>%
    pivot_longer(- title4plot, names_to = "genotype", values_to = "counts") %>%
    ggplot(., aes(x = genotype, y = counts, fill = genotype)) +
    geom_bar(stat = "identity") +
    ggtitle(tmp_df$title4plot)

  print(p)
}
```

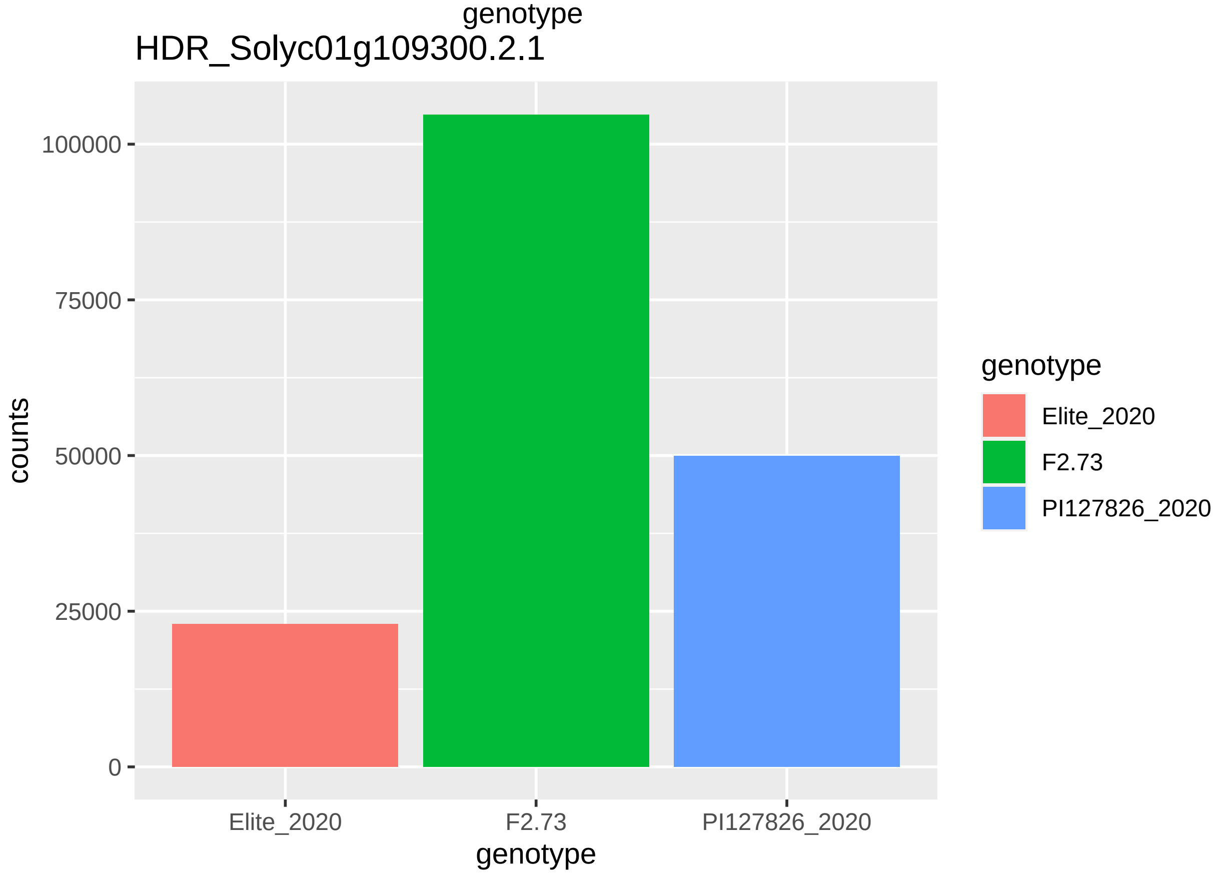



DXS_Solyc01g067890.2.1
counts
4000
3000
2000
1000
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


MVK_Solyc01g098840.2.1
counts
900
600
300
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

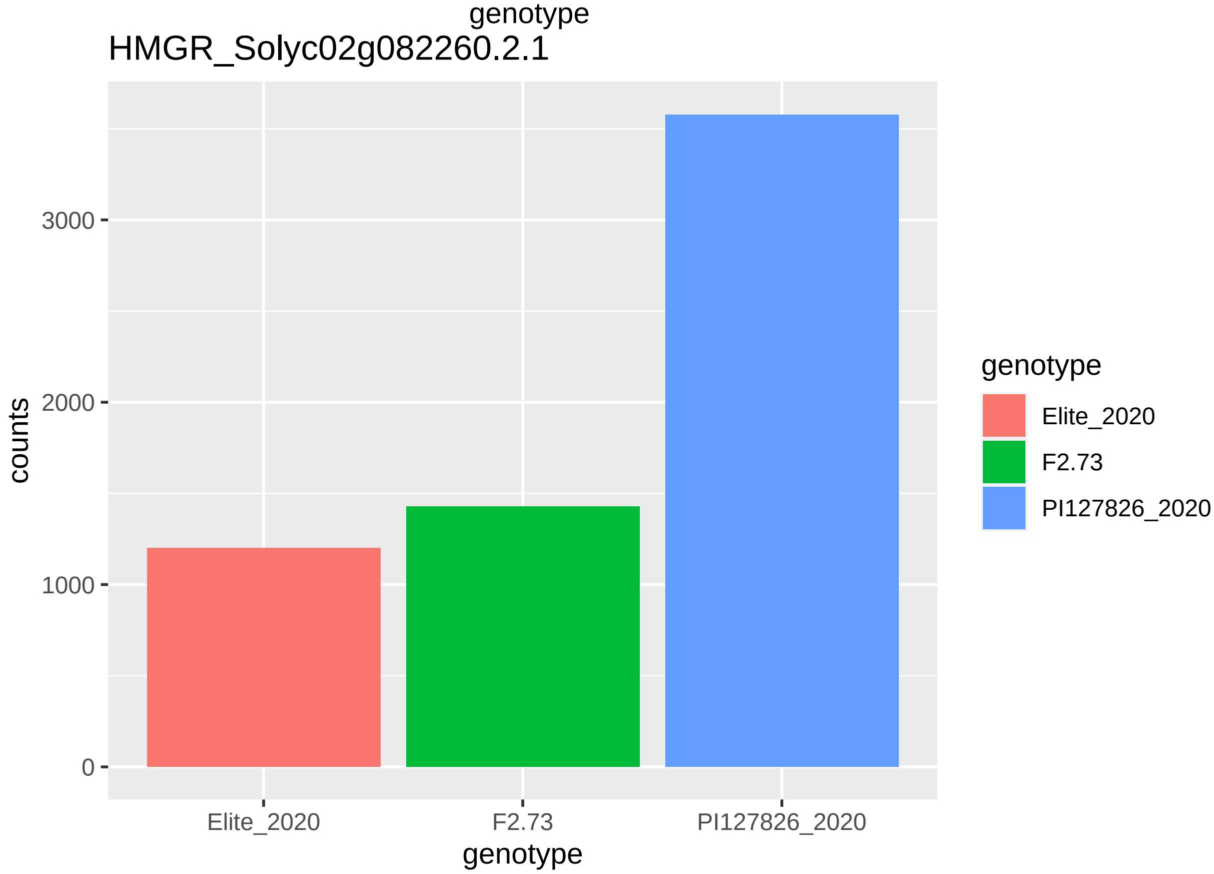


MCT_Solyc01g102820.2.1
3000
2000
1000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

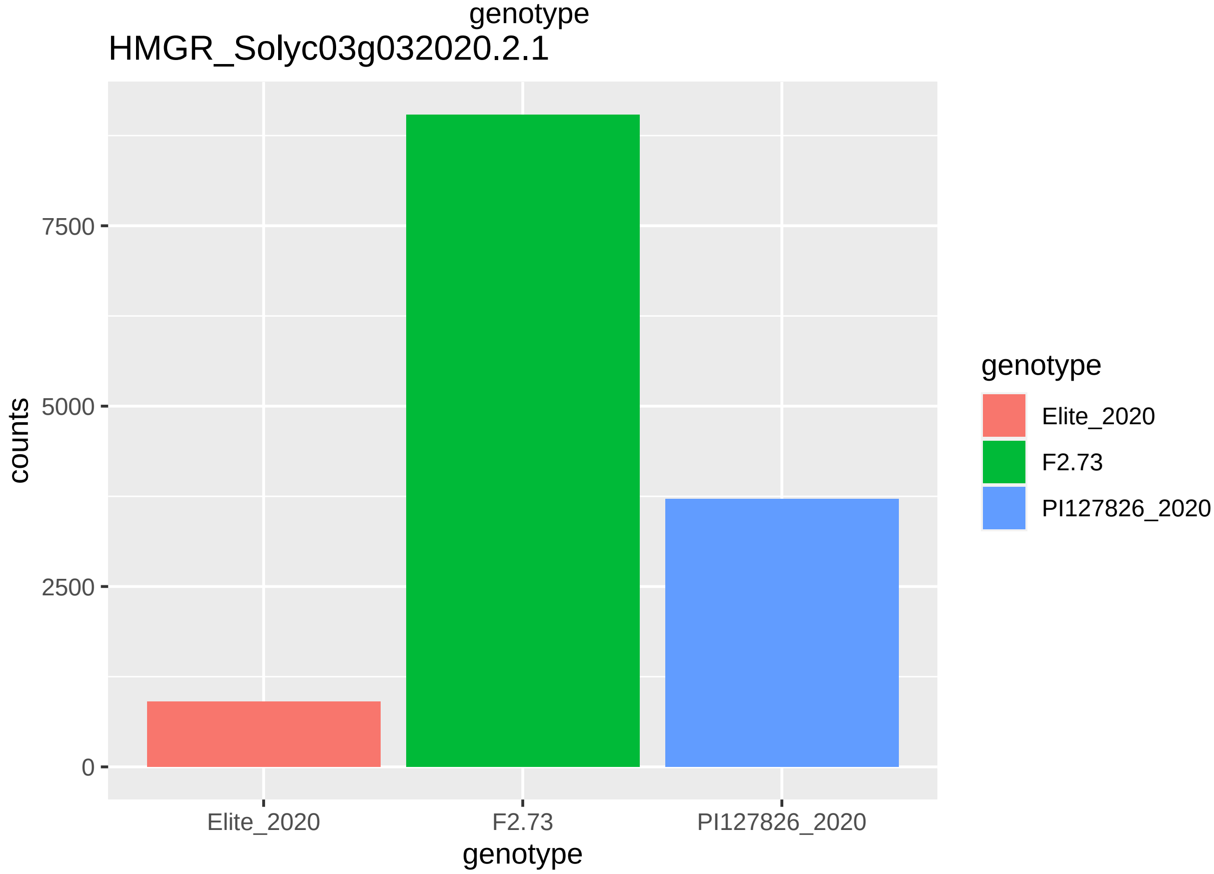


HDR_Solyc01g109300.2.1
100000
75000
50000
25000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

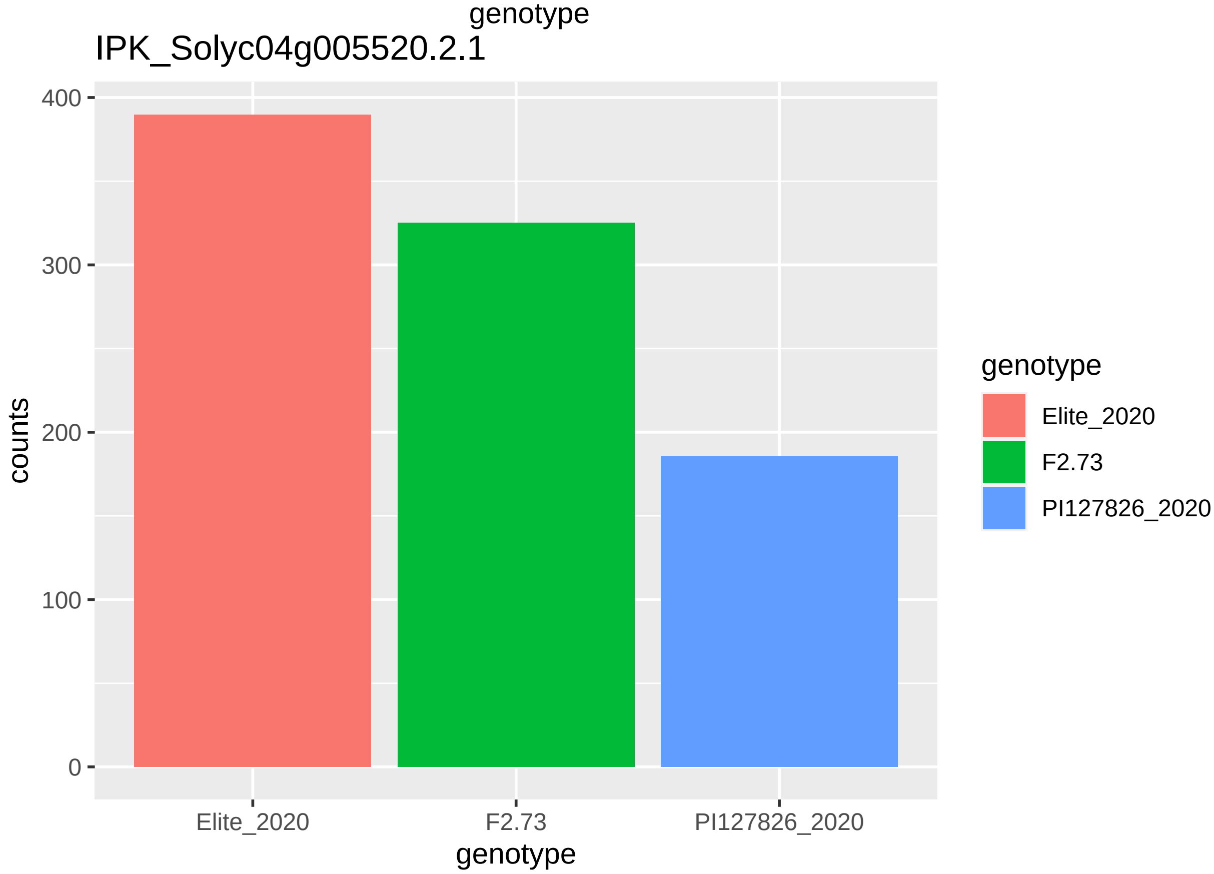

HMGR_Solyc02g038740.2.1
25000
20000
15000
10000
5000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

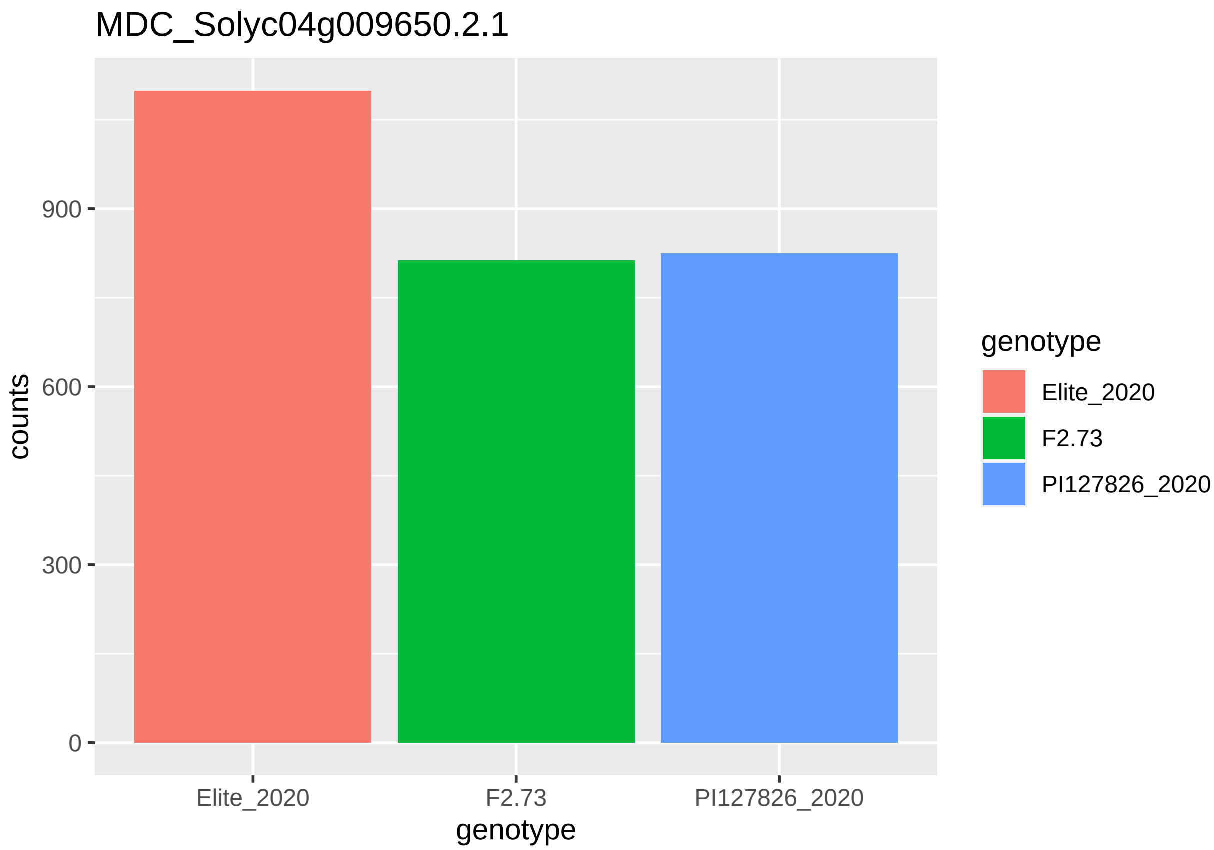

HMGR_Solyc02g082260.2.1
3000
2000
1000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

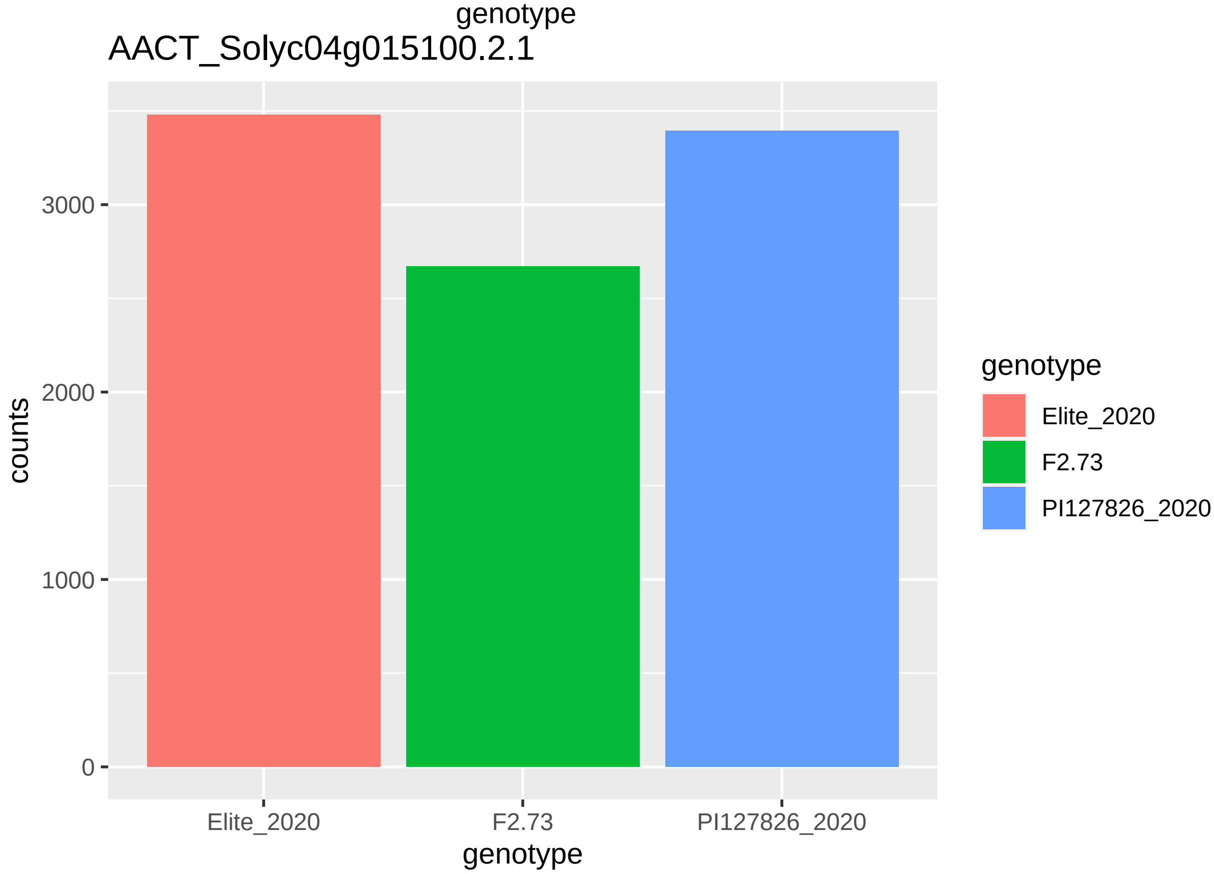

HMGR_Solyc03g032010.2.1
counts
12000
9000
6000
3000
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

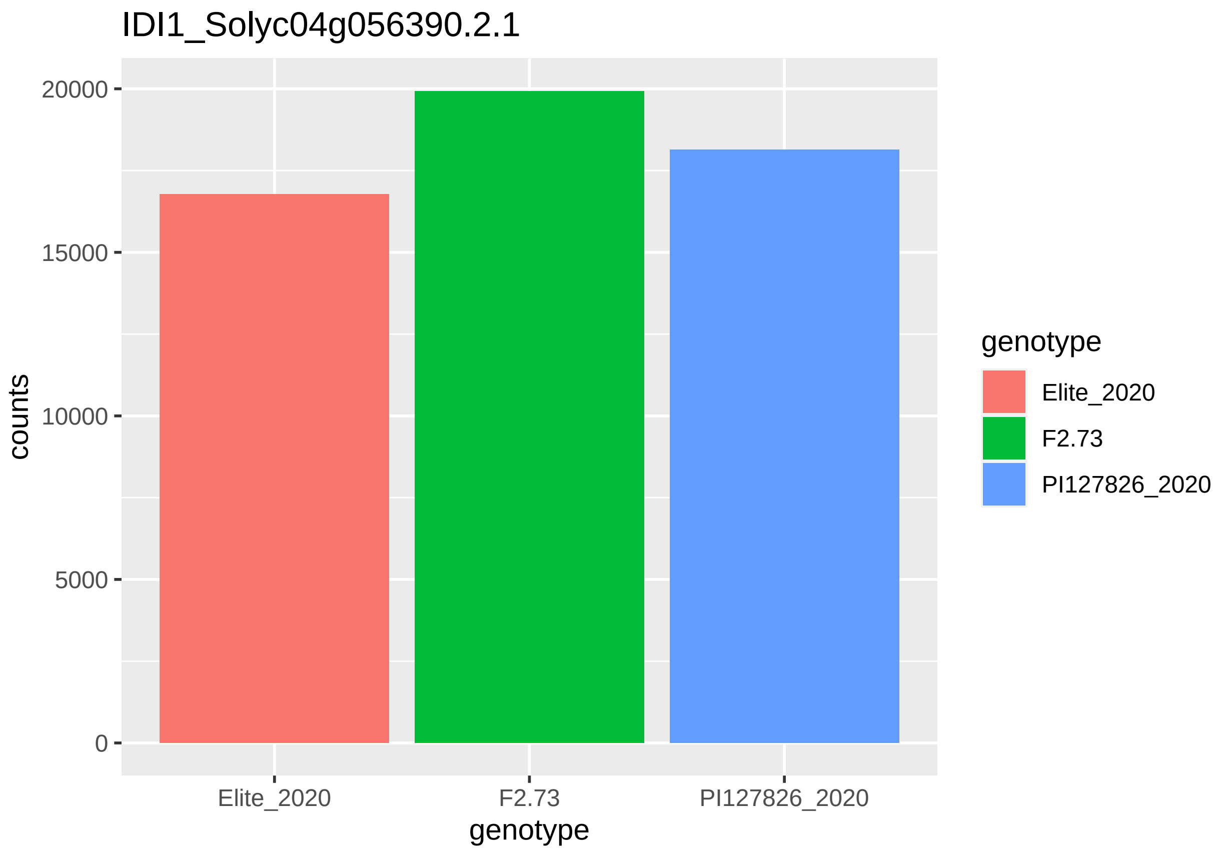

HMGR_Solyc03g032020.2.1
counts
7500
5000
2500
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

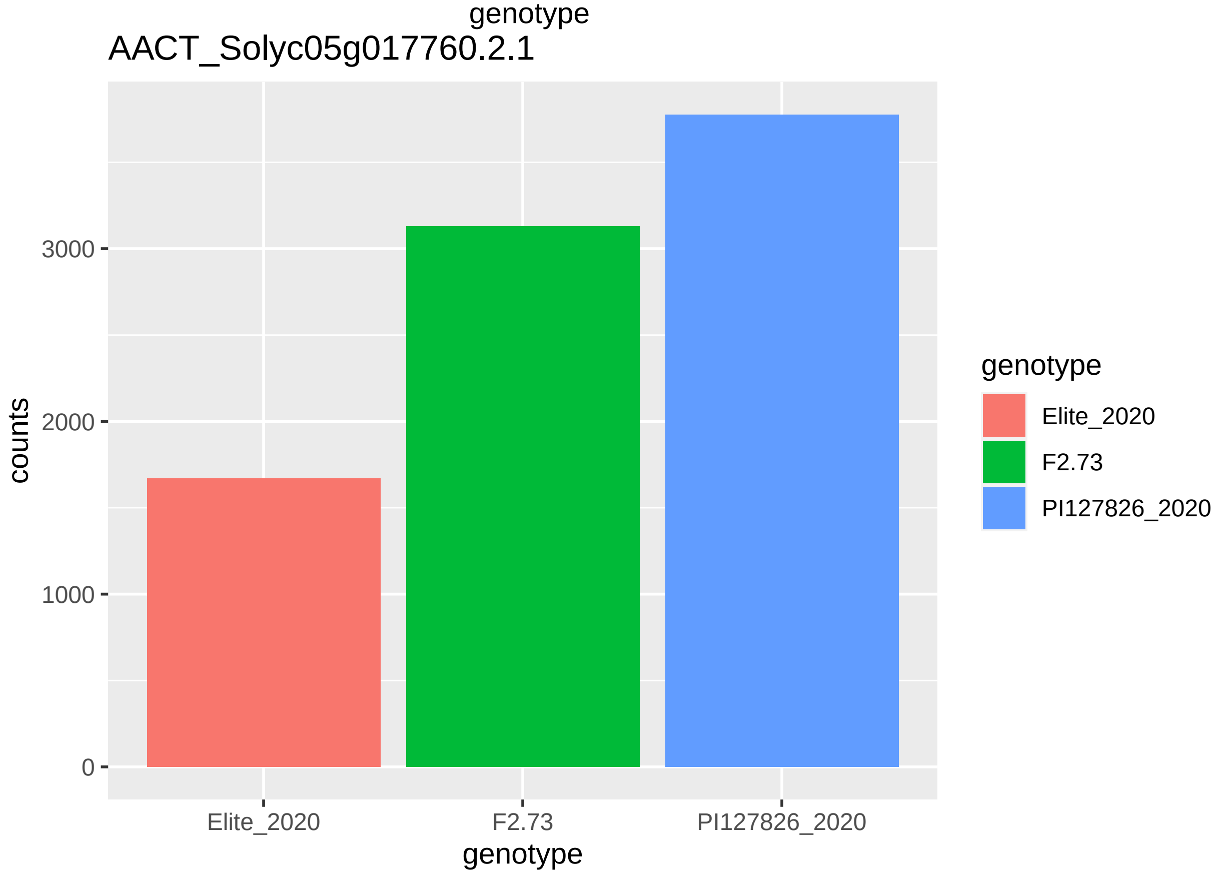

DXR_Solyc03g114340.2.1
counts
15000
10000
5000
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

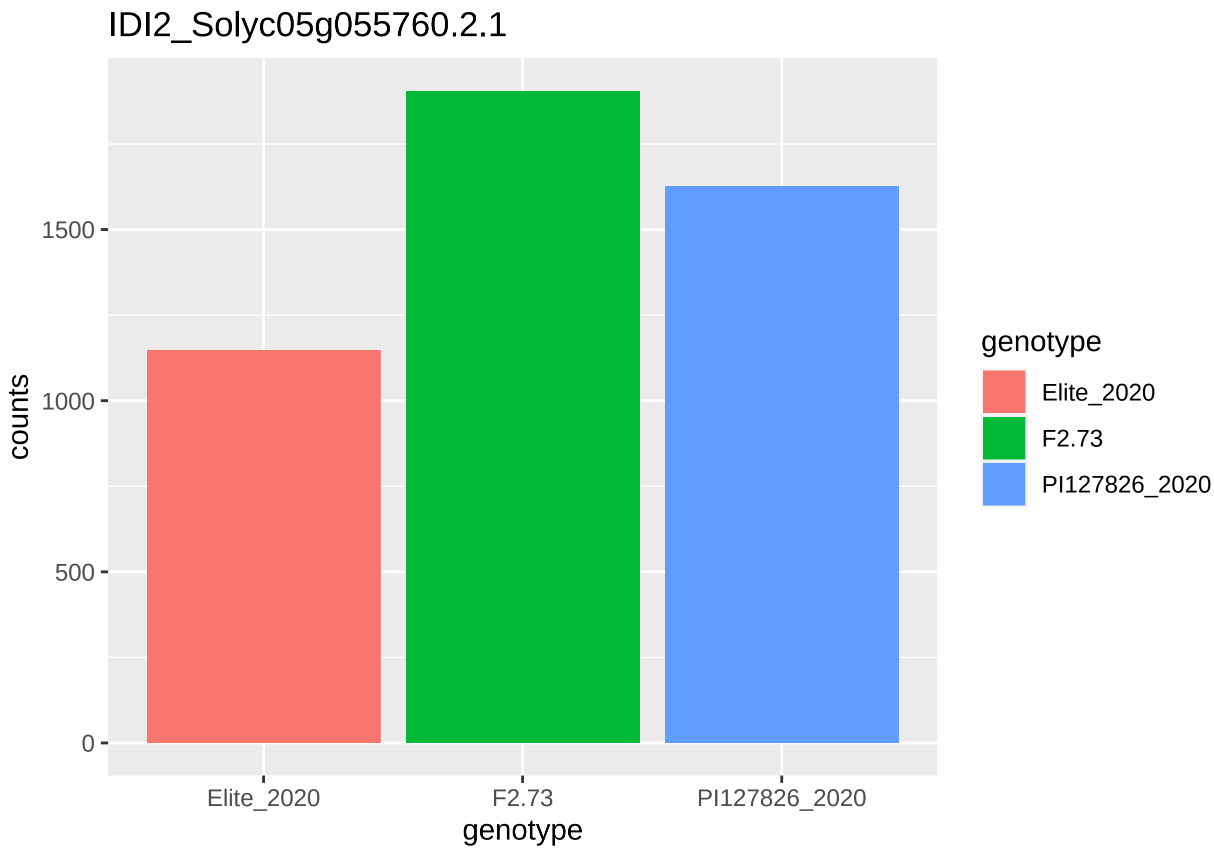

IPK_Solyc04g005520.2.1
counts
400
300
200
100
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

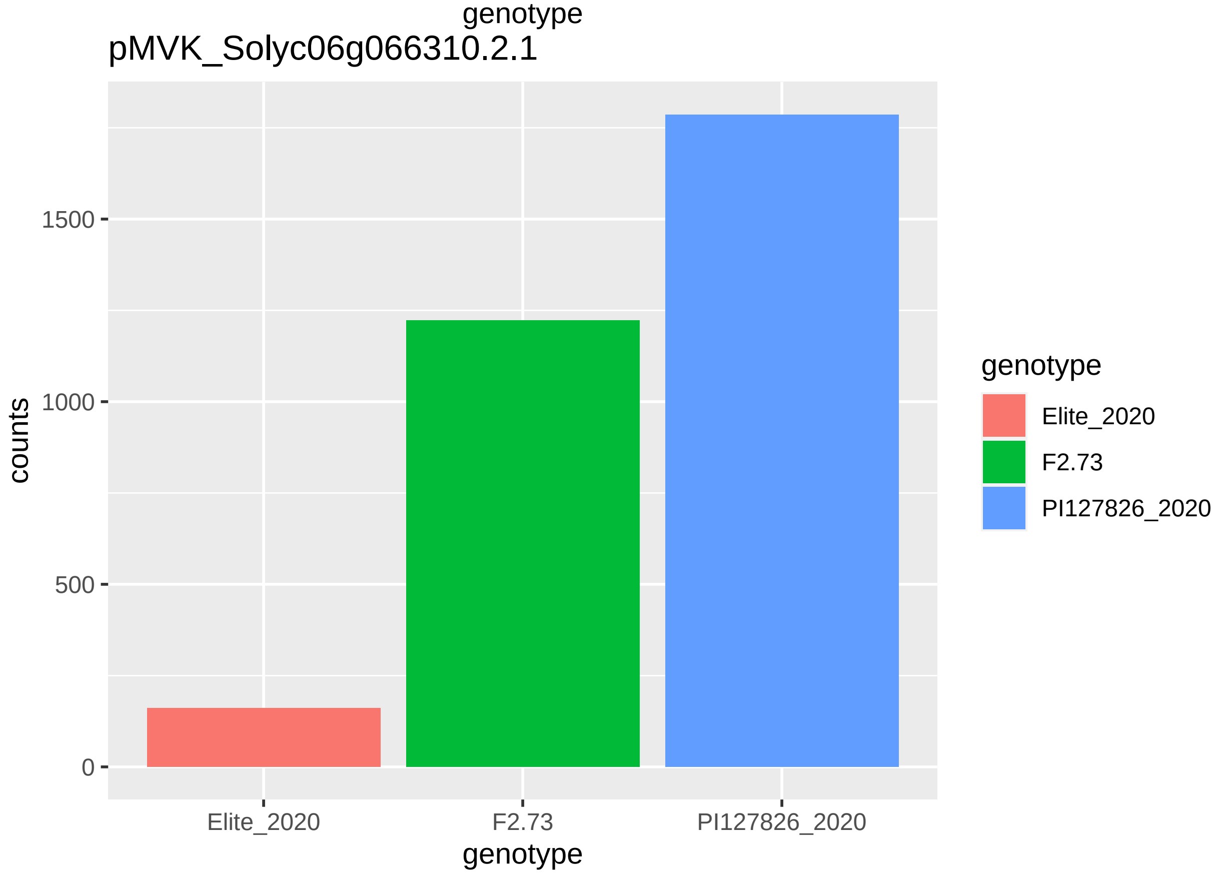

MDC_Solyc04g009650.2.1
counts
900
600
300
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


AACT_Solyc04g015100.2.1
counts
3000
2000
1000
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

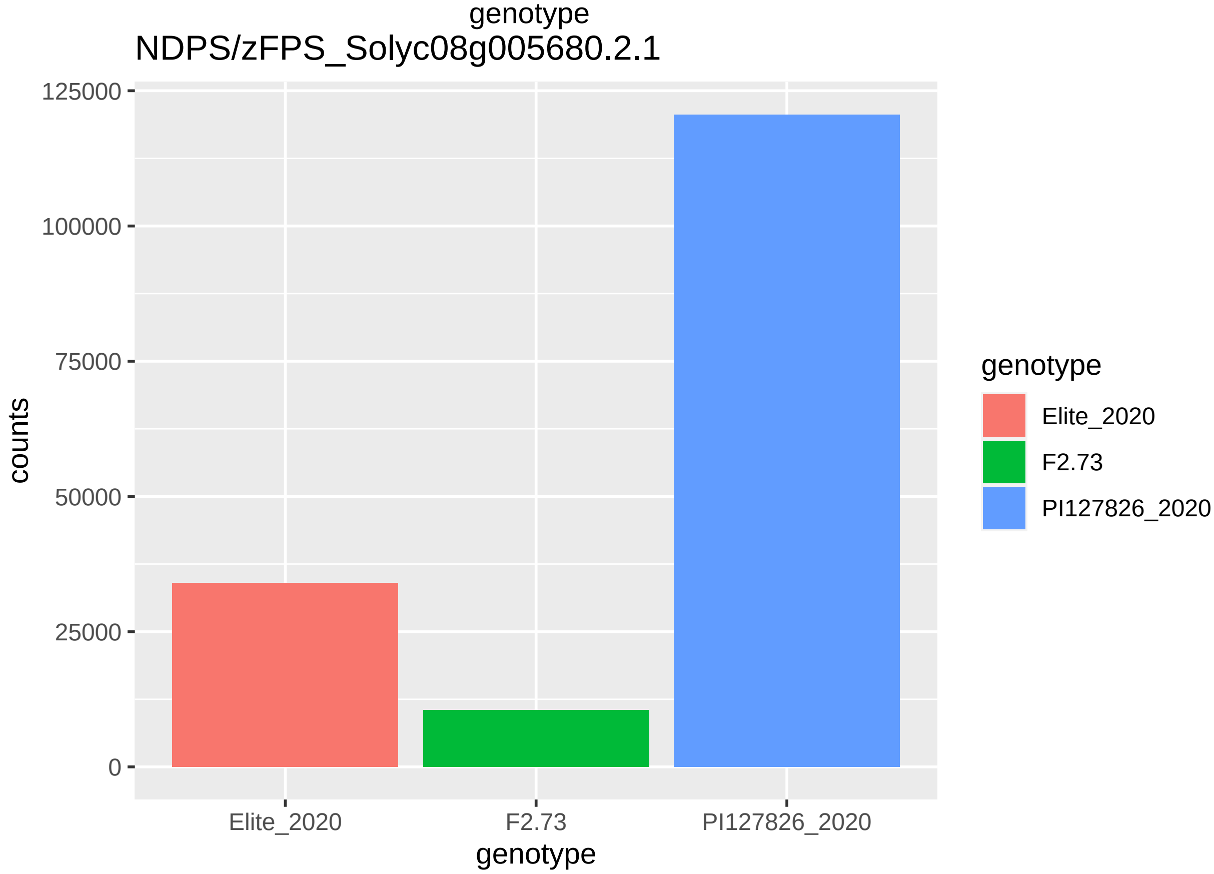

IDI1_Solyc04g056390.2.1
20000
15000
10000
5000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

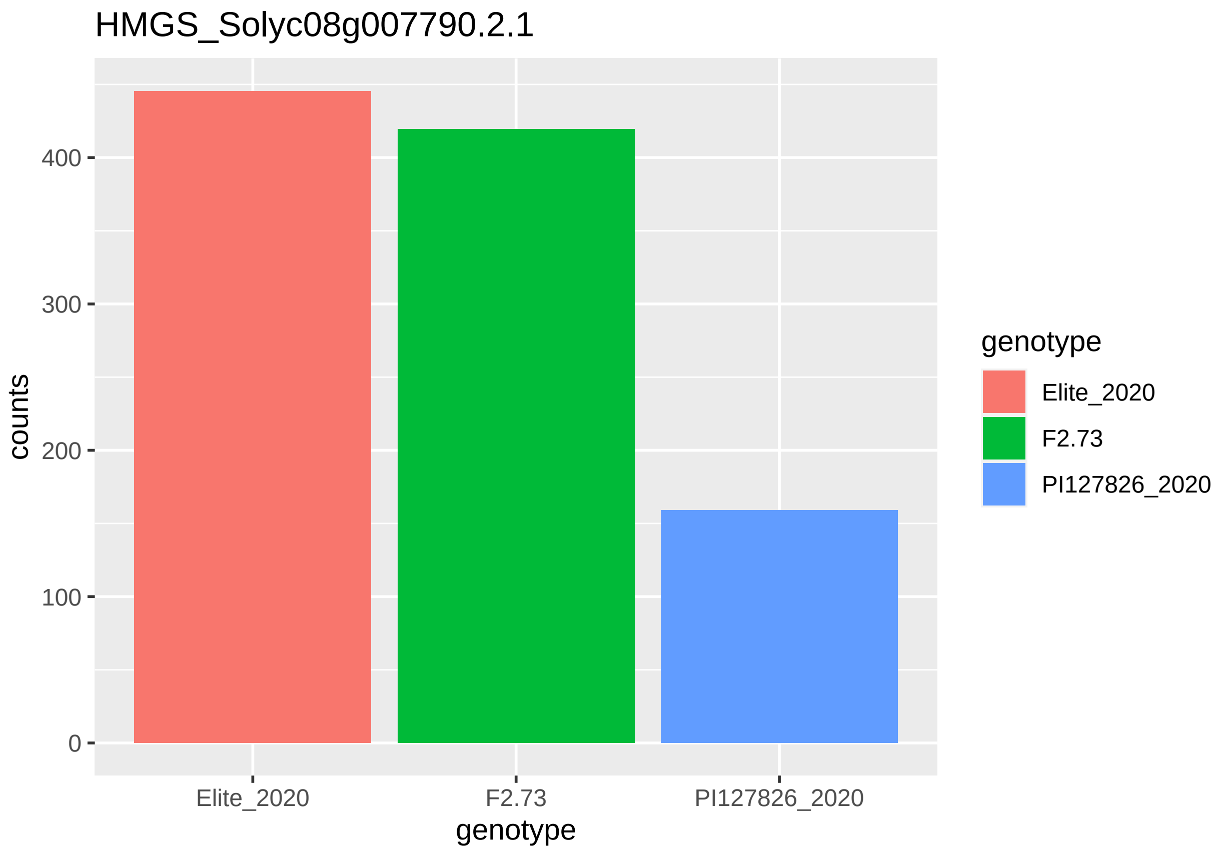

AACT_Solyc05g017760.2.1
3000
2000
1000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

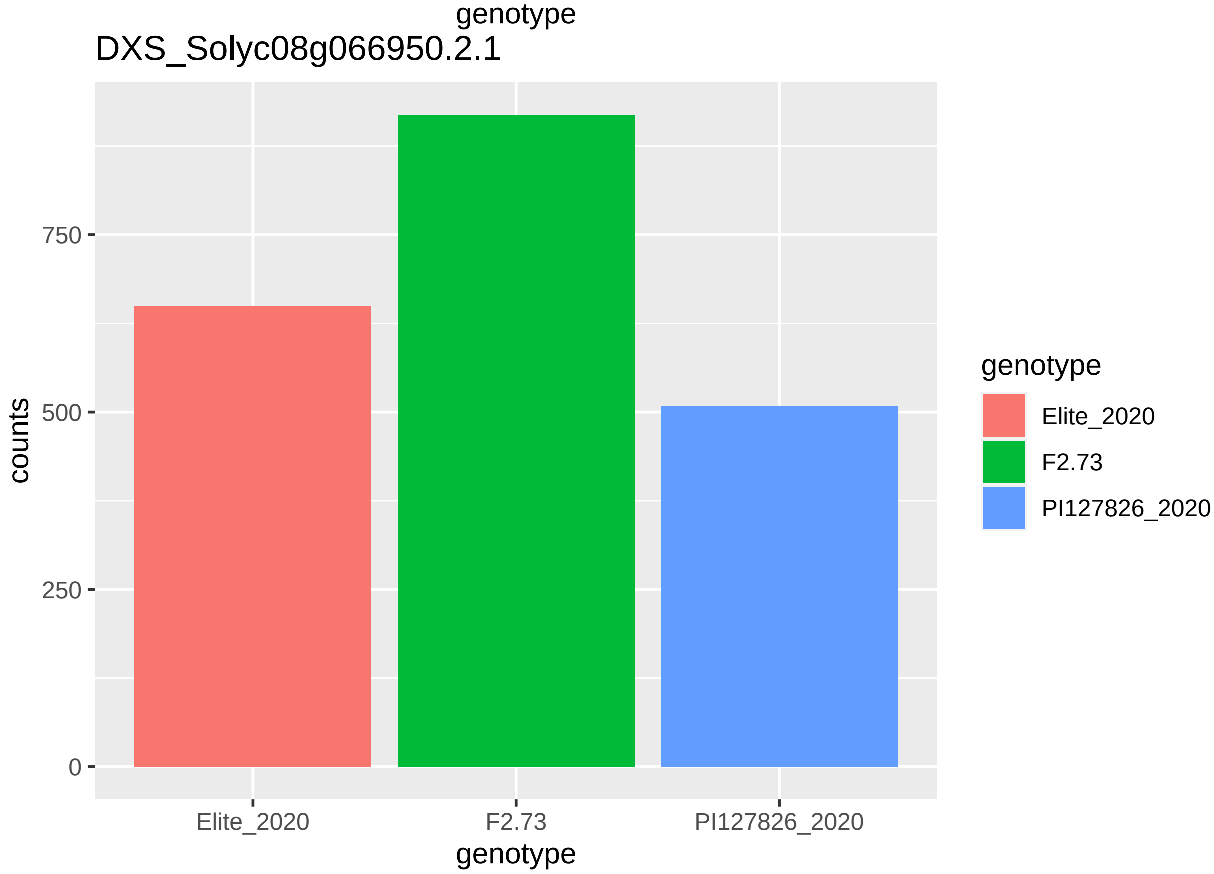

IDI2_Solyc05g055760.2.1
counts
1500
1000
500
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

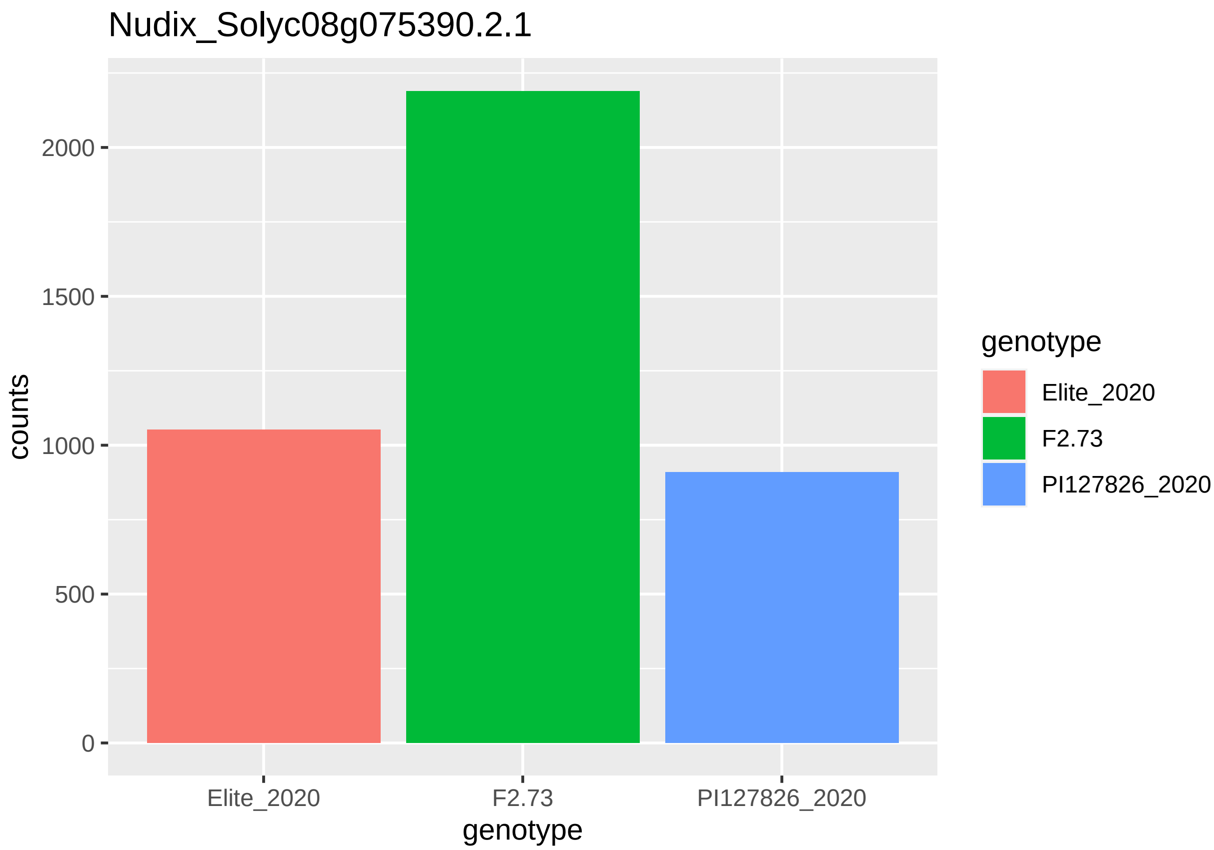

pMVK_Solyc06g066310.2.1
counts
1500
1000
500
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

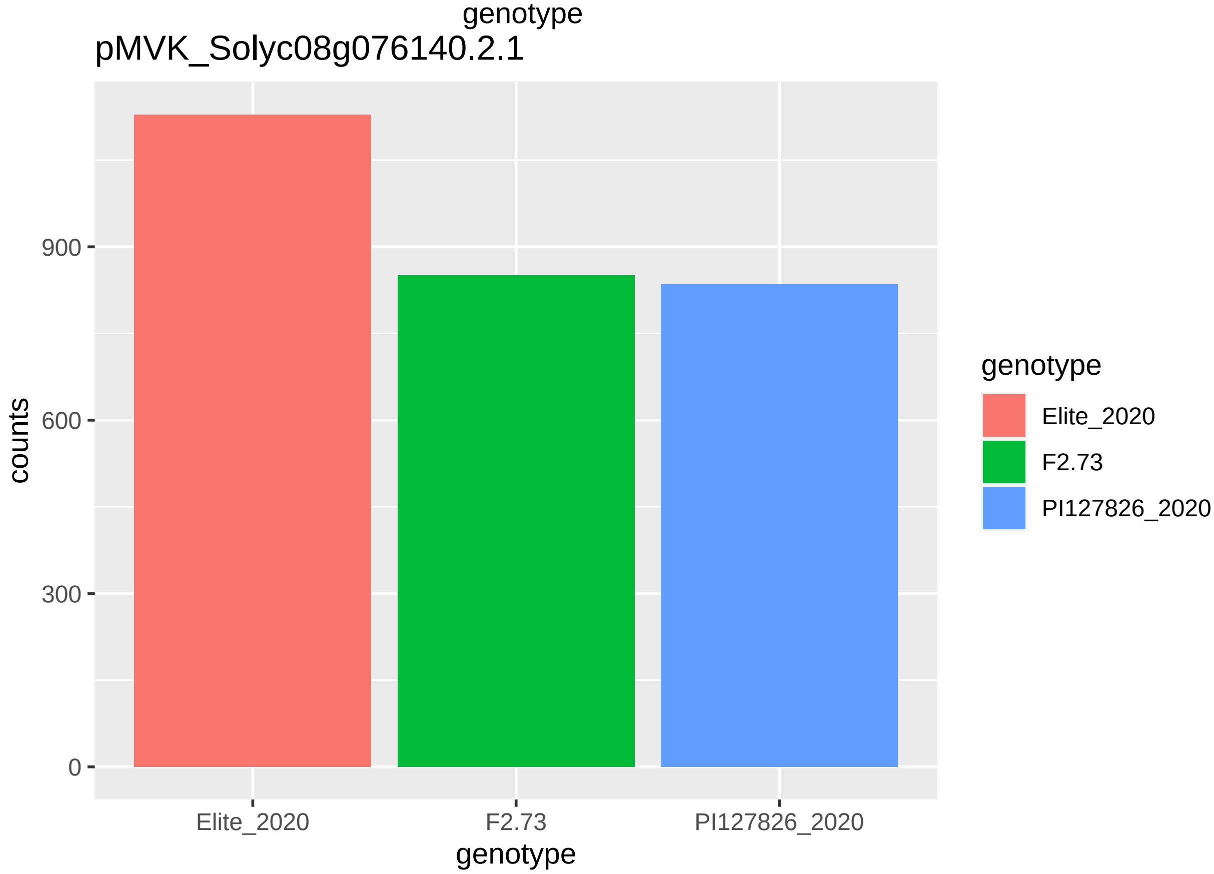

AACT_Solyc07g045350.2.1
15000
10000
5000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

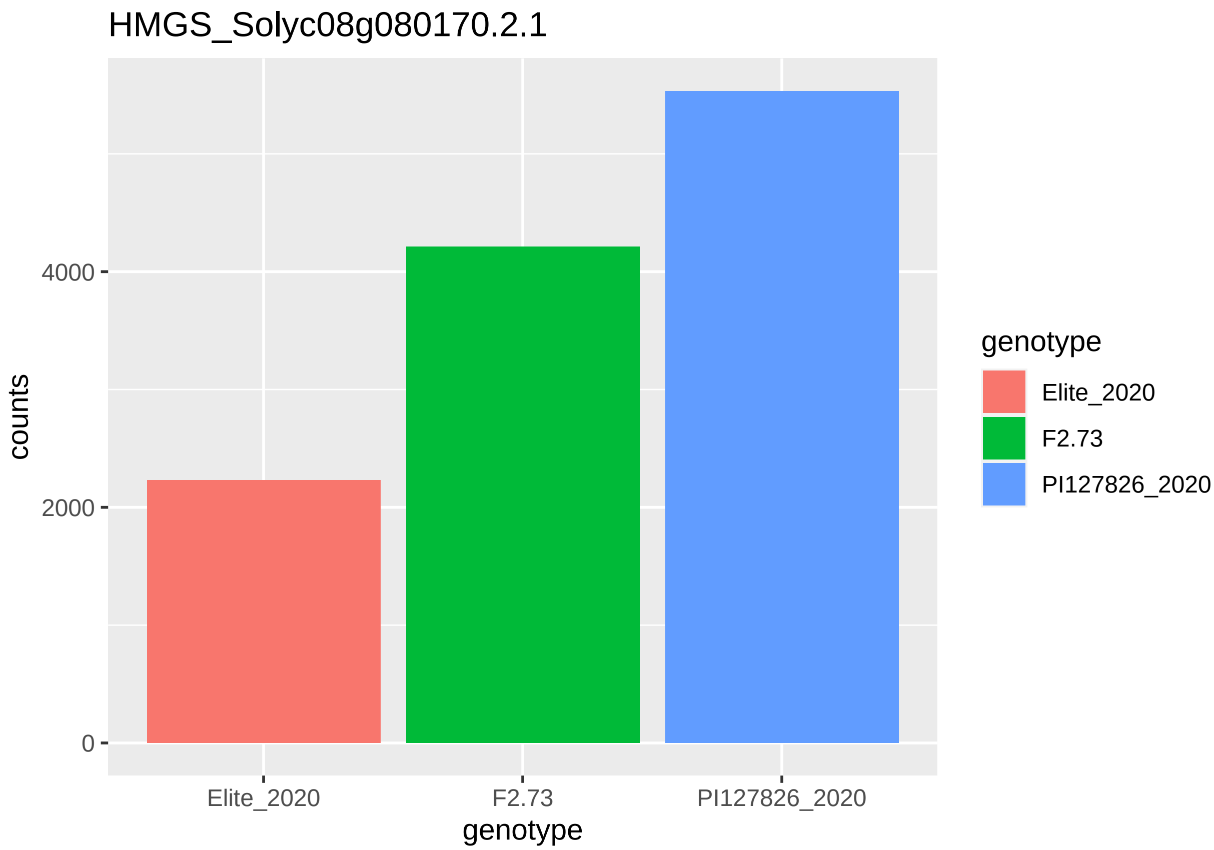

NDPS/zFPS_Solyc08g005680.2.1
125000
100000
75000
50000
25000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

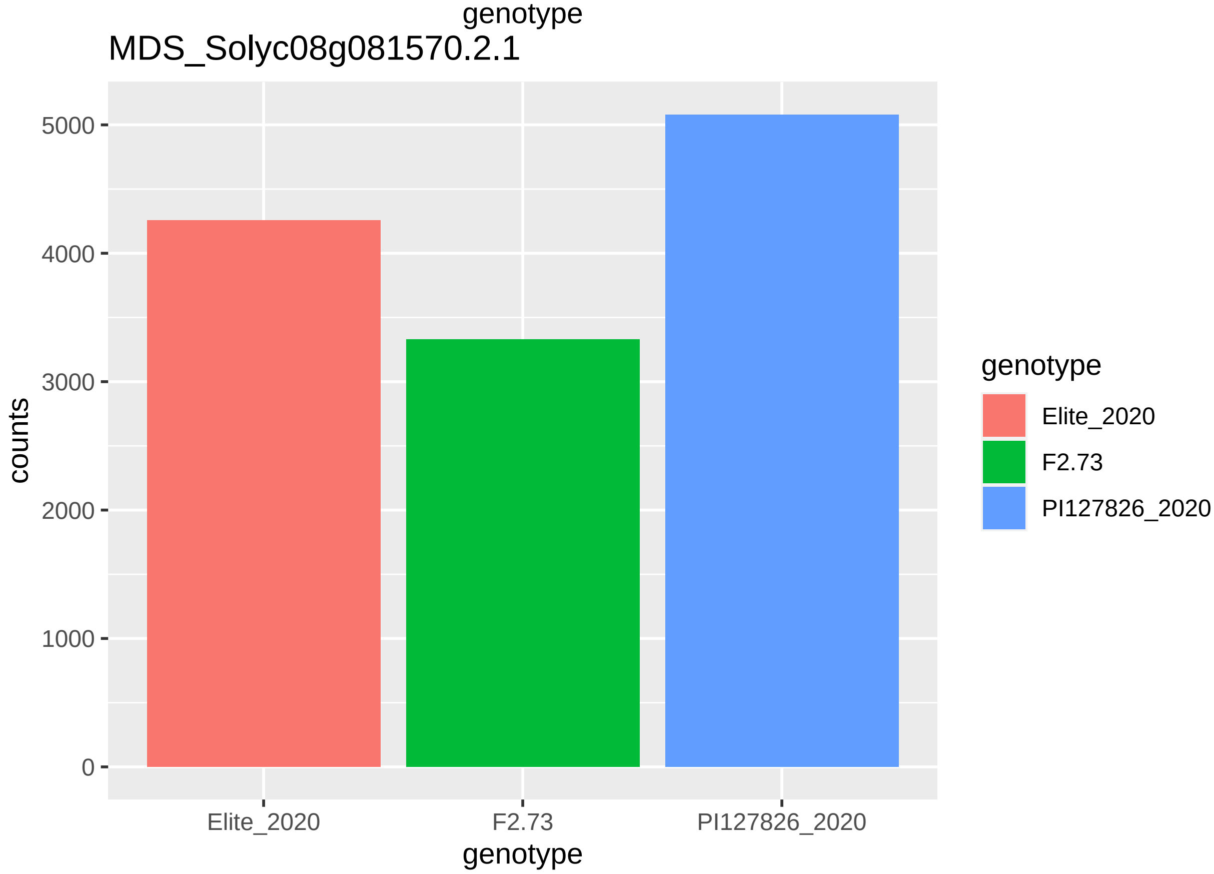

HMGS_Solyc08g007790.2.1
counts
400
300
200
100
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


DXS_Solyc08g066950.2.1
counts
750
500
250
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

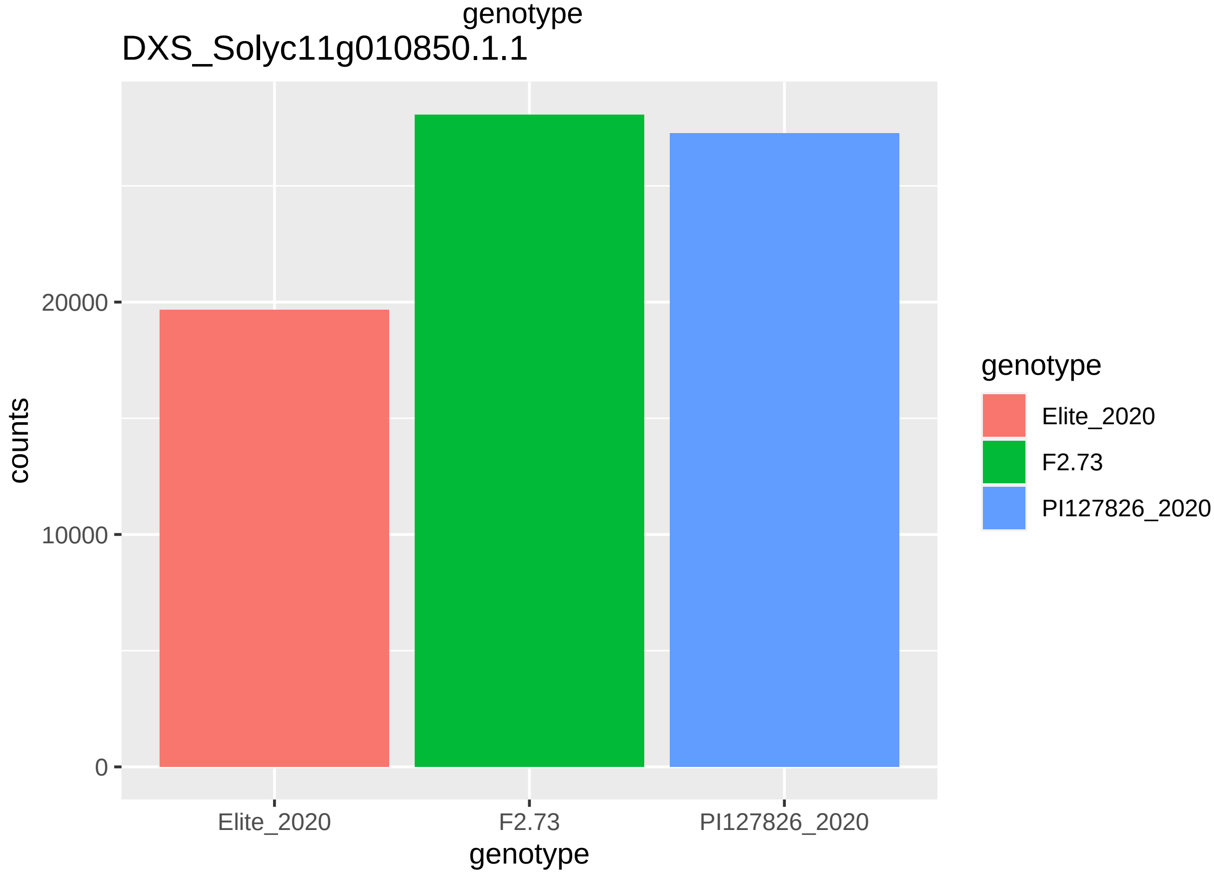

Nudix_Solyc08g075390.2.1
counts
2000
1500
1000
500
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


pMVK_Solyc08g076140.2.1
counts
900
600
300
0
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


HMGS_Solyc08g080170.2.1
counts
0
2000
4000
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


MDS_Solyc08g081570.2.1
counts
0
1000
2000
3000
4000
5000
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


MDC_Solyc11g007020.1.1
5000
4000
3000
2000
1000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020


DXS_Solyc11g010850.1.1
20000
10000
0
counts
Elite_2020
F2.73
PI127826_2020
genotype
genotype
Elite_2020
F2.73
PI127826_2020

HDS_Solyc11g069380.1.1

HMGS_Solyc12g056450.1.1