Differential expression (F2-127, Elite and PI127826 2020 plants)

Marc Galland

10/17/2021

Contents

1 Data import **2**
1.1 Import scaled counts . . . 2

2 QC plots **3**
2.1 Plot density counts . . . 3
2.2 Plot log2ratio F2.127 vs Elite . . . 3
2.3 Plot log2ratio PI127826 vs Elite . . . 4

3 Compute DE genes based on log2ratio Z-score **6**
3.1 Calculate log2ratios . . . 6
3.2 Calculate Z-scores and associated p-values . . . 6
3.3 Add original counts and annotations . . . 7
3.4 Write to CSV file . . . 7

4 MEP and MVA pathway gene analysis **8**
4.1 Import MEP and MVA gene identifiers . . . 8
4.2 Filter for significant DE genes . . . 8
4.3 Keep only MEP and MVA genes significant . . . 8
4.4 Plot all MEP and MVA genes . . . 9

1 Data import

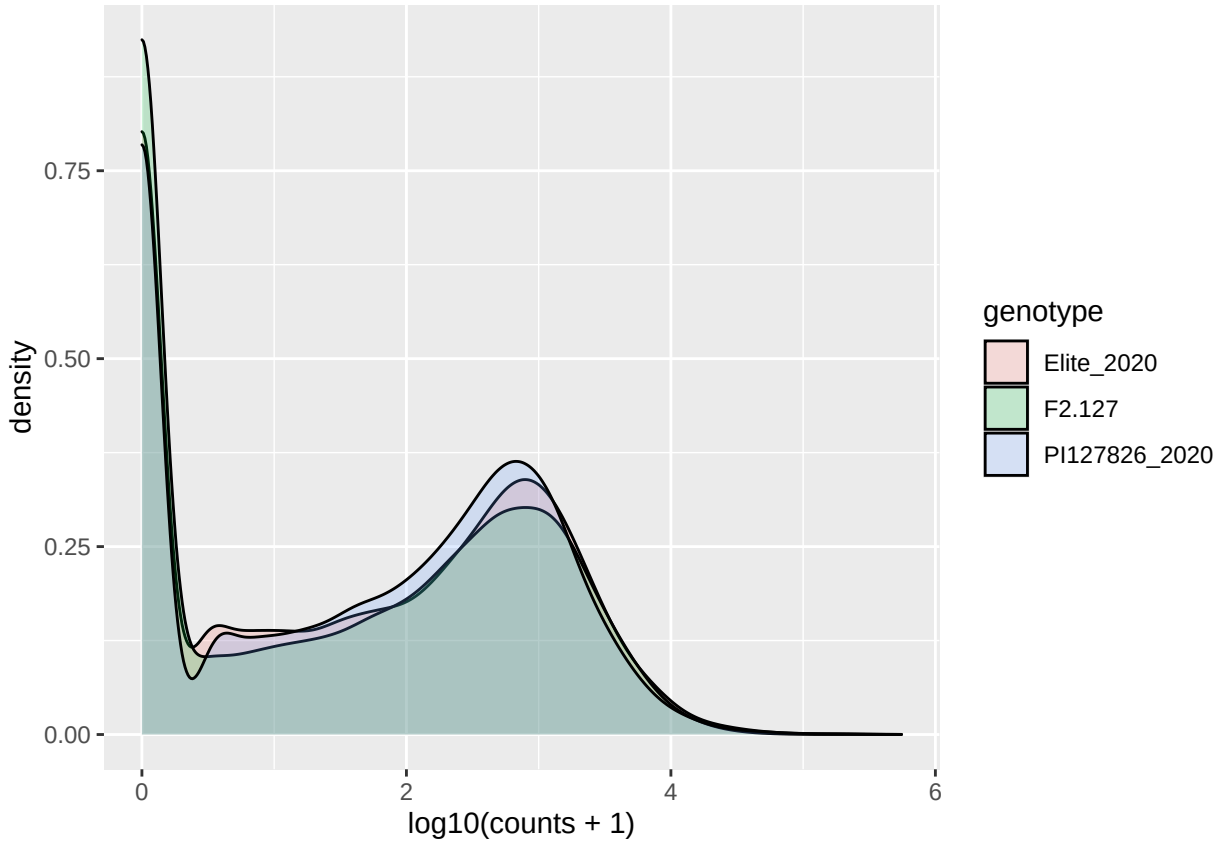
1.1 Import scaled counts

```
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", "F2.127", "Elite_2020", "PI127826_2020")
```

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 = 0.2)
```



The density of genes woth low count values (log10 = 0.3) is lower for F2.127 but seems to be comparable for other genes.

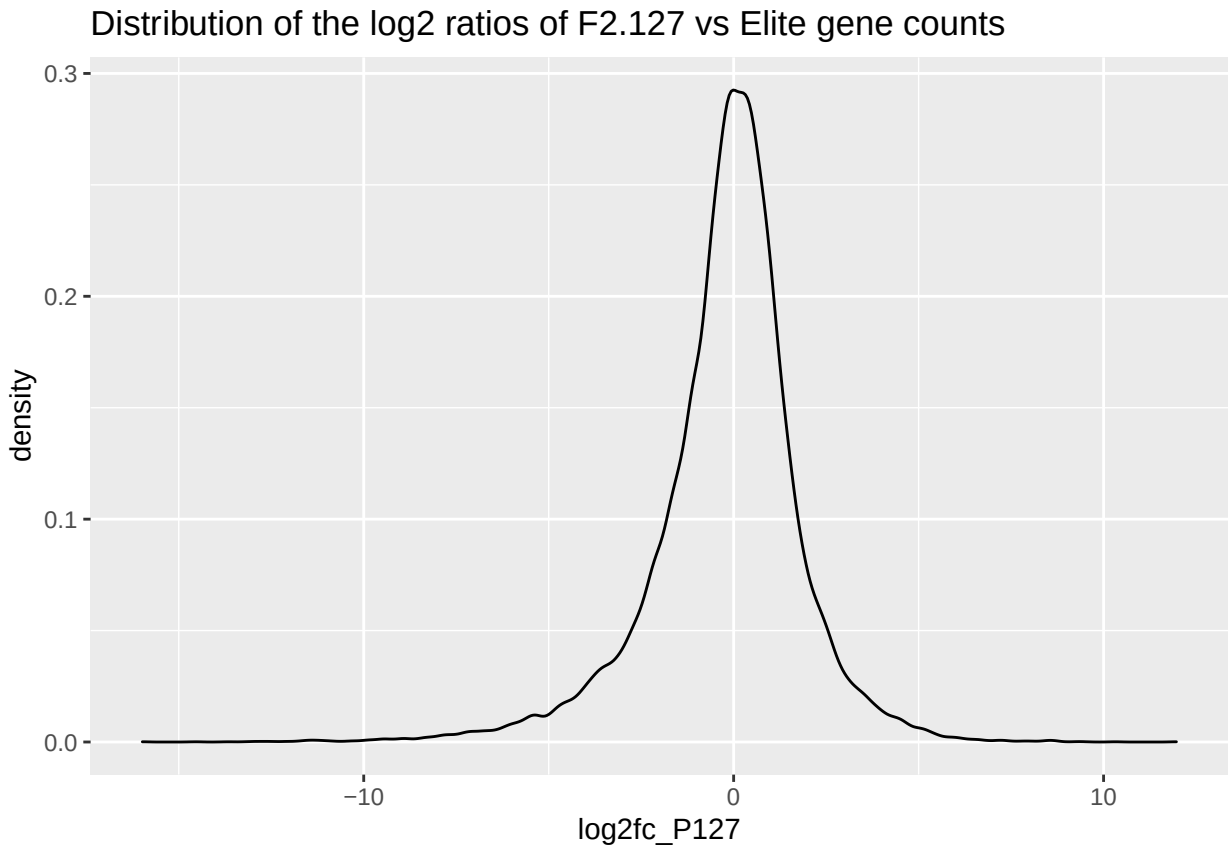
2.2 Plot log2ratio F2.127 vs Elite

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_P127 = log2(`F2.127`/Elite_2020)) %>%
  ggplot(aes(x = log2fc_P127)) +
  geom_density() +
  ggtitle("Distribution of the log2 ratios of F2.127 vs Elite gene counts")
```

```
## Warning: Removed 5208 rows containing non-finite values (stat_density).
```

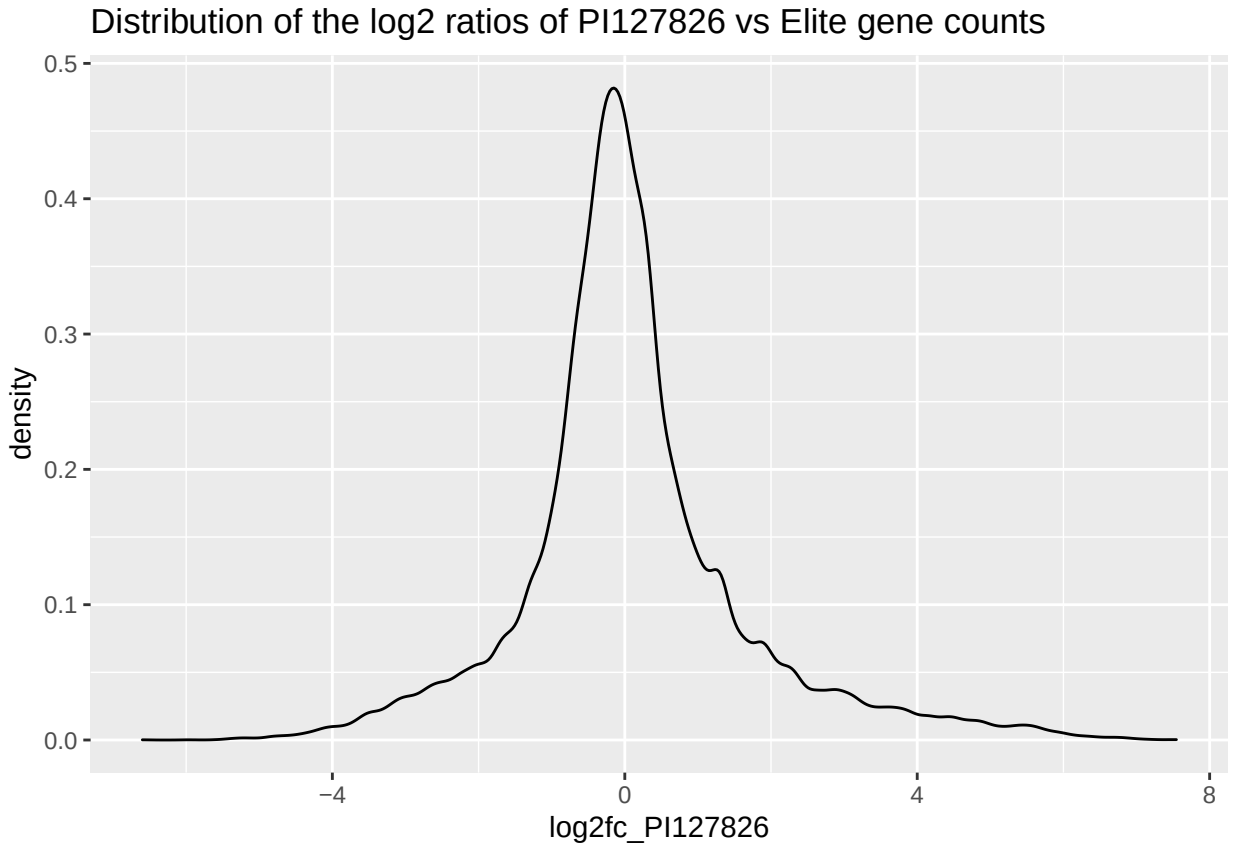


The log2ratio distribution looks OK.

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 3768 rows containing non-finite values (stat_density).
```


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

3 Compute DE genes based on log2ratio Z-score

3.1 Calculate log2ratios

A positive log2ratio for F2.127 and PI127826 means that the gene is more expressed in F2.127 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_P127 = log2(`F2.127`/Elite_2020)) %>%
  mutate(log2ratio_PI127826 = log2(PI127826_2020/Elite_2020)) %>%
  select(gene, log2ratio_P127, log2ratio_PI127826) %>%
  filter(!grepl(pattern = "Inf", x = log2ratio_P127)) %>%
  filter(!grepl(pattern = "Inf", x = log2ratio_PI127826))

head(log2ratio)
```

```
##                 gene log2ratio_P127 log2ratio_PI127826
## 1 Solyc00g005040.2.1    -0.79548801         -1.8945217
## 2 Solyc00g005050.2.1     0.05728413          0.3874636
## 3 Solyc00g005840.2.1    -0.16177155          0.6485176
## 4 Solyc00g005860.1.1    -4.45183514         -3.0148159
## 5 Solyc00g006470.1.1    -8.92408041          1.7963727
## 6 Solyc00g006490.2.1     0.05487591          0.1157466
```

A total of **21036** have a finite log2ratio in both F2.127 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_P127 = scale(log2ratio_P127, center = T, scale = T)) %>%
  mutate(zscore_PI127826 = scale(log2ratio_PI127826, center = T, scale = T)) %>%
  mutate(pval_P127 = pnorm(q = abs(zscore_P127), 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_P127)) %>%
  as_tibble()

head(log2ratio_zscores_pvals)
```

```
## # A tibble: 6 x 7
##   gene  log2ratio_P127 log2ratio_PI127~ zscore_P127[,1] zscore_PI127826~
##   <chr>          <dbl>            <dbl>           <dbl>            <dbl>
## 1 Soly~           9.43            0.500            4.83            0.276
## 2 Soly~           9.25            5.26             4.74            3.50
## 3 Soly~           8.70            1.89             4.46            1.22
## 4 Soly~           8.62            0.307            4.42            0.146
## 5 Soly~           8.62            5.17             4.42            3.44
## 6 Soly~           8.61           -0.693            4.42           -0.531
## # ... with 2 more variables: pval_P127[,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] 21036    12
```

3.4 Write to CSV file

```
dir.create(path = "./tables_F2-127/", showWarnings = F, recursive = T)
write.csv(final,
          file = "tables_F2-127/diff_res_F2.127_or_PI127826_vs_Elite.csv",
          row.names = F,
          quote = F)
```

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

:warning: nothing significant at $\mathbf{p < 0.05}$ so I had to "relax" the p-value threshold to $\mathbf{p < 0.1}$.

```
signif_genes <- filter(final, pval_P127 < 0.1) %>% pull(gene)
```

4.3 Keep only MEP and MVA genes significant

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

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

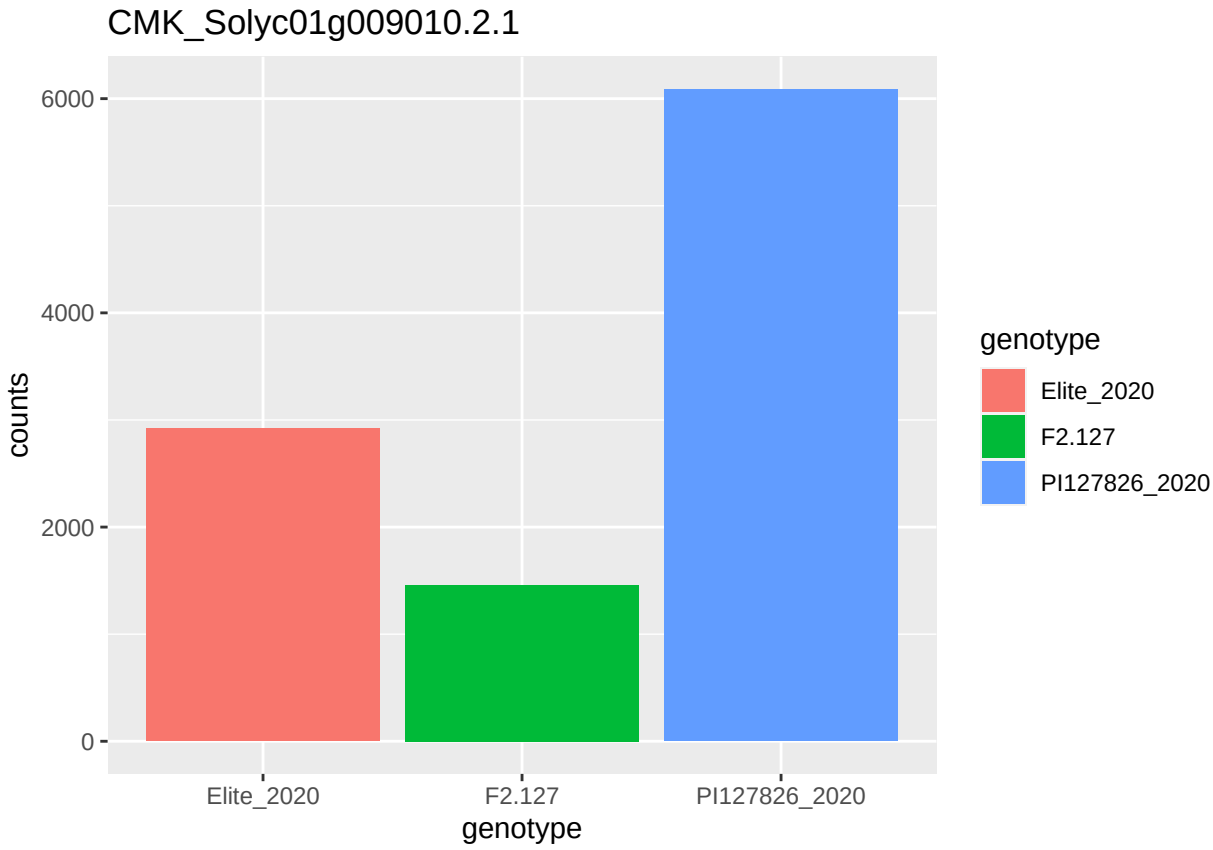
```
mep_mva_gene_signif <-
  mep_mva_genes %>%
  filter(gene %in% signif_genes)

# show table
mep_mva_gene_signif %>%
  select(name, gene, pathway, log2ratio_P127, log2ratio_PI127826, pval_P127, pval_PI127826) %>%
  knitr::kable()
```

name	gene	pathway	log2ratio_P127	log2ratio_PI127826	pval_P127	pval_PI127826
HMGR	Solyc03g032010.2.1	MVA	2.598834	2.9233538	0.07934201	0.02764363
pMVK	Solyc06g066310.2.1	MVA	2.389467	3.4748983	0.09599290	0.01101278
NDPS/zFPS	Solyc08g005680.2.1	MEP	-3.129270	1.8280808	0.07251612	0.11994288
HMGS	Solyc08g080160.2.1	MVA	-3.129907	0.6290403	0.07247217	0.35804197

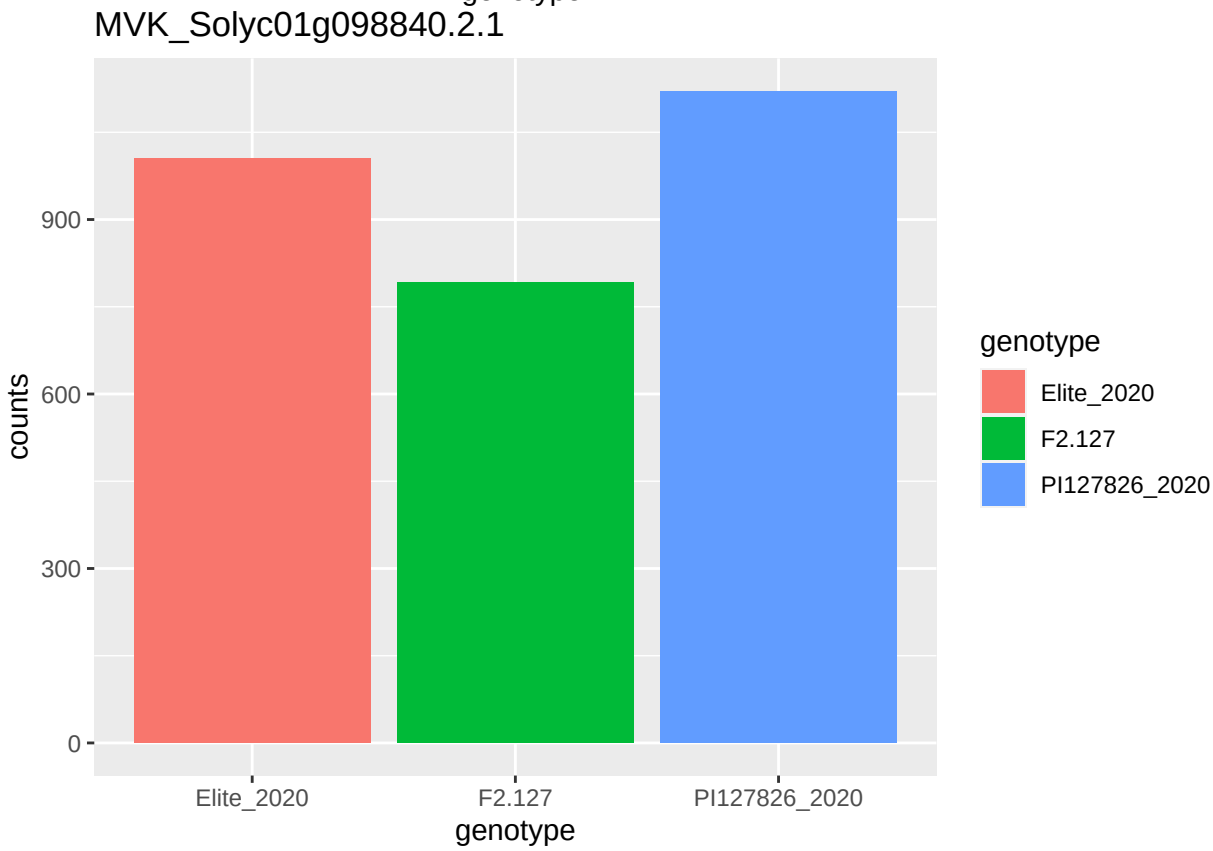
```
write.csv(mep_mva_gene_signif,
            file = "tables_F2-127/mep_mva_gene_signif.csv",
            row.names = F,
            quote = F)
```

4.4 Plot all MEP and MVA genes

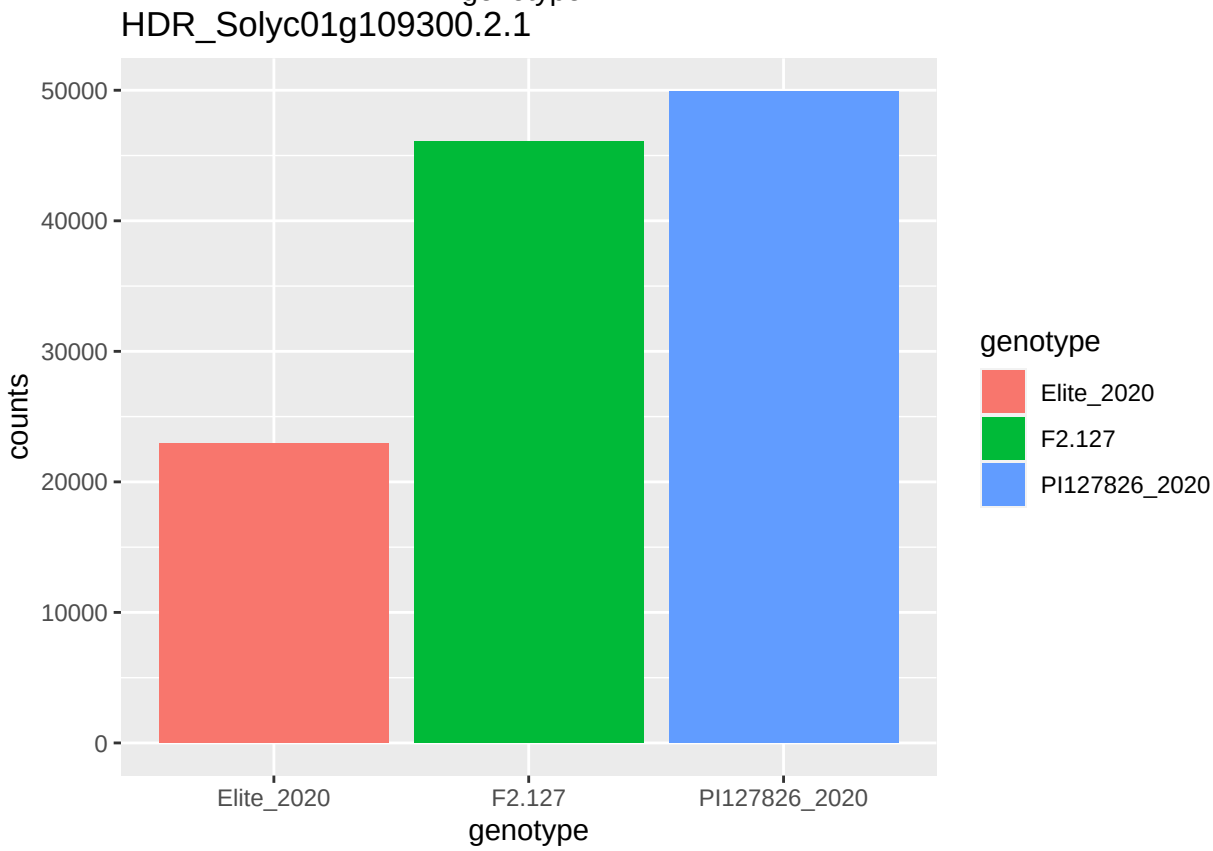
```
for (i in seq_along(mep_mva_genes$gene)){
  tmp_df <- mep_mva_genes[i,]
  tmp_df$title4plot <- paste(tmp_df$name, tmp_df$gene, sep = "_")

  p <-
    tmp_df %>%
    select(title4plot, `F2.127`, 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)
}
```

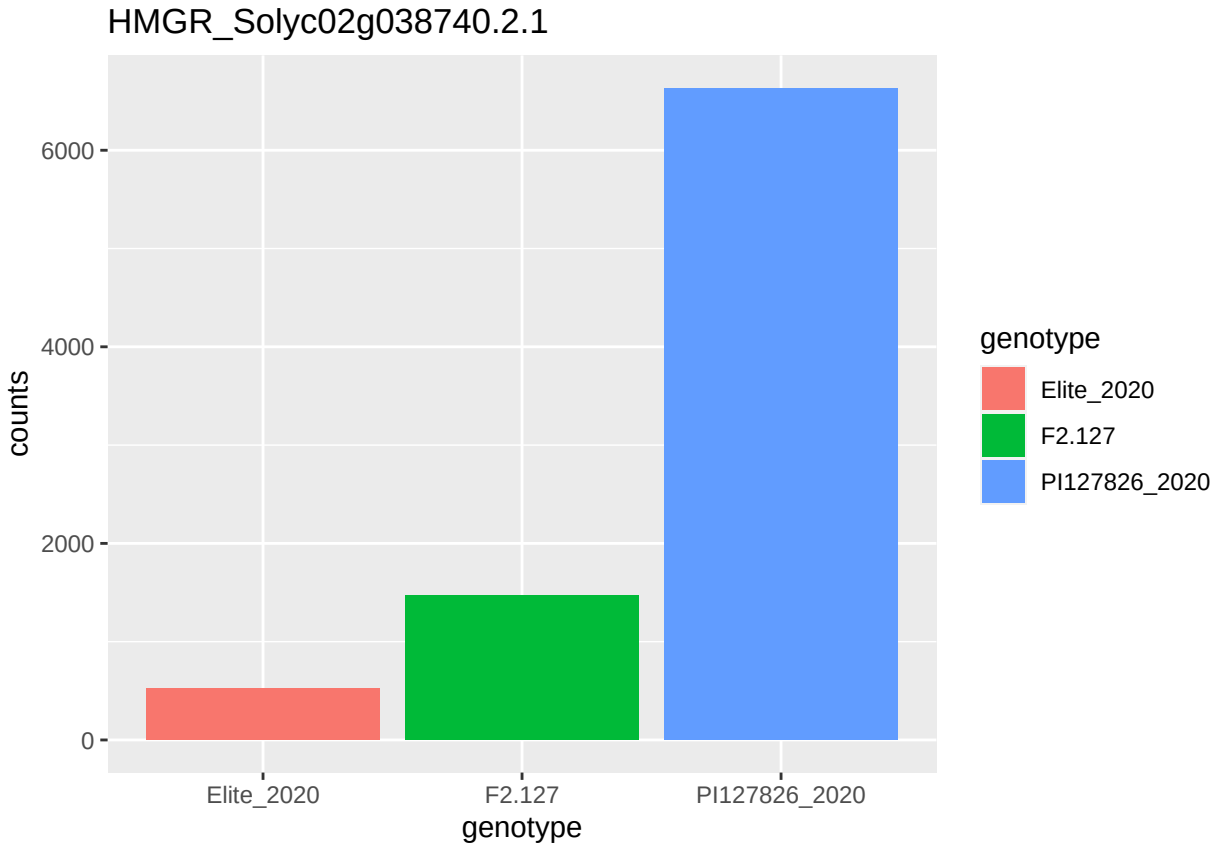


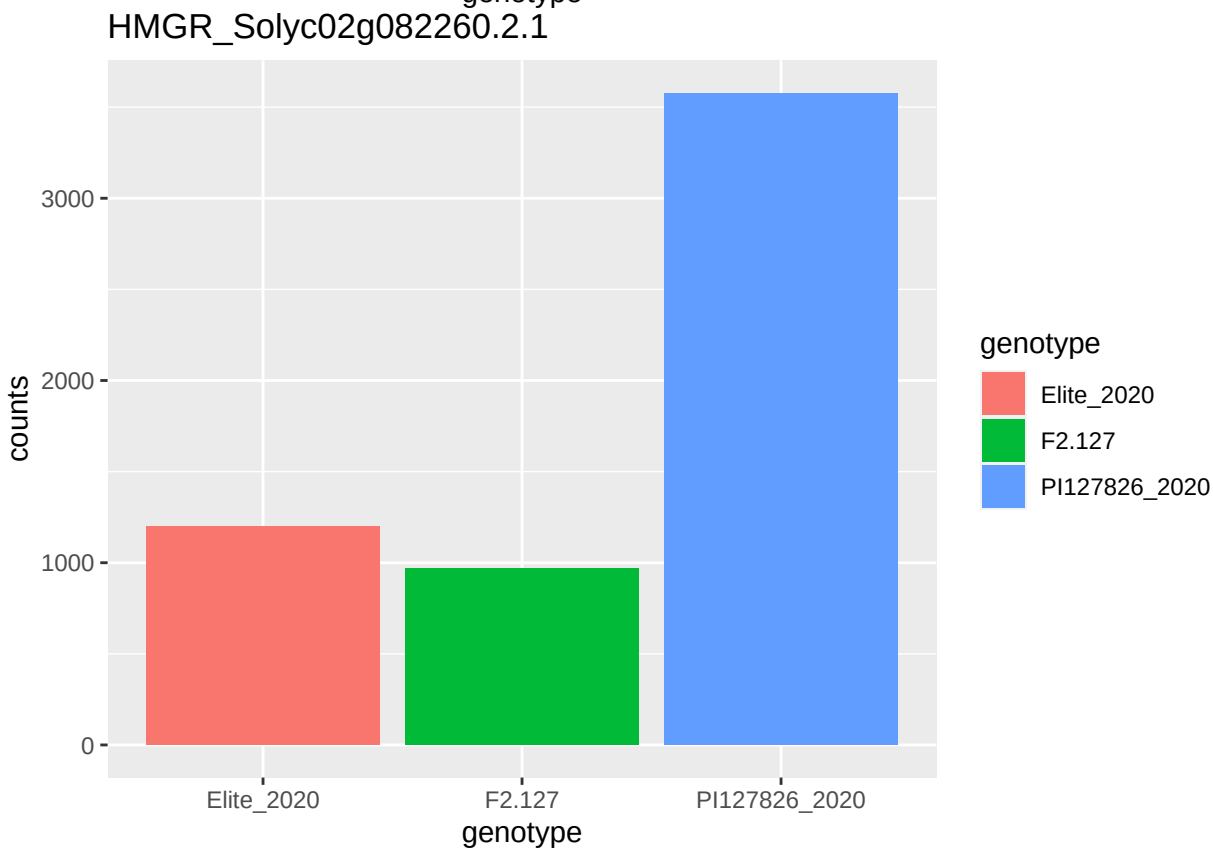

DXS_Solyc01g067890.2.1
counts
4000
3000
2000
1000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


MVK_Solyc01g098840.2.1
counts
900
600
300
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

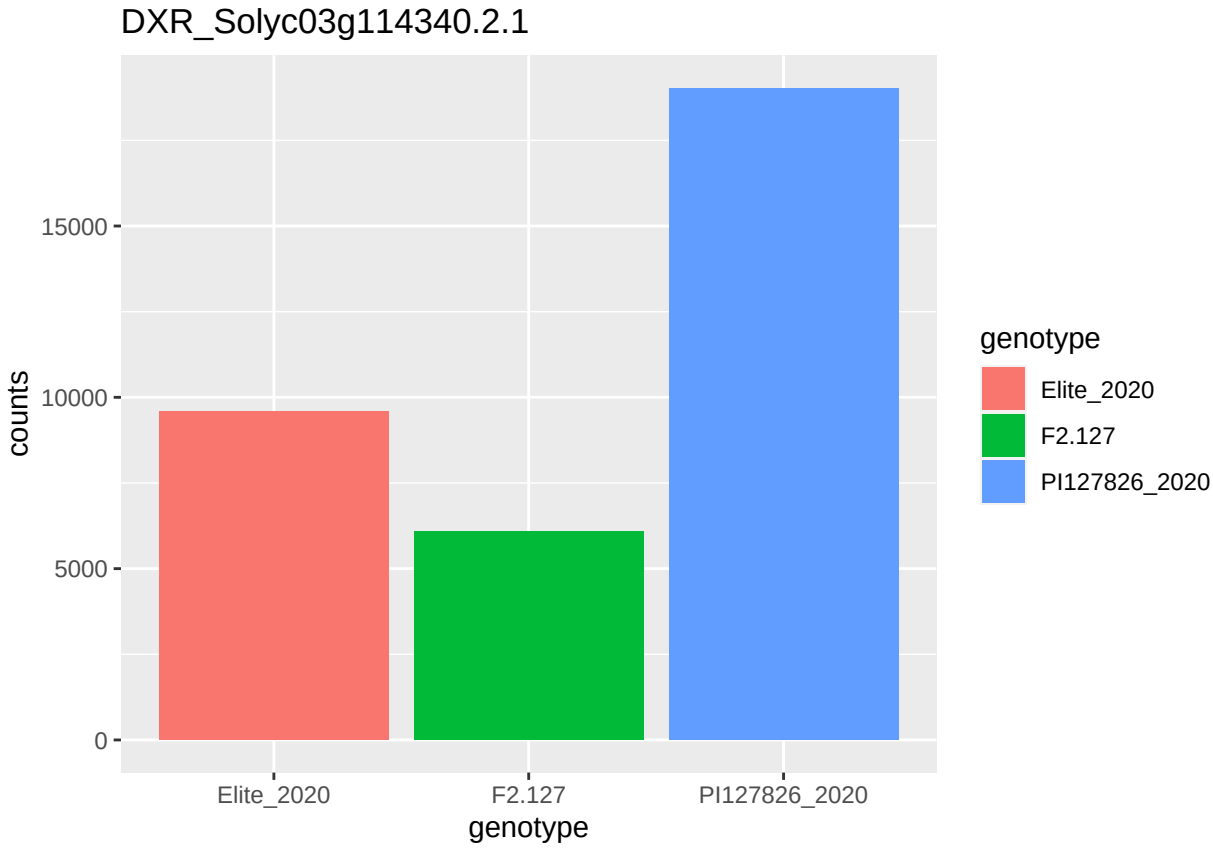

MCT_Solyc01g102820.2.1
2000
1500
1000
500
0
counts
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


HDR_Solyc01g109300.2.1
50000
40000
30000
20000
10000
0
counts
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

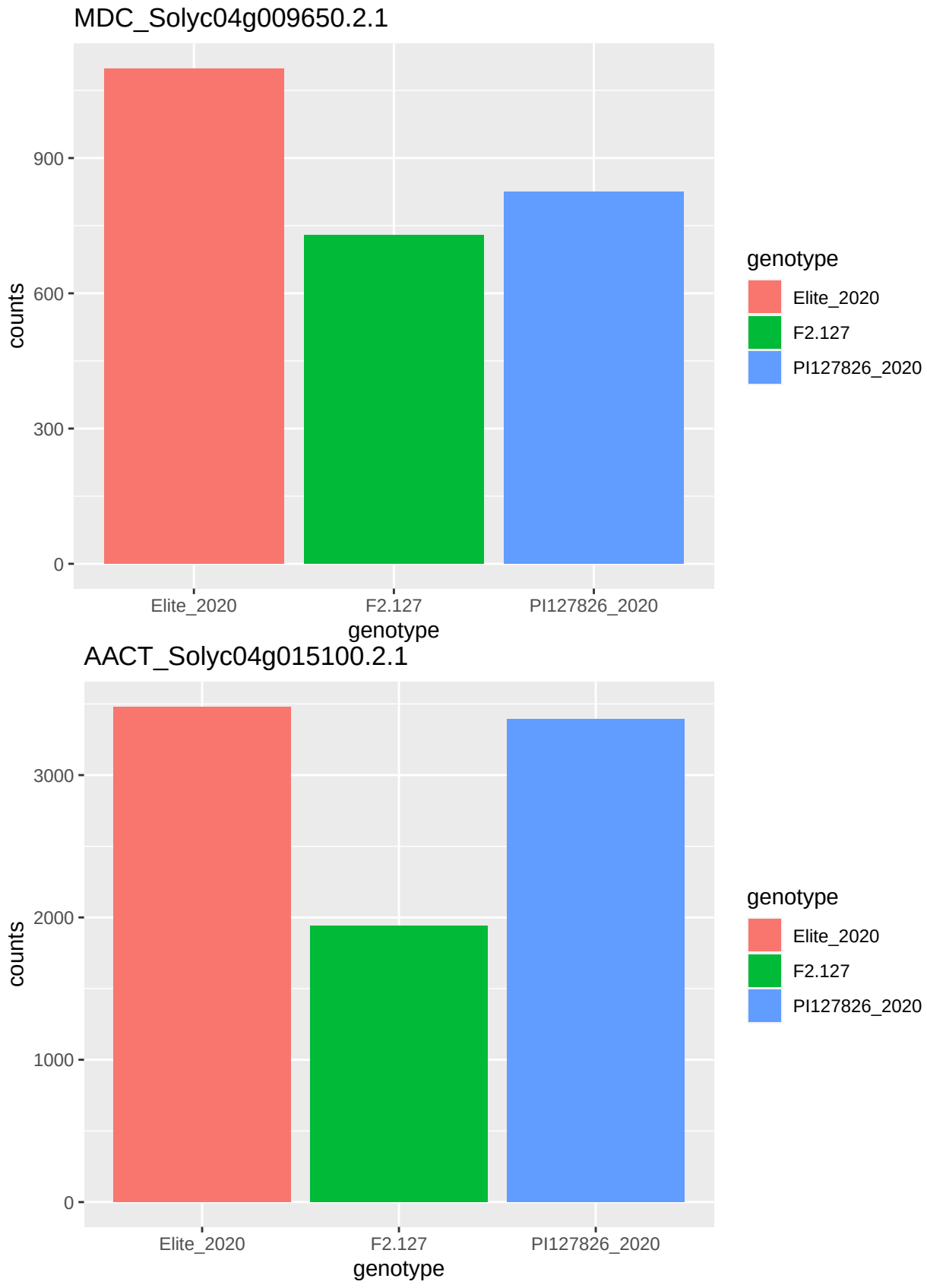

HMGR_Solyc02g038740.2.1
counts
6000
4000
2000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

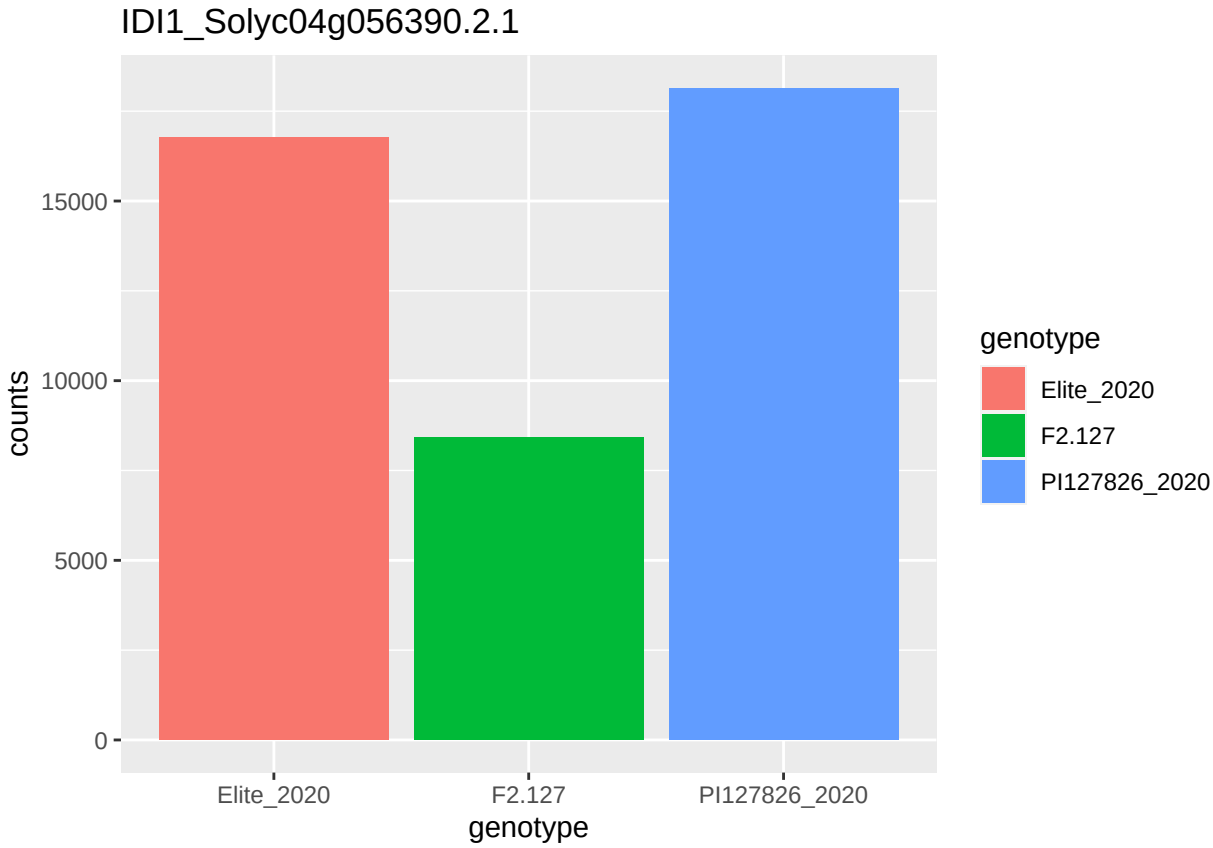

HMGR_Solyc02g082260.2.1
counts
3000
2000
1000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

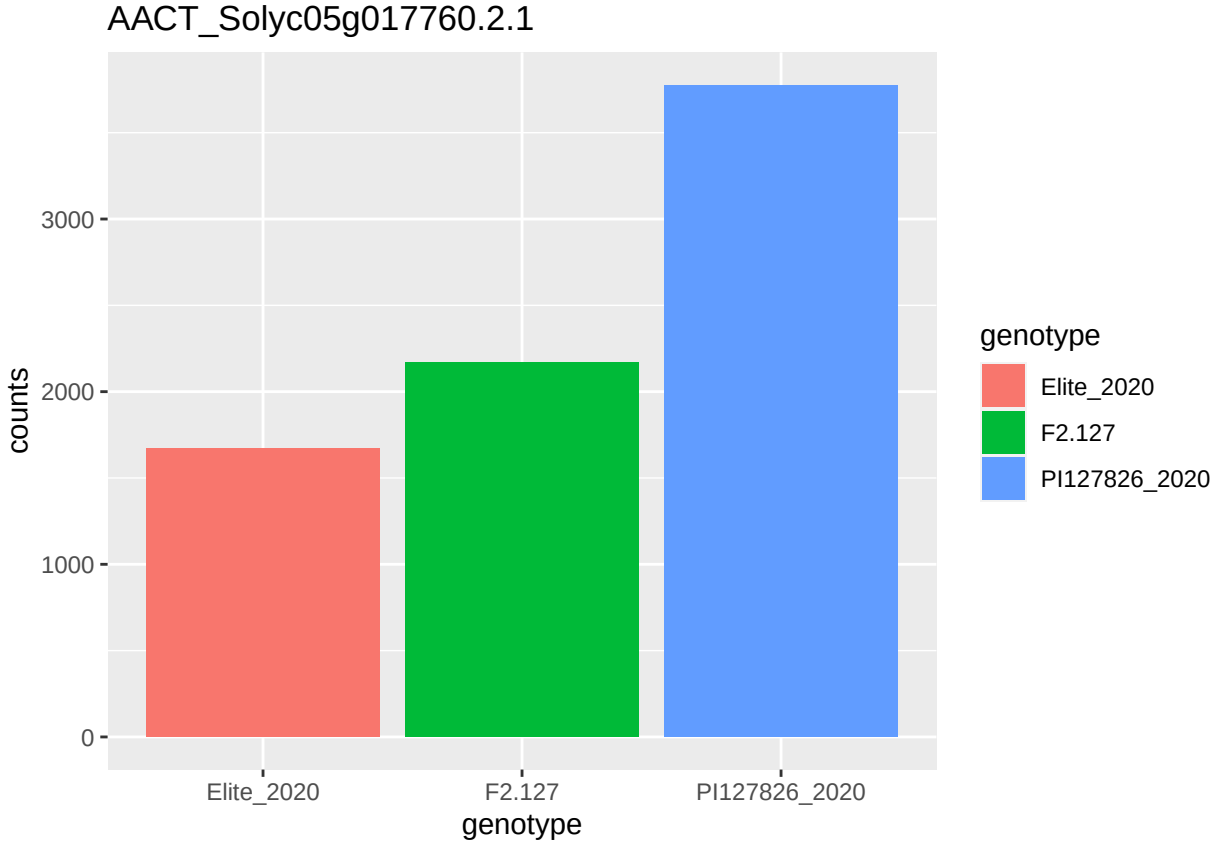

HMGR_Solyc03g032010.2.1
6000
4000
2000
0
counts
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020
HMGR_Solyc03g032020.2.1
4000
3000
2000
1000
0
counts
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

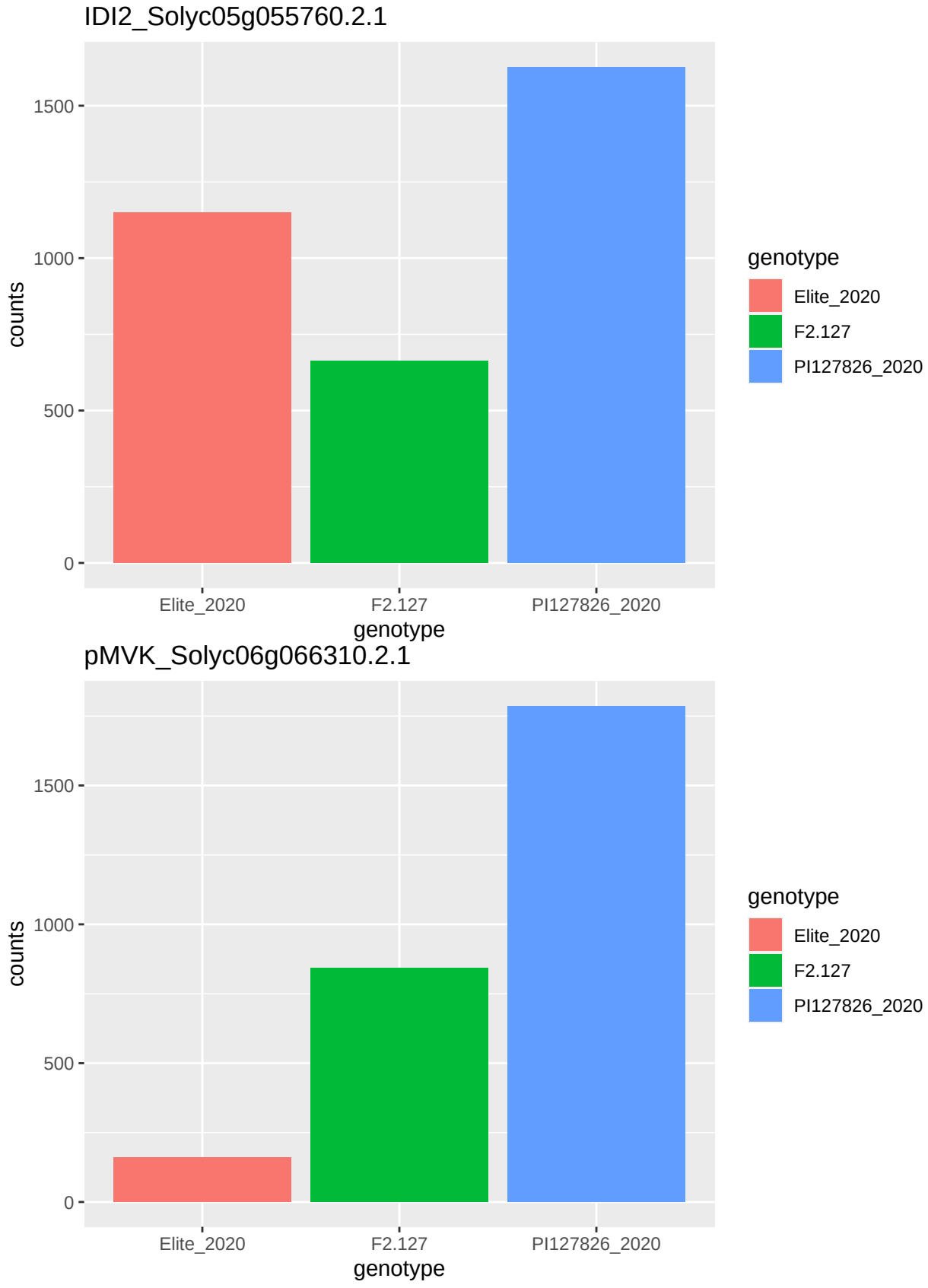

DXR_Solyc03g114340.2.1
counts
15000
10000
5000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

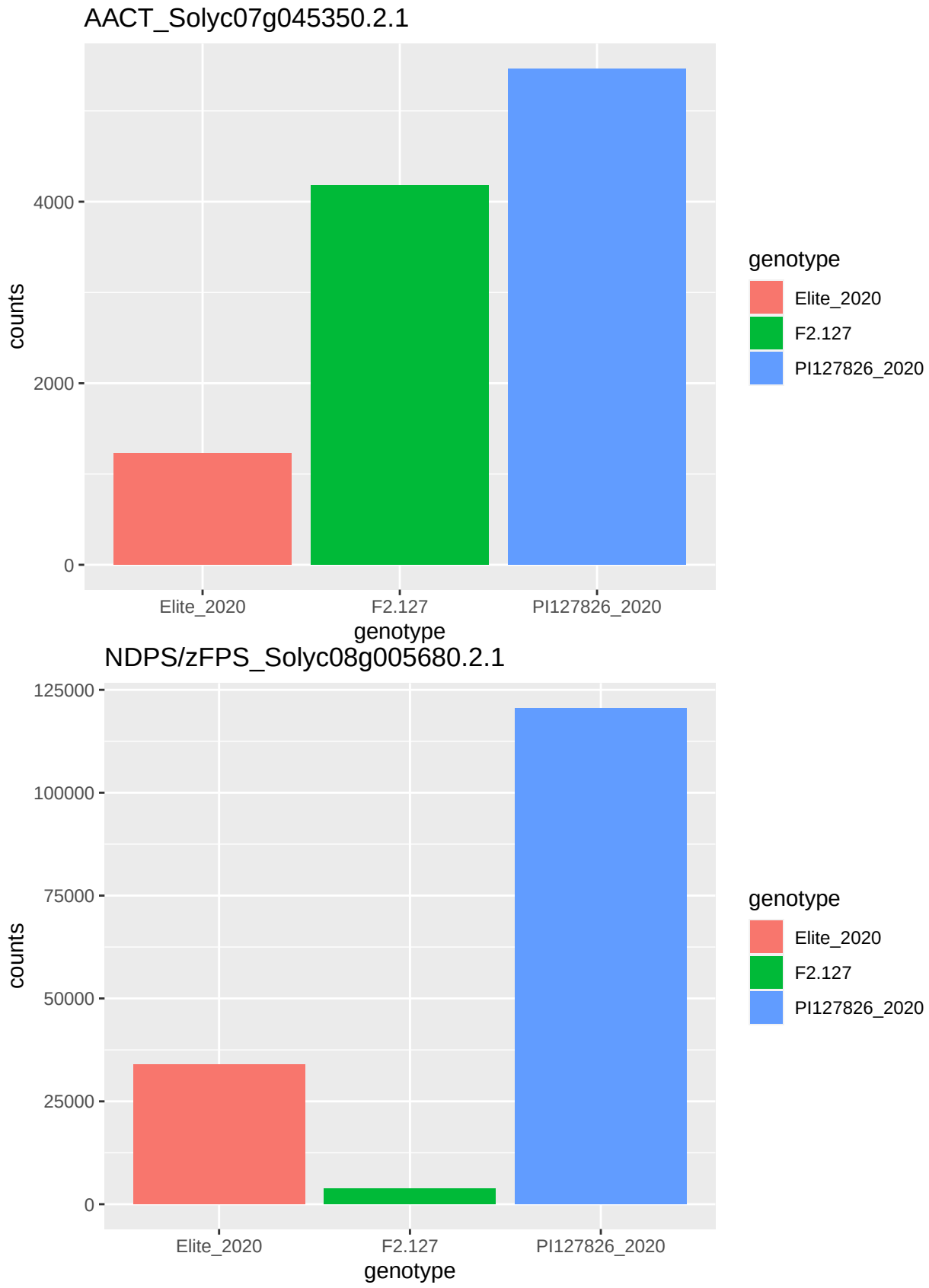

IPK_Solyc04g005520.2.1
counts
400
300
200
100
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

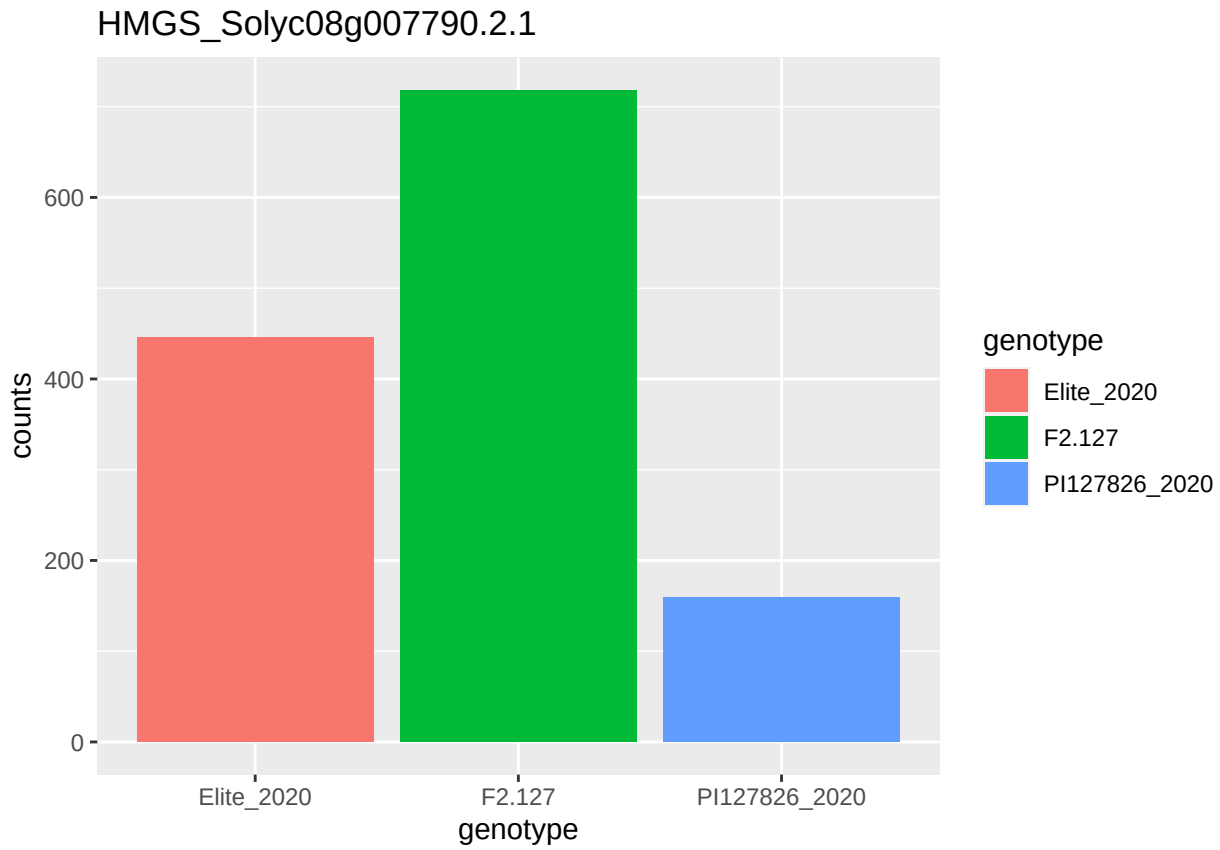

MDC_Solyc04g009650.2.1
900
600
300
0
counts
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020
AACT_Solyc04g015100.2.1
3000
2000
1000
0
counts
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

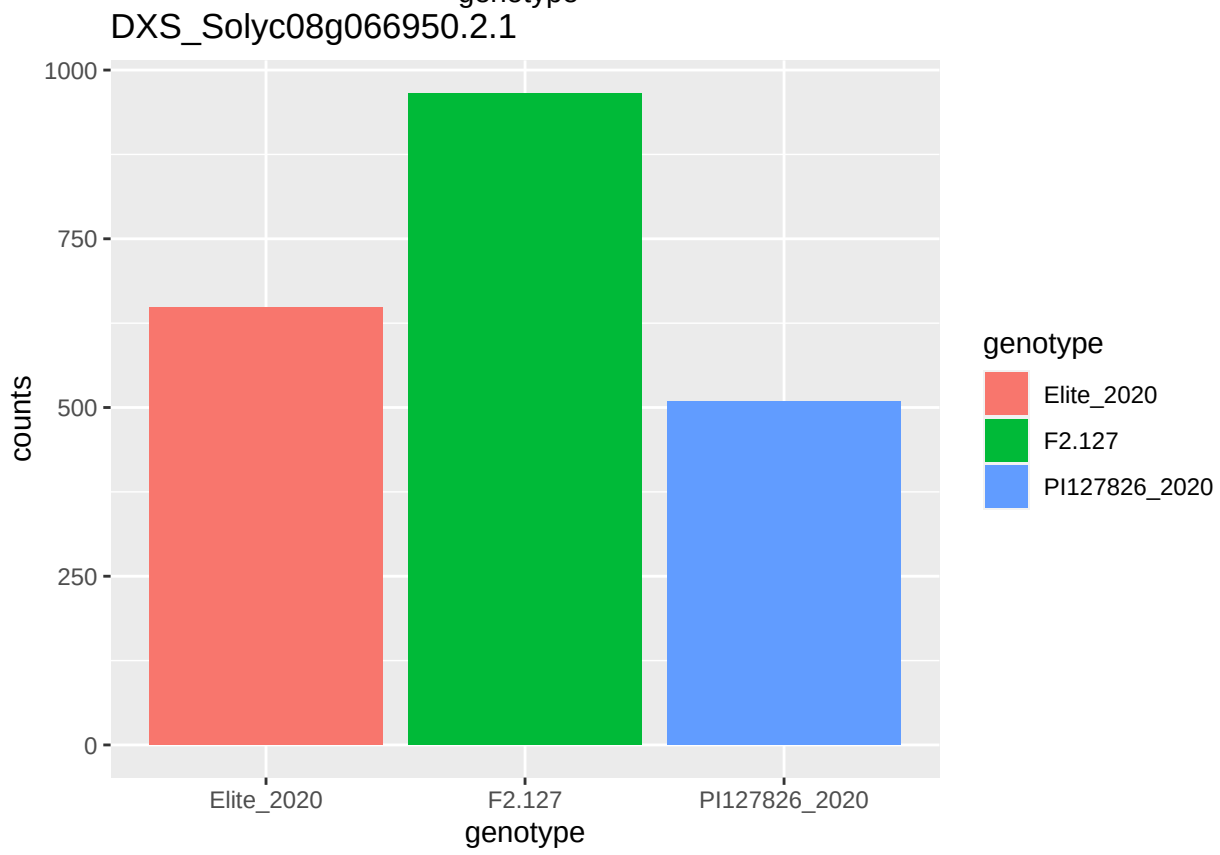

IDI1_Solyc04g056390.2.1
counts
15000
10000
5000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

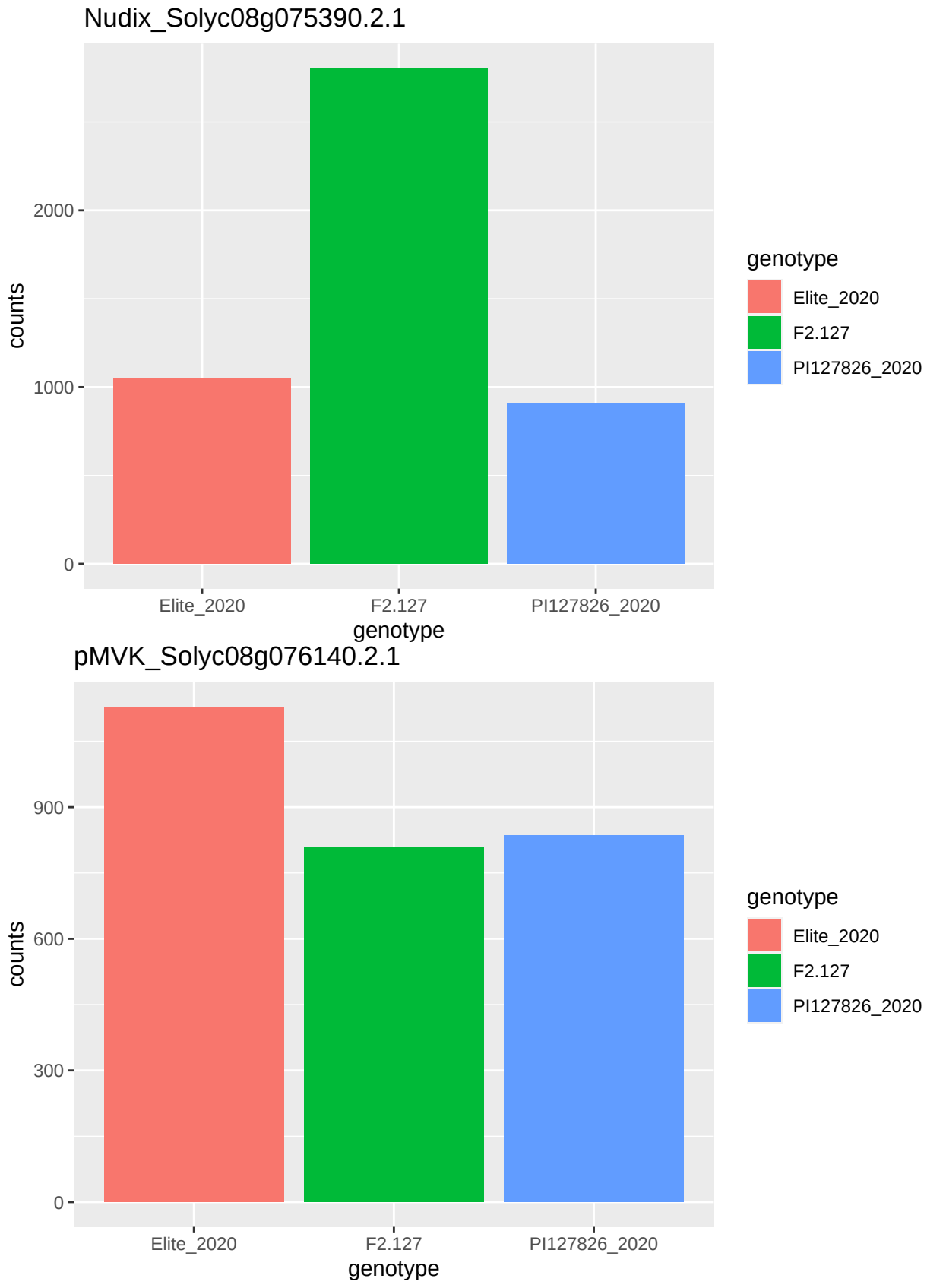

AACT_Solyc05g017760.2.1
counts
3000
2000
1000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

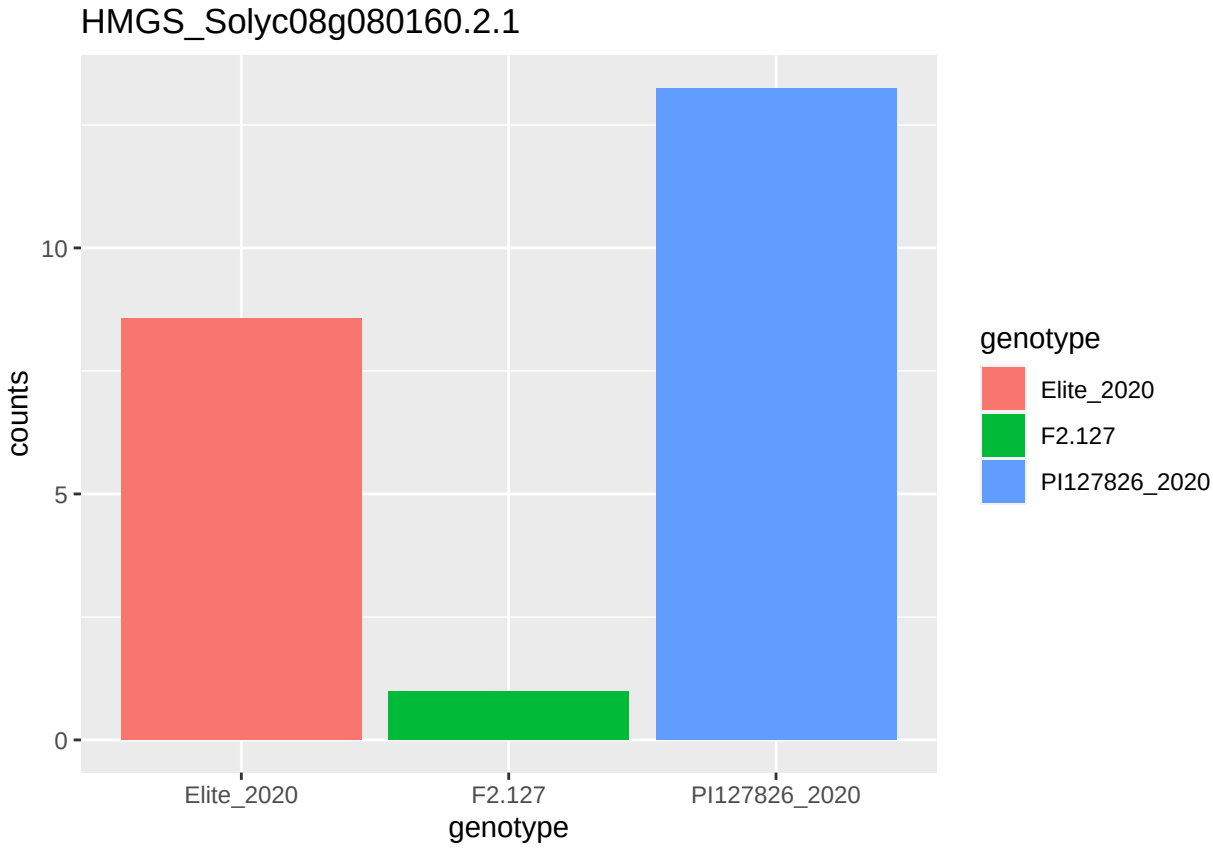

IDI2_Solyc05g055760.2.1
counts
1500
1000
500
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020
pMVK_Solyc06g066310.2.1
counts
1500
1000
500
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

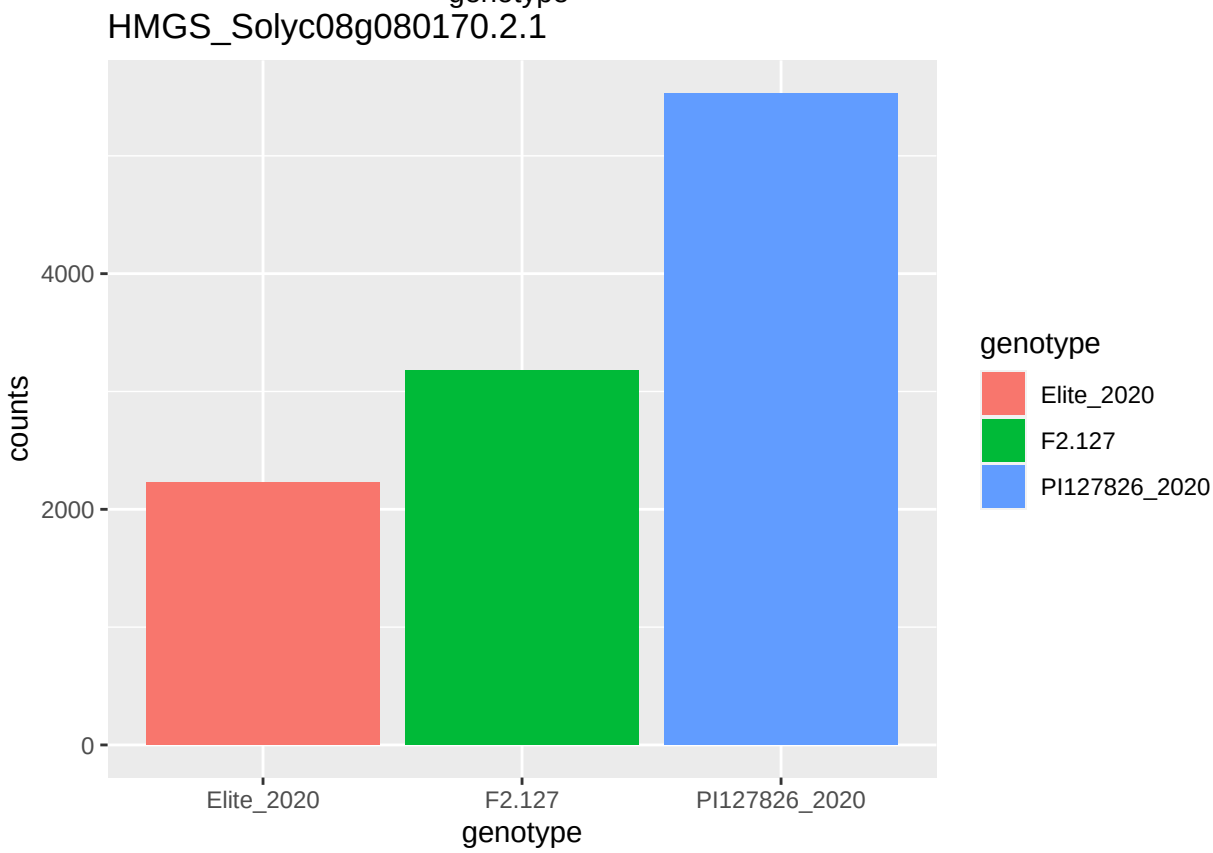

AACT_Solyc07g045350.2.1
counts
4000
2000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020
NDPS/zFPS_Solyc08g005680.2.1
counts
125000
100000
75000
50000
25000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

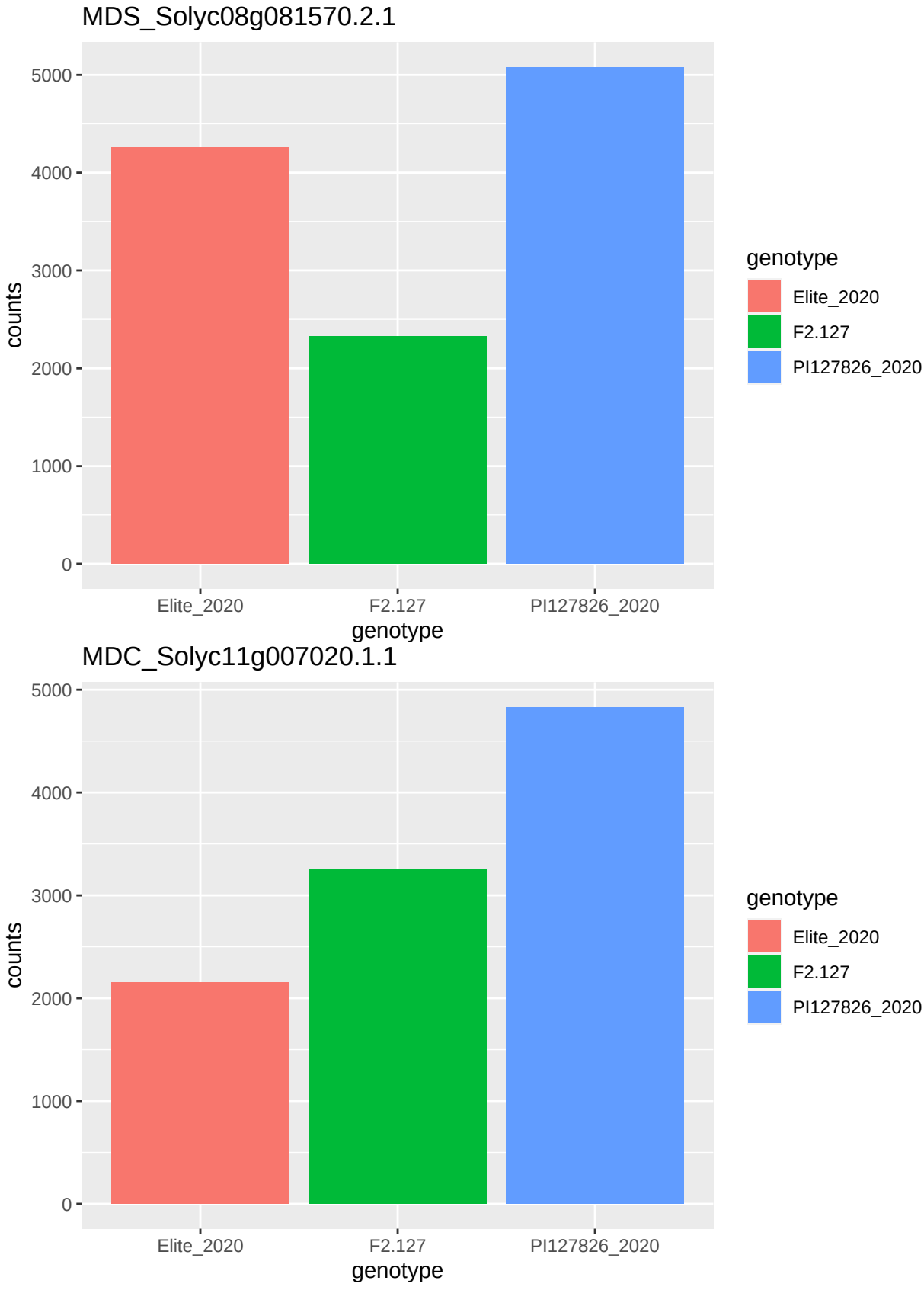

HMGS_Solyc08g007790.2.1
counts
600
400
200
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

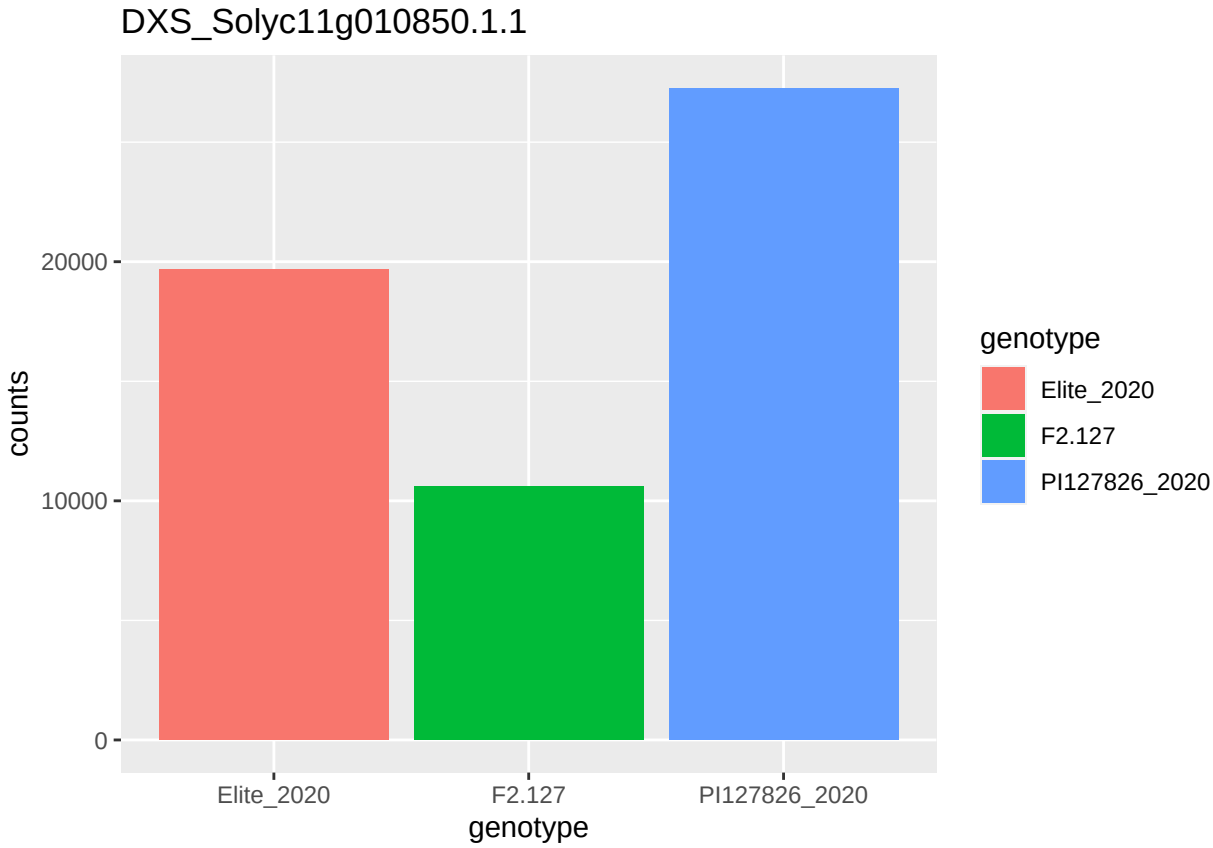

DXS_Solyc08g066950.2.1
counts
1000
750
500
250
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


Nudix_Solyc08g075390.2.1
counts
2000
1000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020
pMVK_Solyc08g076140.2.1
counts
900
600
300
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020

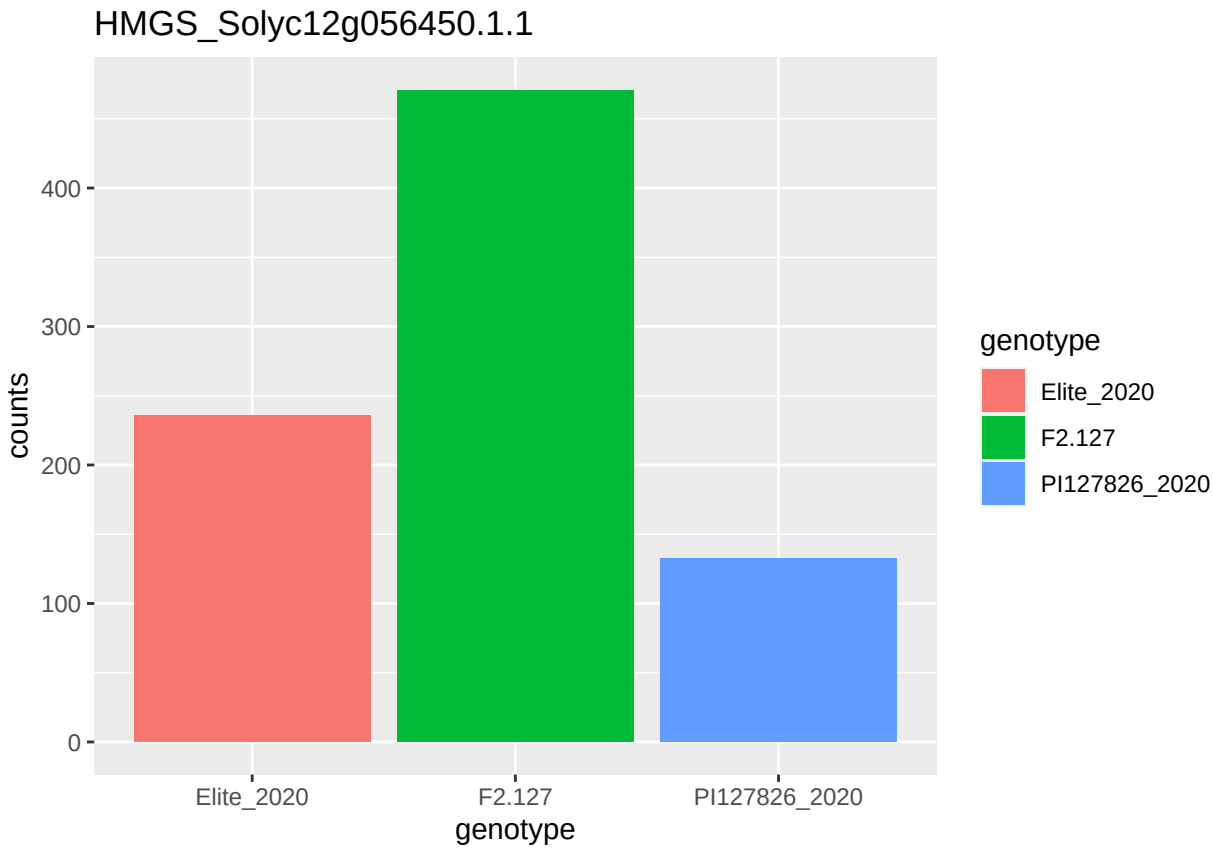

HMGS_Solyc08g080160.2.1
counts
10
5
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


HMGS_Solyc08g080170.2.1
counts
4000
2000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


MDS_Solyc08g081570.2.1
counts
5000
4000
3000
2000
1000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020
MDC_Solyc11g007020.1.1
counts
5000
4000
3000
2000
1000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


DXS_Solyc11g010850.1.1
counts
20000
10000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


HDS_Solyc11g069380.1.1
counts
30000
20000
10000
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020


HMGS_Solyc12g056450.1.1
counts
400
300
200
100
0
Elite_2020
F2.127
PI127826_2020
genotype
genotype
Elite_2020
F2.127
PI127826_2020