# Differential expression (F2-28, 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) %>%
  mutate(gene = gsub(pattern = "mRNA:", replacement = "", x = gene)) %>%
  dplyr::select("gene", "F2-28_concat", "Elite_2020", "PI127826_2020")
```

## 1.2 Mapping summary

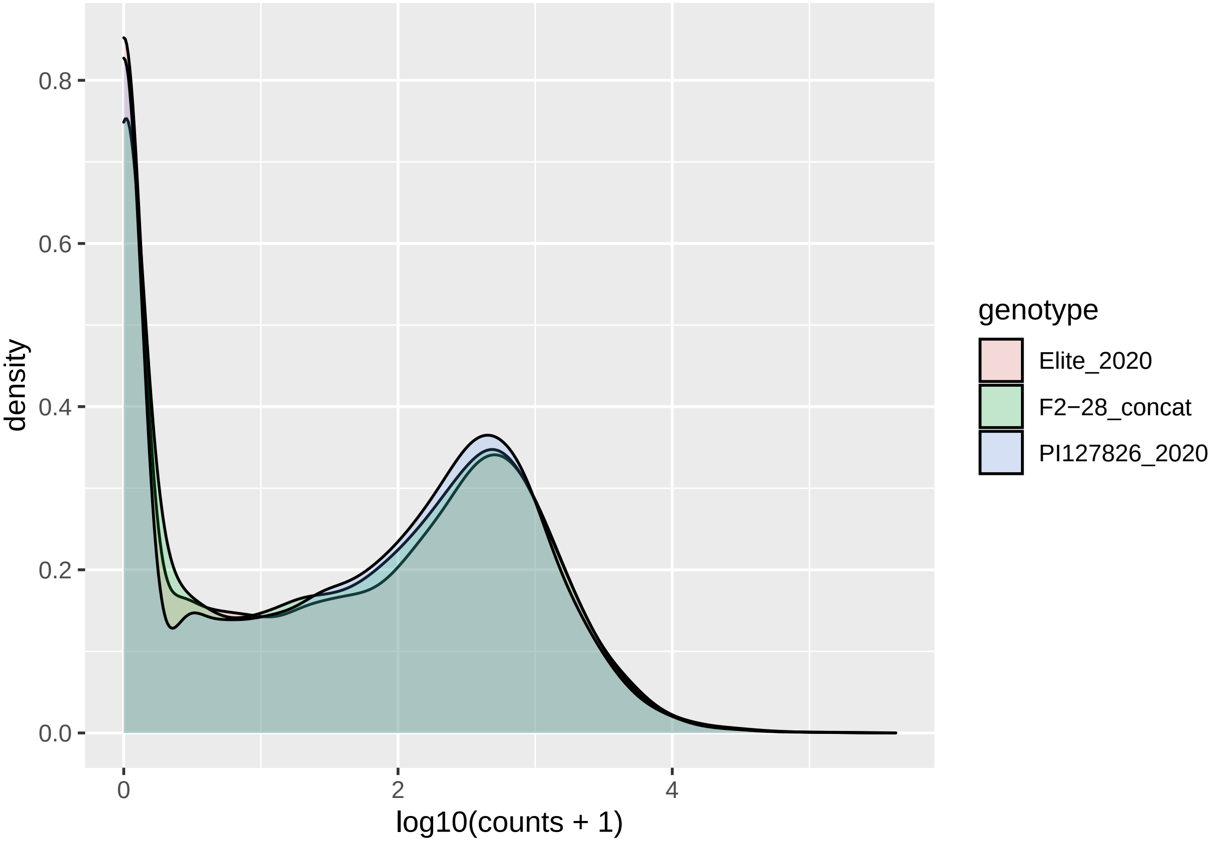
```
read.csv("../Supplemental_data_RNA-seq/mapping_summary.csv",
         col.names = c("row_id", "attribute", "Elite", "F2-28", "PI127826"),
         stringsAsFactors = F,
         check.names = F) %>%
  knitr::kable()
```

| row_id | attribute | Elite | F2-28 | PI127826 |
|---:|---|---|---|---|
| 0 | Started job on \| | Oct 18 10:33:50 | Oct 18 10:33:50 | Oct 18 10:33:50 |
| 1 | Started mapping on \| | Oct 18 10:33:51 | Oct 18 10:33:51 | Oct 18 10:33:51 |
| 2 | Finished on \| | Oct 18 10:35:12 | Oct 18 10:39:59 | Oct 18 10:35:41 |
| 3 | Mapping speed, Million of reads per hour \| | 881.26 | 813.11 | 613.73 |
| 4 | Number of input reads \| | 19828365 | 83118240 | 18752904 |
| 5 | Average input read length \| | 75 | 75 | 75 |
| 6 | UNIQUE READS: | | | |
| 7 | Uniquely mapped reads number \| | 18055736 | 61661853 | 14198867 |
| 8 | Uniquely mapped reads % \| | 91.06% | 74.19% | 75.72% |
| 9 | Average mapped length \| | 75.93 | 75.97 | 75.94 |
| 10 | Number of splices: Total \| | 2884753 | 9859004 | 2184309 |
| 11 | Number of splices: Annotated (sjdb) \| | 2657391 | 9134650 | 2021184 |
| 12 | Number of splices: GT/AG \| | 2848414 | 9700348 | 2143184 |
| 13 | Number of splices: GC/AG \| | 27981 | 92798 | 22373 |
| 14 | Number of splices: AT/AC \| | 876 | 2987 | 710 |
| 15 | Number of splices: Non-canonical \| | 7482 | 62871 | 18042 |
| 16 | Mismatch rate per base, % \| | 0.48% | 1.15% | 1.55% |
| 17 | Deletion rate per base \| | 0.01% | 0.07% | 0.09% |
| 18 | Deletion average length \| | 1.68 | 2.03 | 2.14 |
| 19 | Insertion rate per base \| | 0.01% | 0.03% | 0.04% |
| 20 | Insertion average length \| | 1.49 | 1.72 | 1.75 |
| 21 | MULTI-MAPPING READS: | | | |
| 22 | Number of reads mapped to multiple loci \| | 768089 | 3757232 | 1700643 |
| 23 | % of reads mapped to multiple loci \| | 3.87% | 4.52% | 9.07% |
| 24 | Number of reads mapped to too many loci \| | 4801 | 10938 | 2723 |
| 25 | % of reads mapped to too many loci \| | 0.02% | 0.01% | 0.01% |
| 26 | UNMAPPED READS: | | | |
| 27 | Number of reads unmapped: too many mismatches \| | 445826 | 5091357 | 1666717 |
| 28 | % of reads unmapped: too many mismatches \| | 2.25% | 6.13% | 8.89% |
| 29 | Number of reads unmapped: too short \| | 276153 | 4291129 | 928705 |
| 30 | % of reads unmapped: too short \| | 1.39% | 5.16% | 4.95% |
| 31 | Number of reads unmapped: other \| | 277760 | 8305731 | 255249 |
| 32 | % of reads unmapped: other \| | 1.40% | 9.99% | 1.36% |
| 33 | CHIMERIC READS: | | | |
| 34 | Number of chimeric reads \| | 0 | 0 | 0 |
| 35 | % of chimeric reads \| | 0.00% | 0.00% | 0.00% |

# 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 F2-28 but seems to be comparable for other genes.

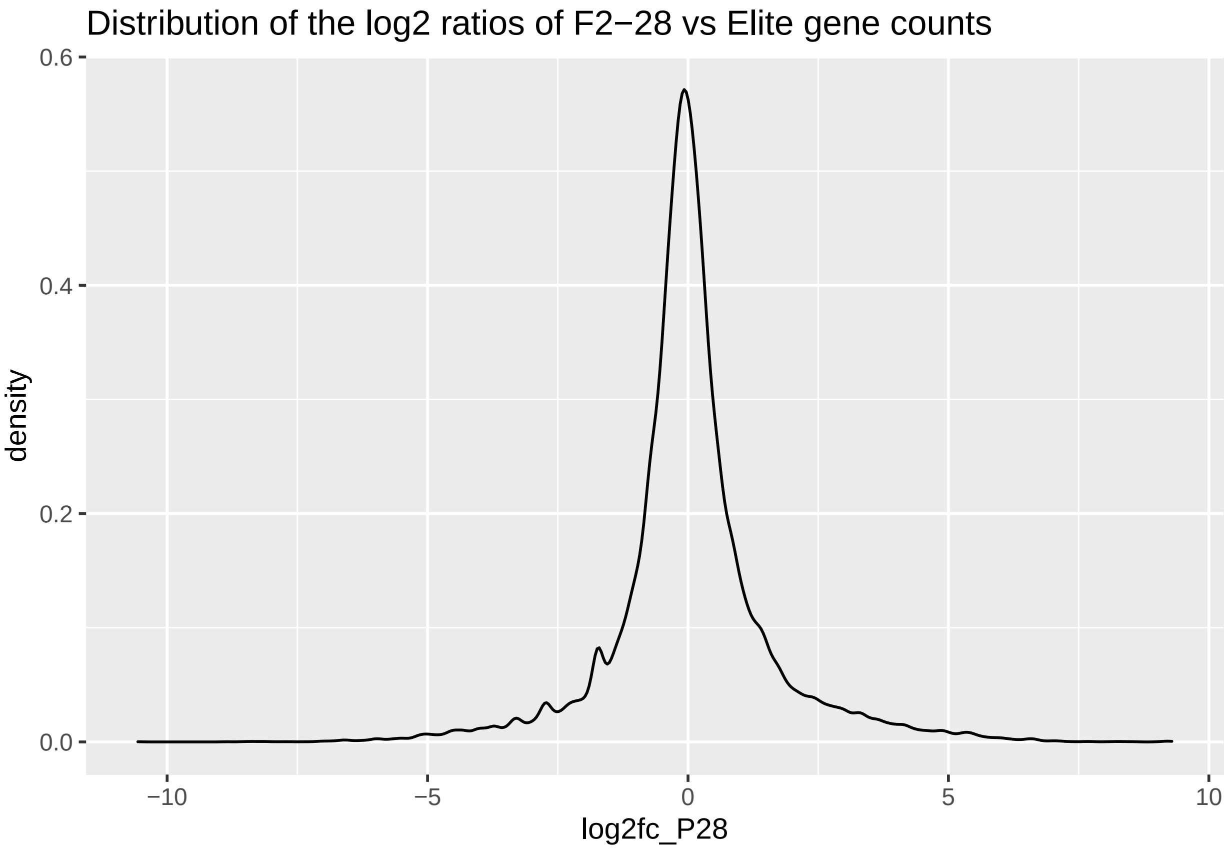
## 2.2 Plot log2ratio F2-28 vs Elite

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_P28 = log2(`F2-28_concat`/Elite_2020)) %>%
  ggplot(aes(x = log2fc_P28)) +
  geom_density() +
  ggtitle("Distribution of the log2 ratios of F2-28 vs Elite gene counts")
```

```
## Warning: Removed 3847 rows containing non-finite values (stat_density).
```

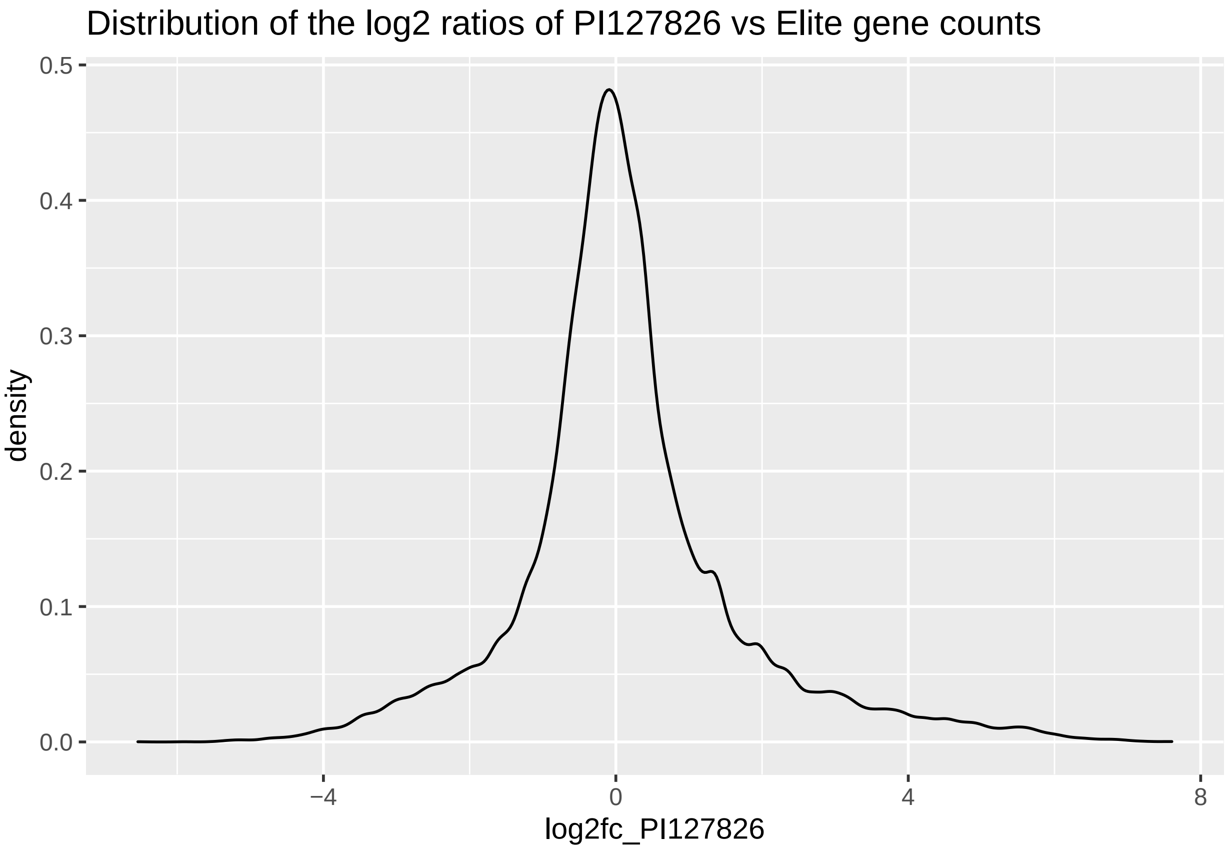


The log2ratio distribution looks OK.

### 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 4589 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 F2-28 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_P28 = log2(`F2-28_concat`/Elite_2020)) %>%
  mutate(log2ratio_PI127826 = log2(PI127826_2020/Elite_2020)) %>%
  select(gene, log2ratio_P28, log2ratio_PI127826) %>%
  filter(!grepl(pattern = "Inf", x = log2ratio_P28)) %>%
  filter(!grepl(pattern = "Inf", x = log2ratio_PI127826))

head(log2ratio)
```

```
##                  gene log2ratio_P28 log2ratio_PI127826
## 1 Solyc00g005040.2.1    -0.9349052         -1.8336997
## 2 Solyc00g005050.2.1     0.1267105          0.4482855
## 3 Solyc00g005080.1.1     0.5886568          0.3679341
## 4 Solyc00g005150.1.1     1.8213175          2.4553970
## 5 Solyc00g005840.2.1     1.1218088          0.7093396
## 6 Solyc00g005860.1.1    -1.2478445         -2.9539940
```

A total of **22793** have a finite log2ratio in both F2-28 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_P28 = scale(log2ratio_P28, center = T, scale = T)) %>%
  mutate(zscore_PI127826 = scale(log2ratio_PI127826, center = T, scale = T)) %>%
  mutate(pval_P28 = pnorm(q = abs(zscore_P28), 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=FALS
  arrange(desc(log2ratio_P28)) %>%
  as_tibble()

head(log2ratio_zscores_pvals)
```

```
## # A tibble: 6 x 7
##   gene  log2ratio_P28 log2ratio_PI127~ zscore_P28[,1] zscore_PI127826~
##   <chr>         <dbl>            <dbl>          <dbl>            <dbl>
## 1 Soly~          9.29             6.73           6.18             4.23
## 2 Soly~          9.22             6.02           6.13             3.77
## 3 Soly~          9.21             6.86           6.12             4.31
## 4 Soly~          9.18             3.37           6.10             2.06
## 5 Soly~          9.05             5.21           6.02             3.25
## 6 Soly~          8.56             3.88           5.69             2.39
## # ... with 2 more variables: pval_P28[,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] 22793    12
```

## 3.4 Write to CSV file

```
write.csv(final,
          file = "tables/diff_res_F2-28orPI127826_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

Should be significant ($p < 0.05$) in **either** PI127826 vs Elite *AND* F2-28 vs Elite.

```
signif_genes <- filter(final, pval_P28 < 0.05 | pval_PI127826 < 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"
```

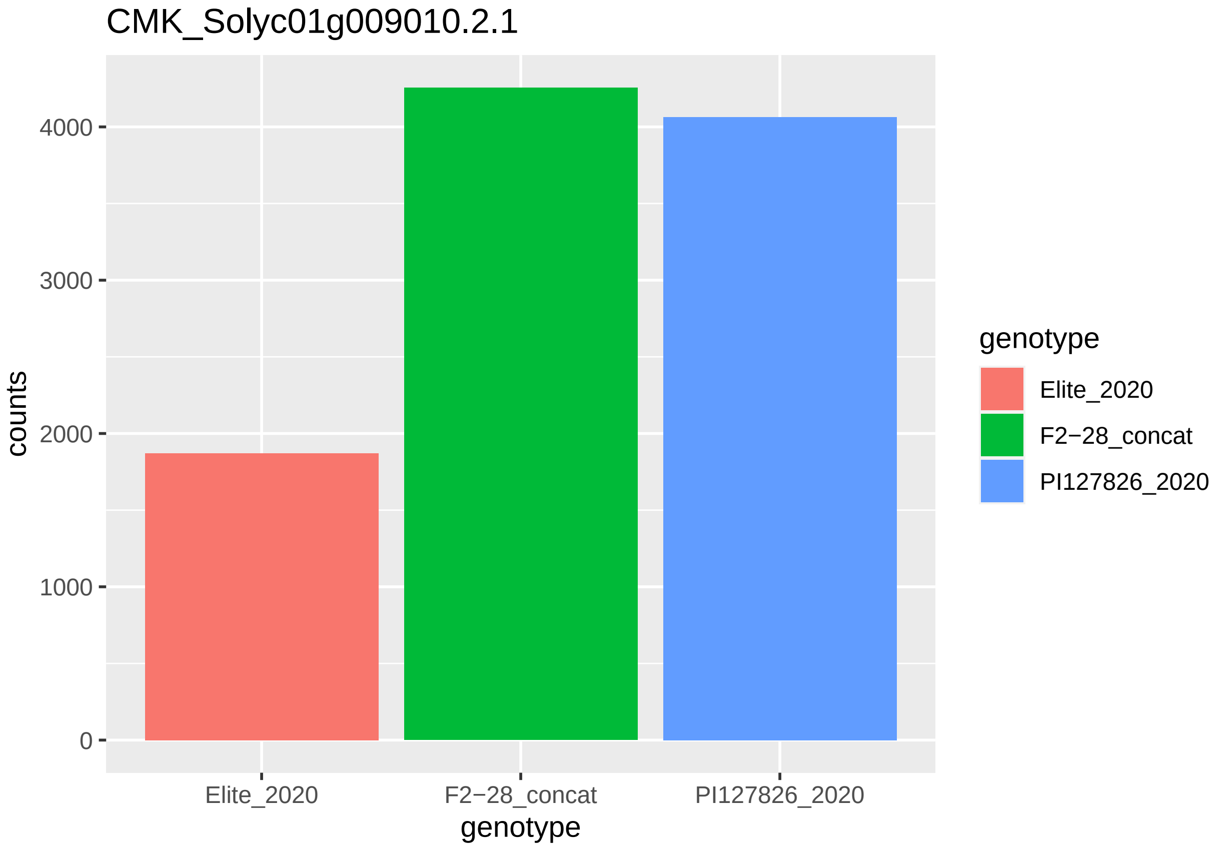
```
mep_mva_gene_signif <-
  mep_mva_genes %>%
  filter(gene %in% signif_genes)

# show table
mep_mva_gene_signif %>%
  select(name, gene, pathway, pval_P28, pval_PI127826) %>%
  knitr::kable()
```

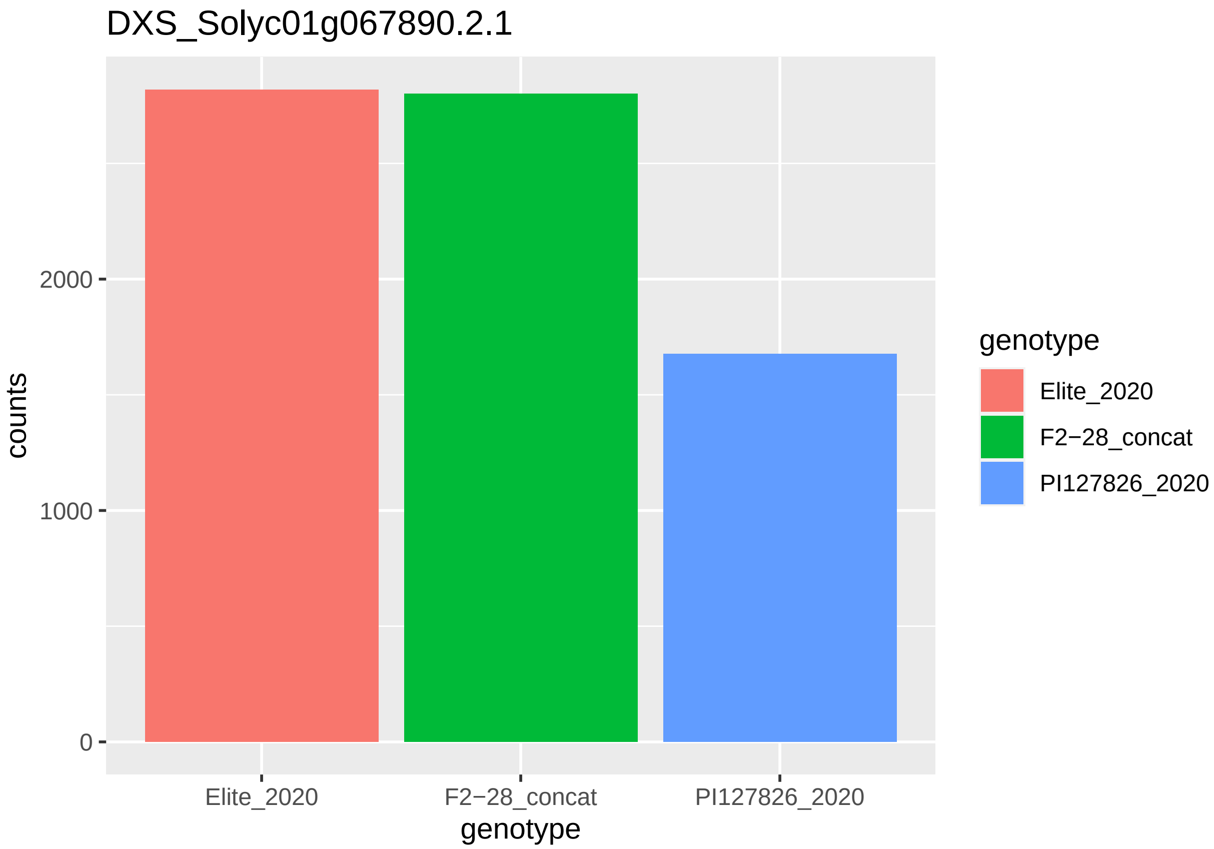
| name | gene | pathway | pval_P28 | pval_PI127826 |
|---|---|---|---:|---:|
| HMGR | Solyc02g038740.2.1 | MVA | 0.08318780 | 0.01081294 |
| HMGR | Solyc03g032010.2.1 | MVA | 0.13853560 | 0.03476849 |
| pMVK | Solyc06g066310.2.1 | MVA | 0.01070946 | 0.01499932 |

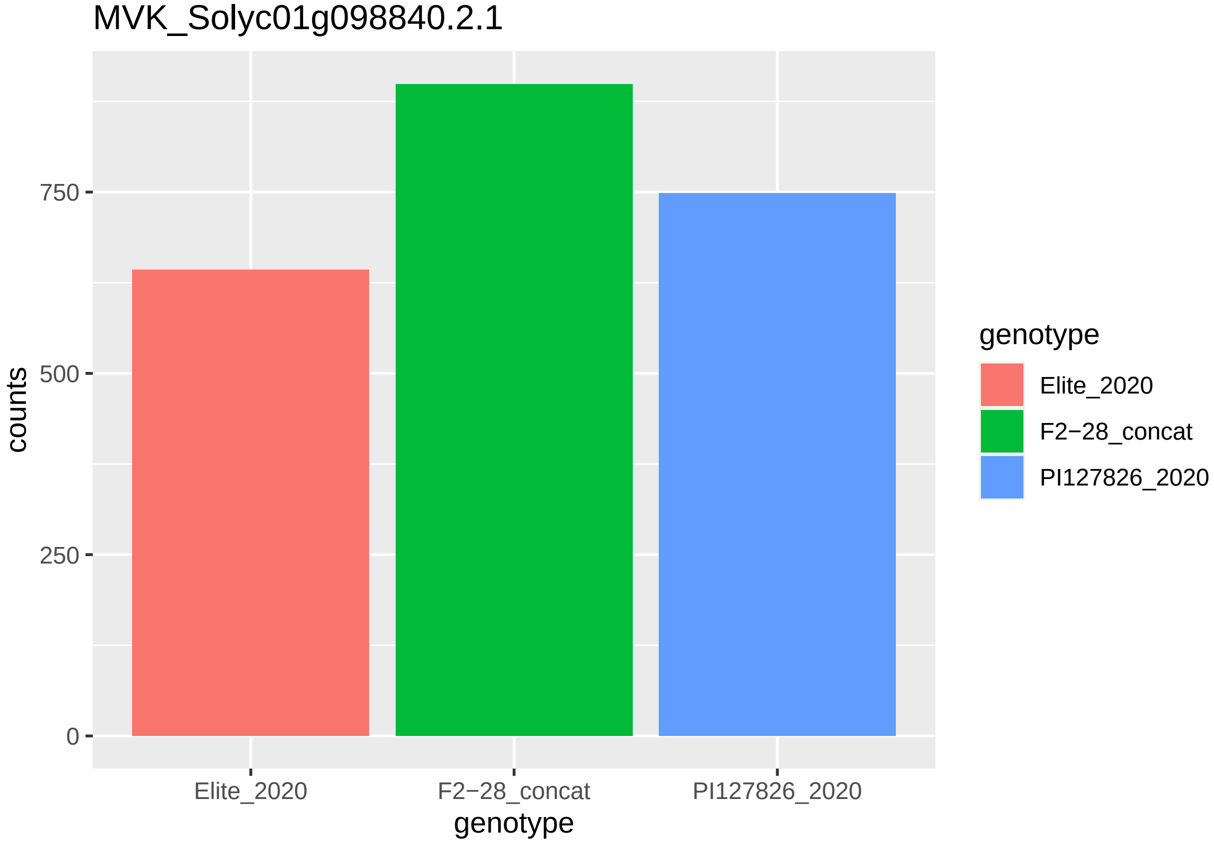
```
write.csv(mep_mva_gene_signif,
            file = "tables/mep_mva_gene_signif.csv",
            row.names = F,
            quote = F)
```

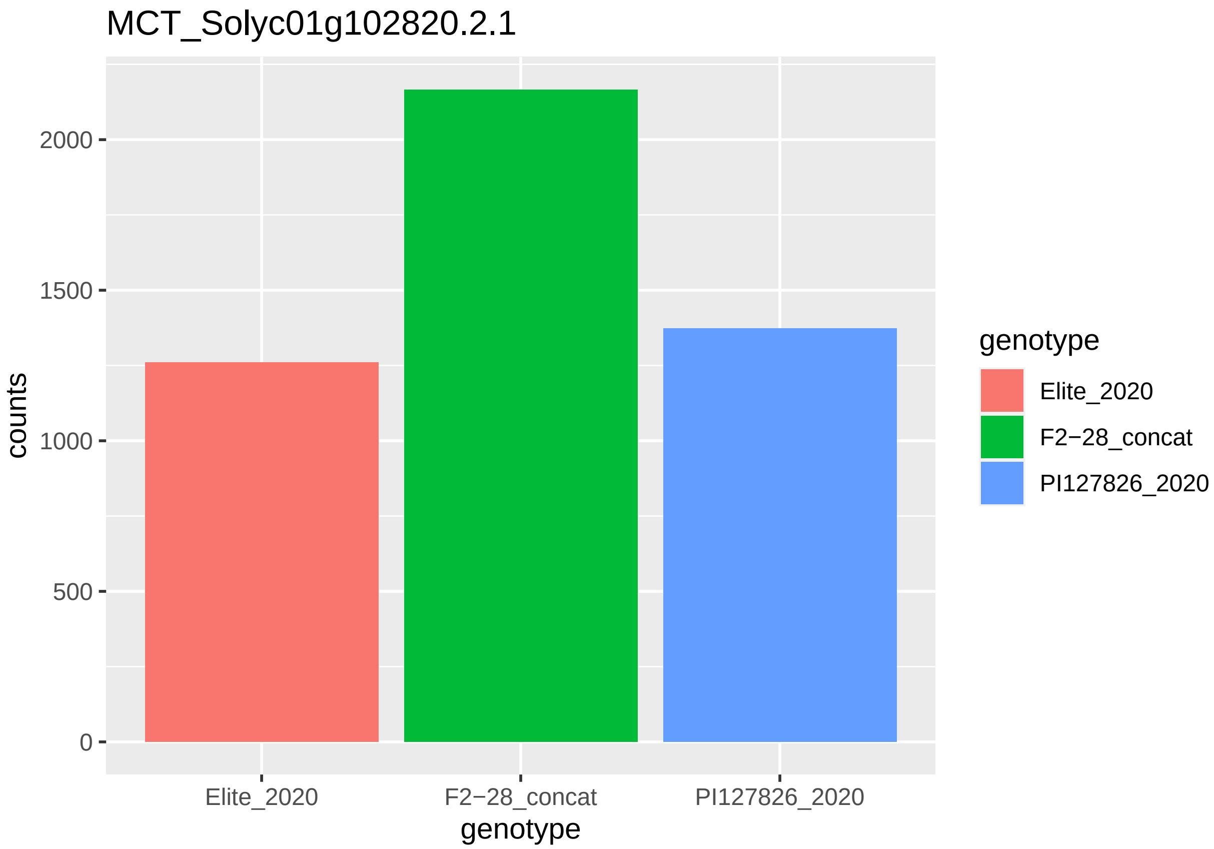
### 4.4 Plot all MEP and MVA genes

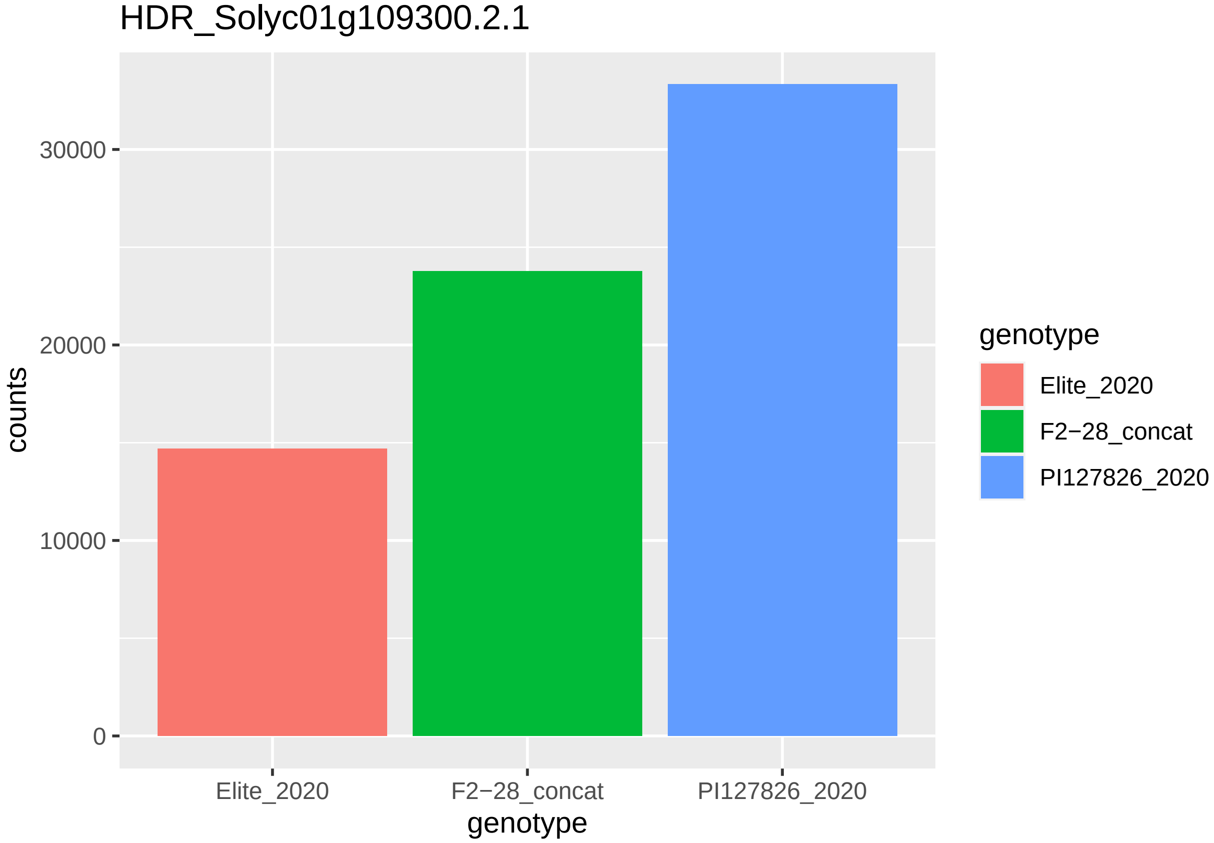
```r
for (i in seq_along(mep_mva_genes$gene)){
  tmp_df <- mep_mva_genes[i,]
  tmp_df$title4plot <- paste(tmp_df$name, tmp_df$gene, sep = "_")

  p <-
    tmp_df %>%
#    mutate(plot_title = paste(name, gene, sep = "_")) %>%
    select(title4plot, `F2-28_concat`, 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)
}
```

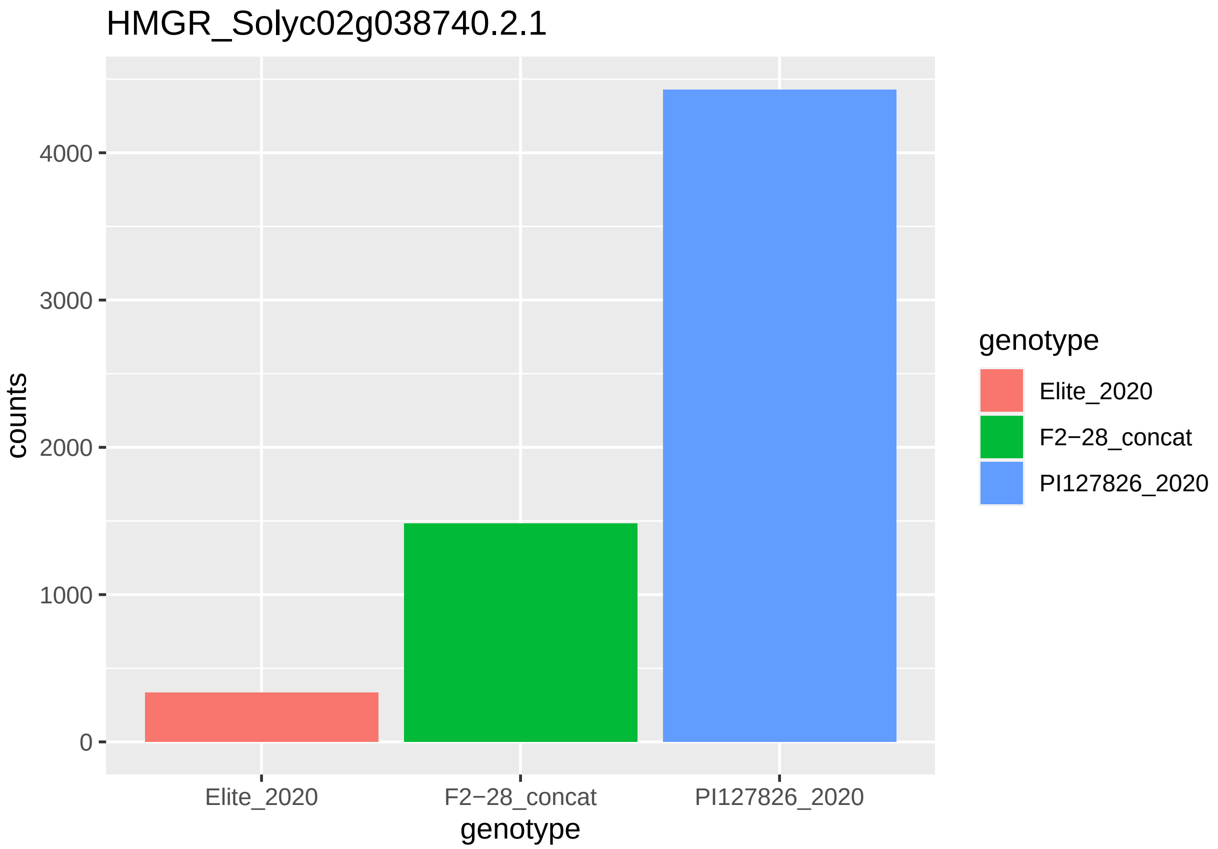


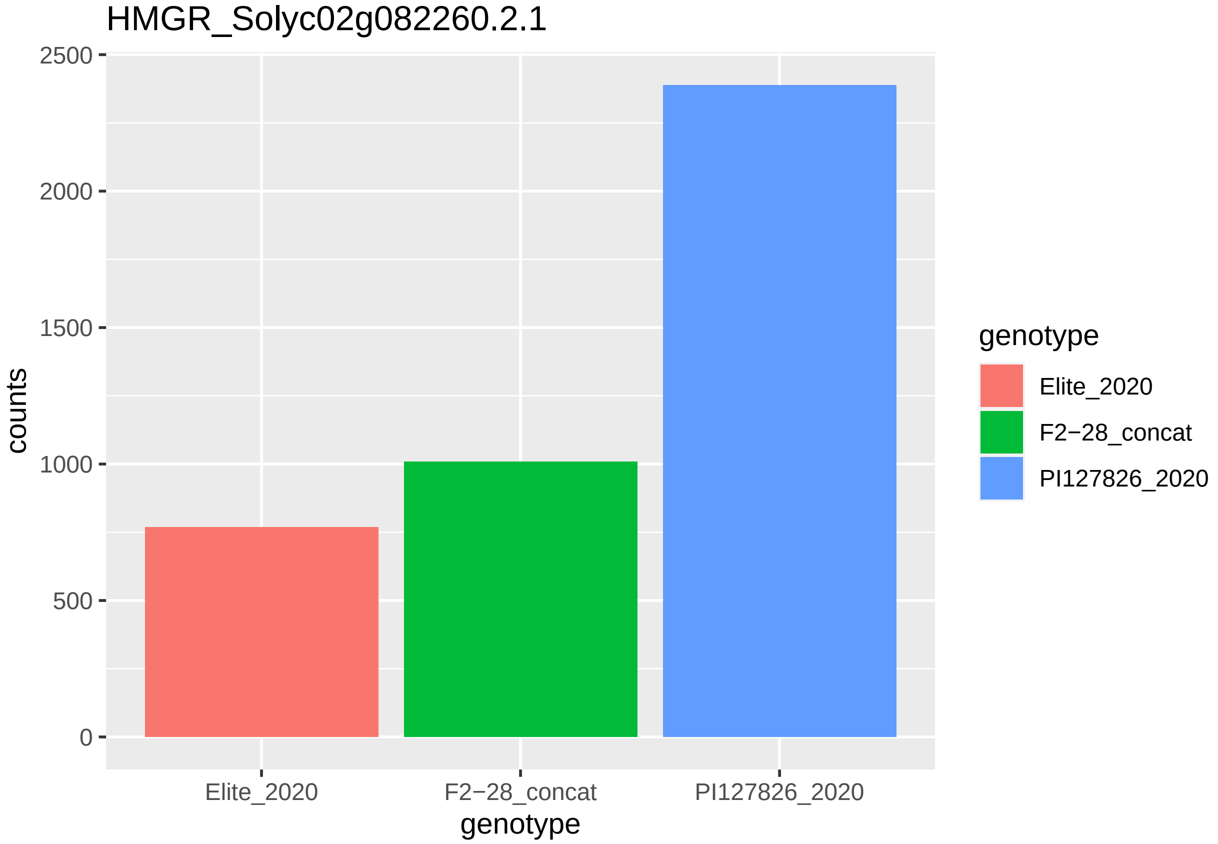

DXS_Solyc01g067890.2.1
counts
2000
1000
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
🟥 Elite_2020
🟩 F2−28_concat
🟦 PI127826_2020

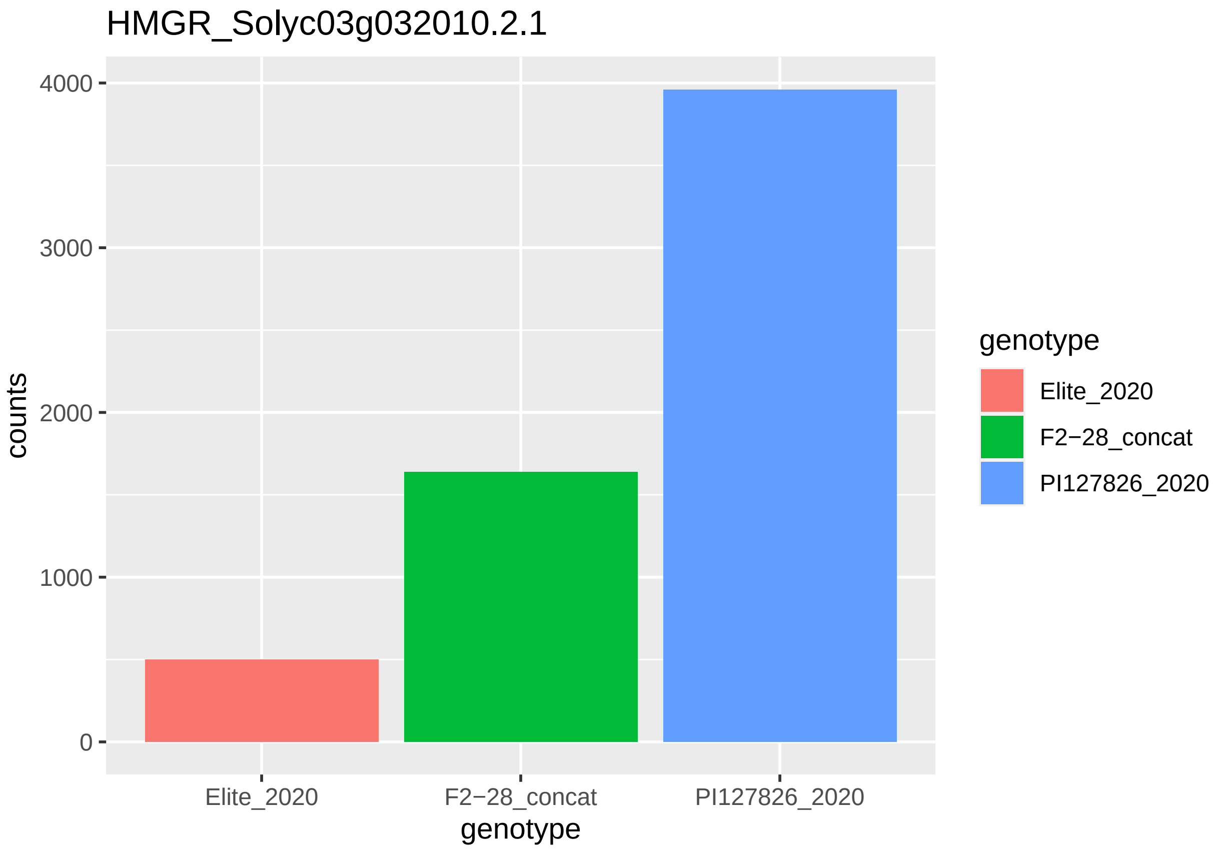

MVK_Solyc01g098840.2.1
counts
750
500
250
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
🟥 Elite_2020
🟩 F2−28_concat
🟦 PI127826_2020

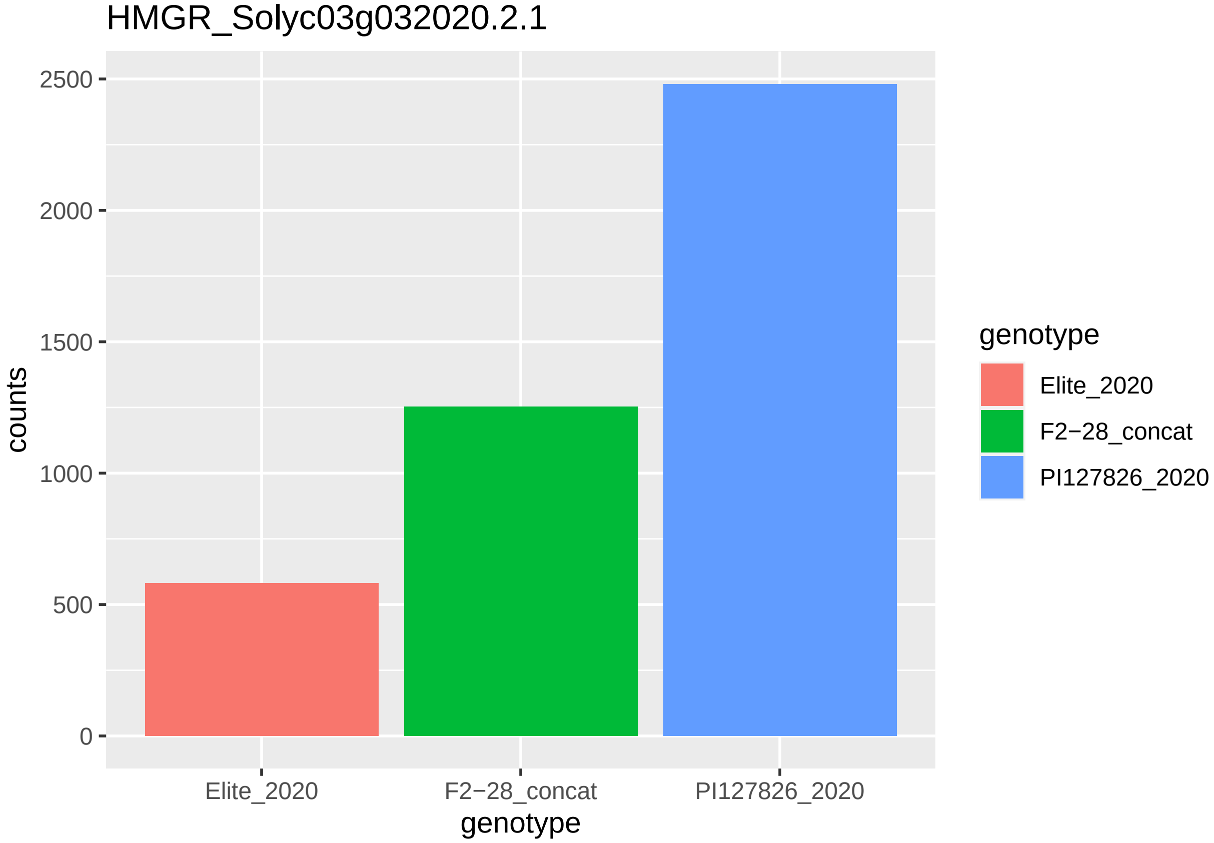

MCT_Solyc01g102820.2.1
counts
2000
1500
1000
500
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

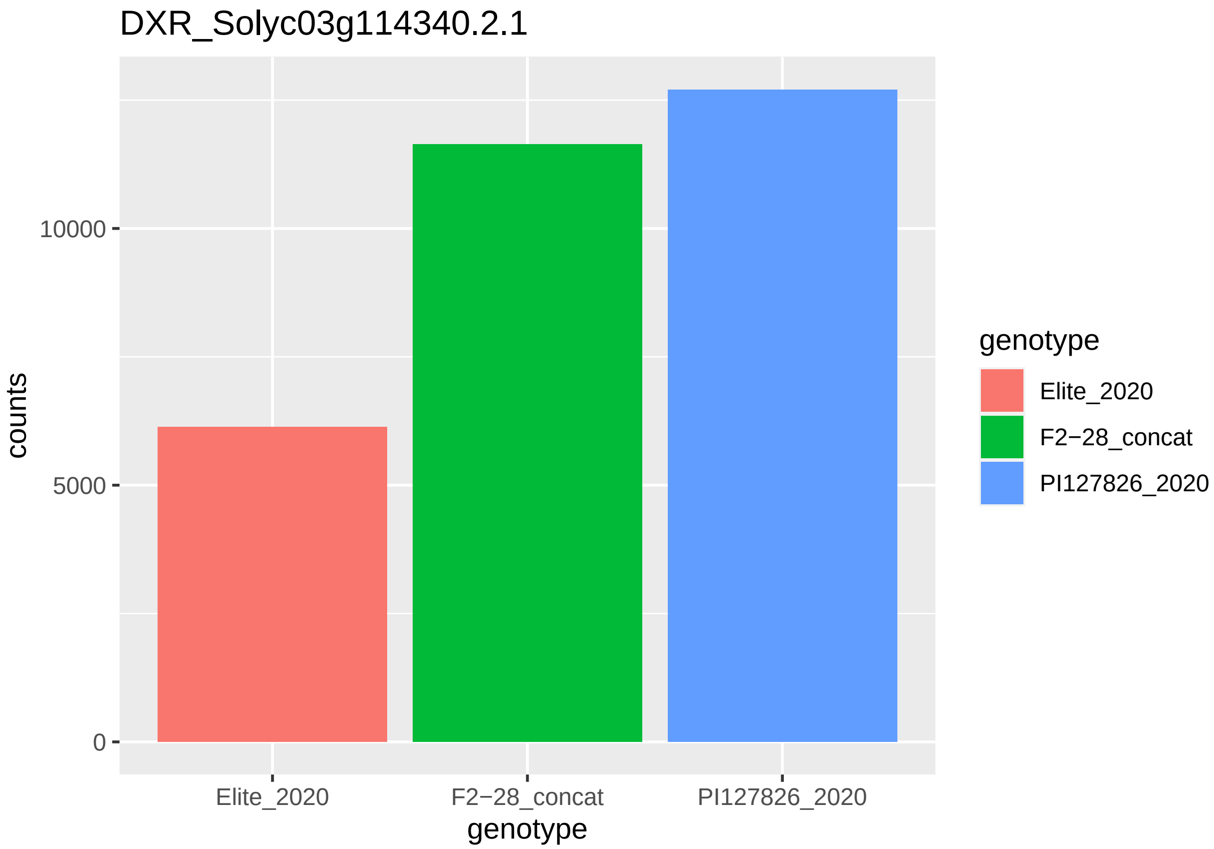

HDR_Solyc01g109300.2.1
counts
30000
20000
10000
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

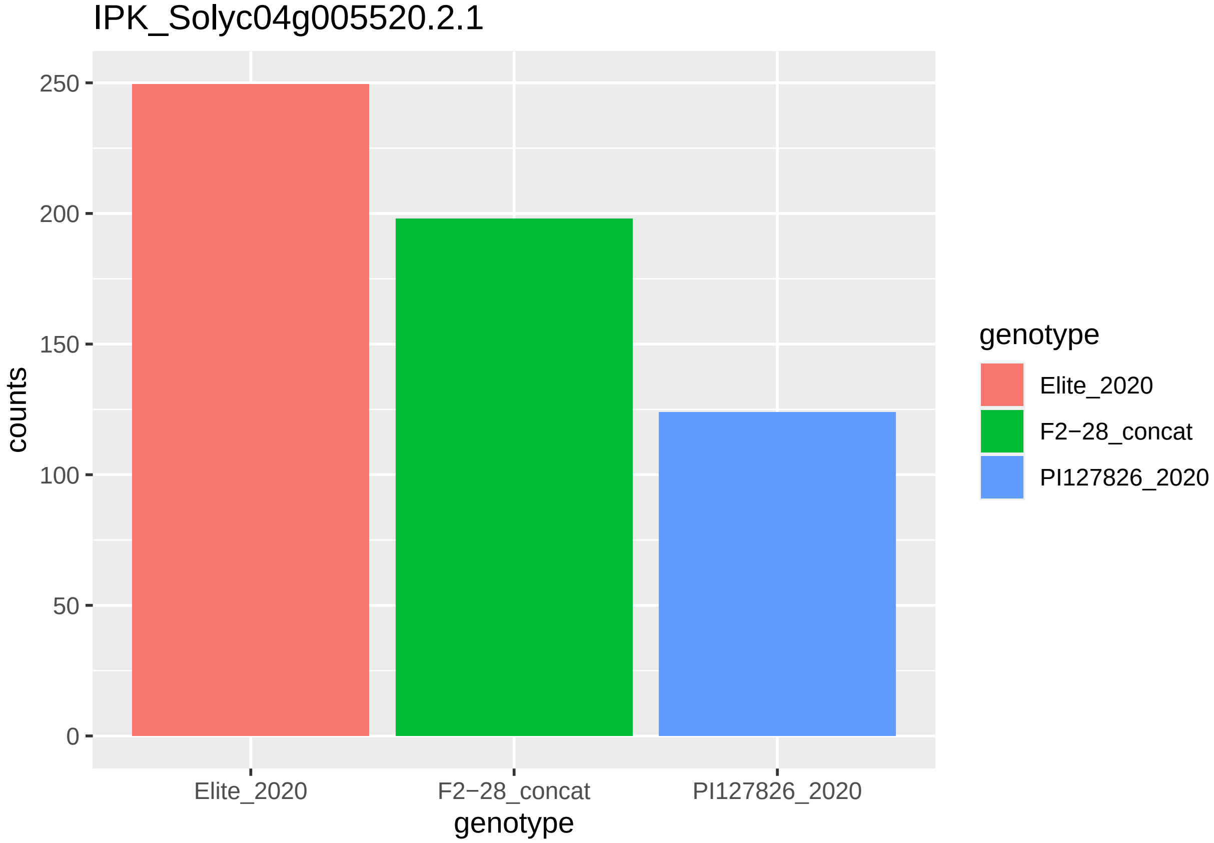

HMGR_Solyc02g038740.2.1

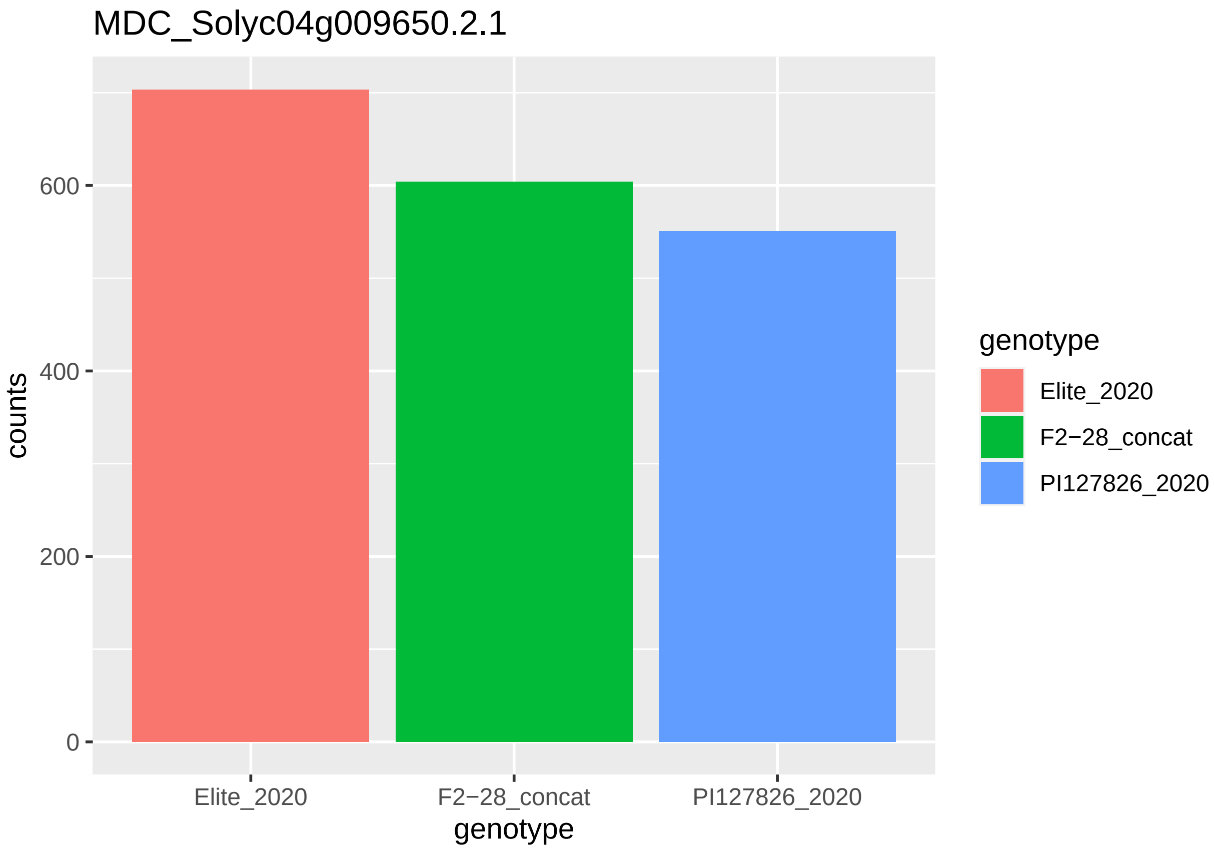

HMGR_Solyc02g082260.2.1

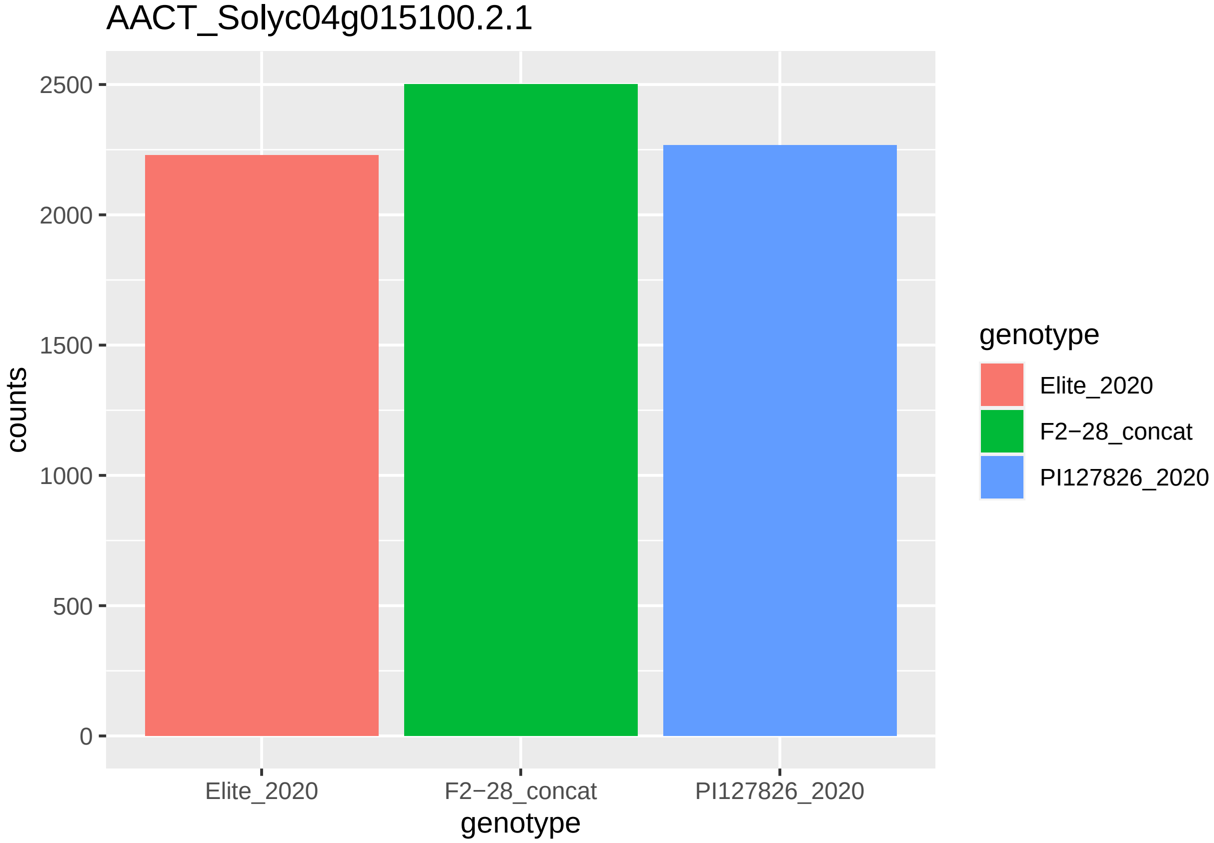

HMGR_Solyc03g032010.2.1

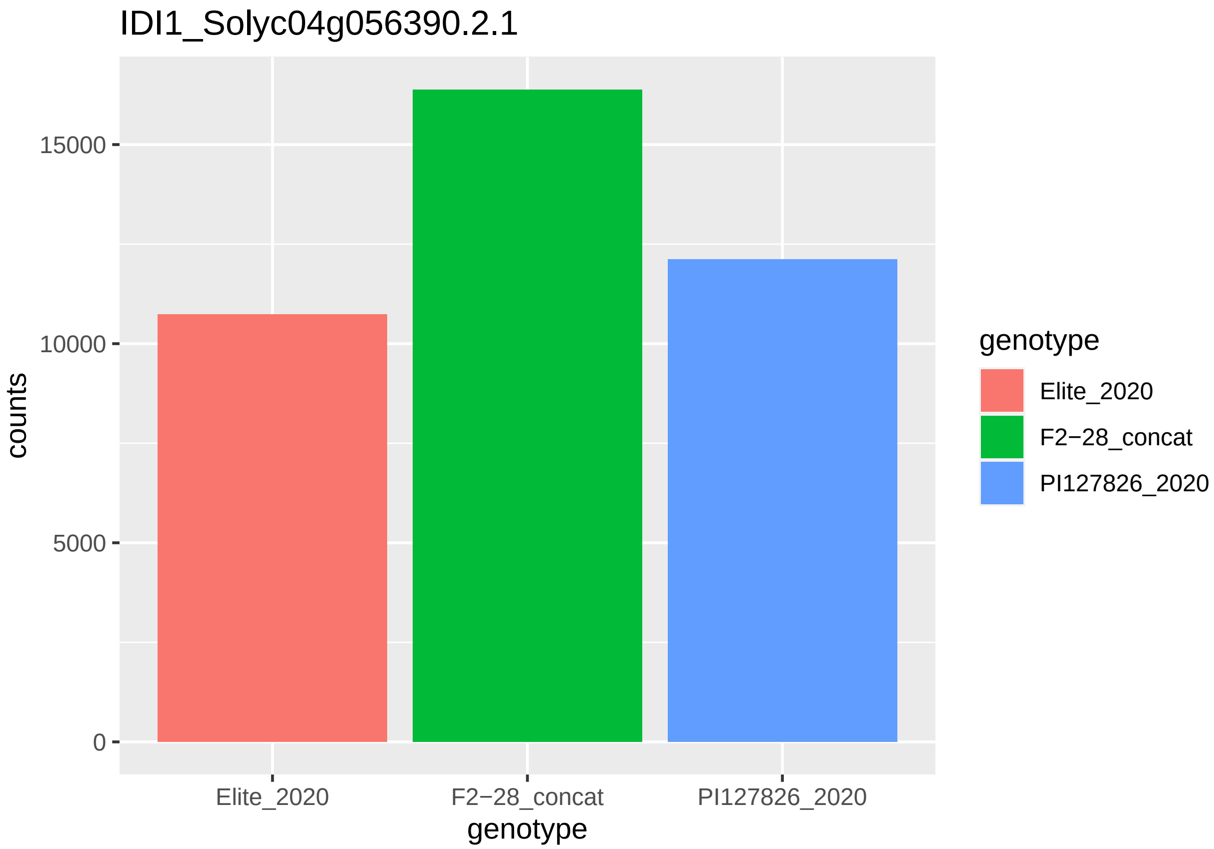

HMGR_Solyc03g032020.2.1

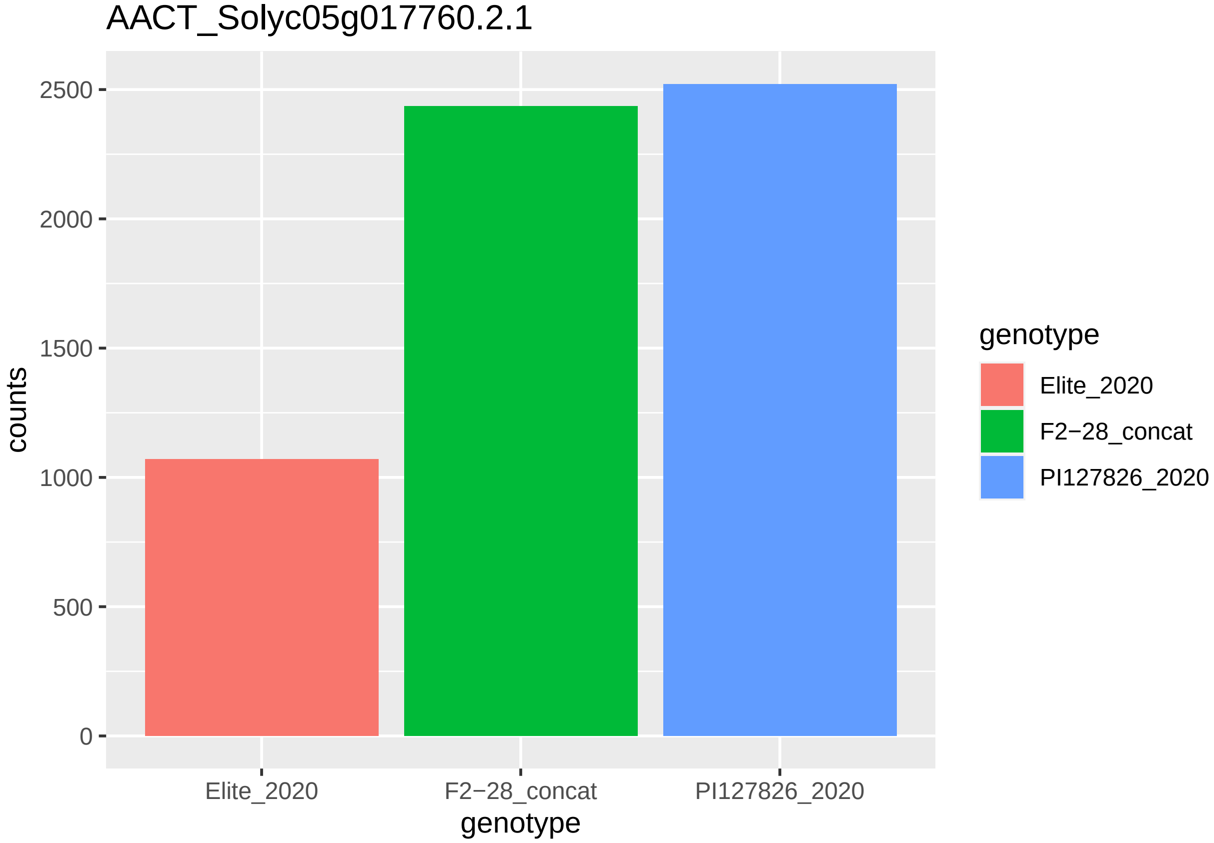

DXR_Solyc03g114340.2.1
counts
10000
5000
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

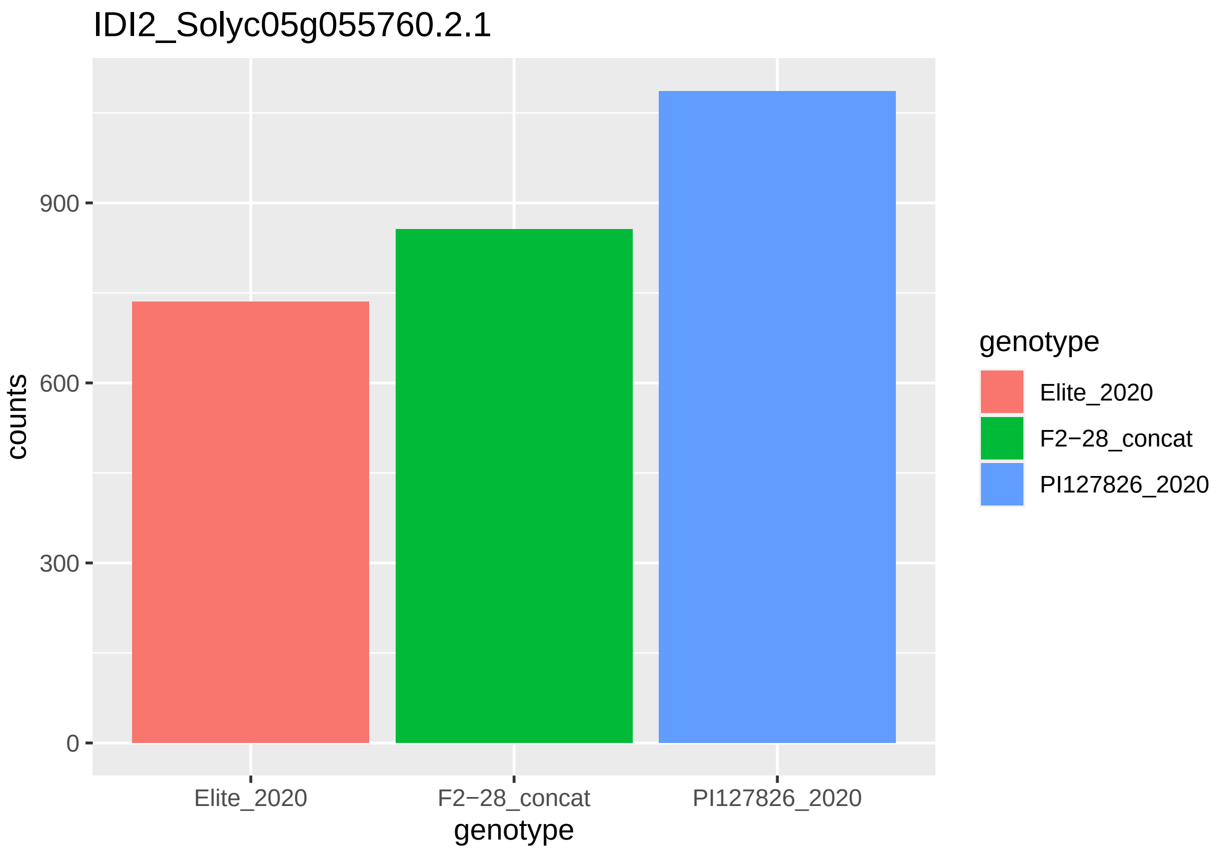

IPK_Solyc04g005520.2.1
counts
250
200
150
100
50
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

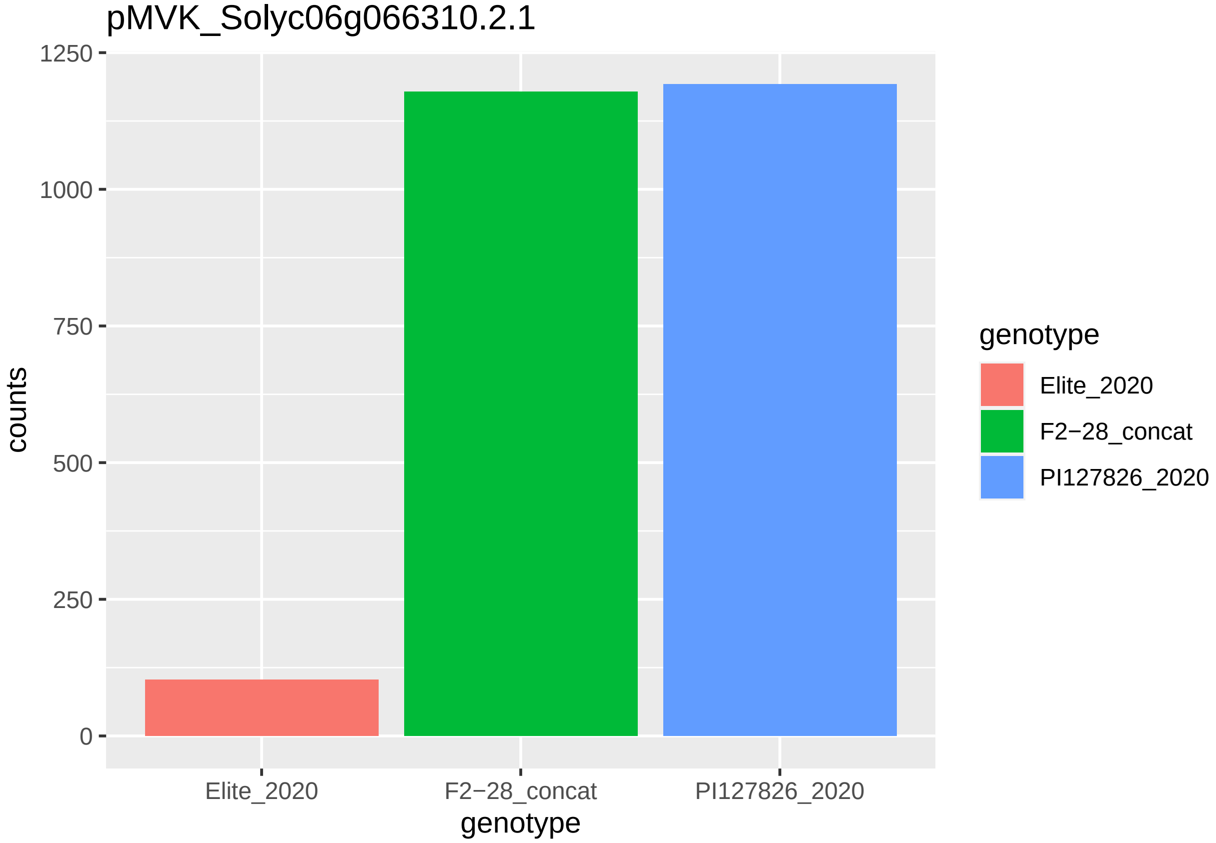

MDC_Solyc04g009650.2.1
counts
600
400
200
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

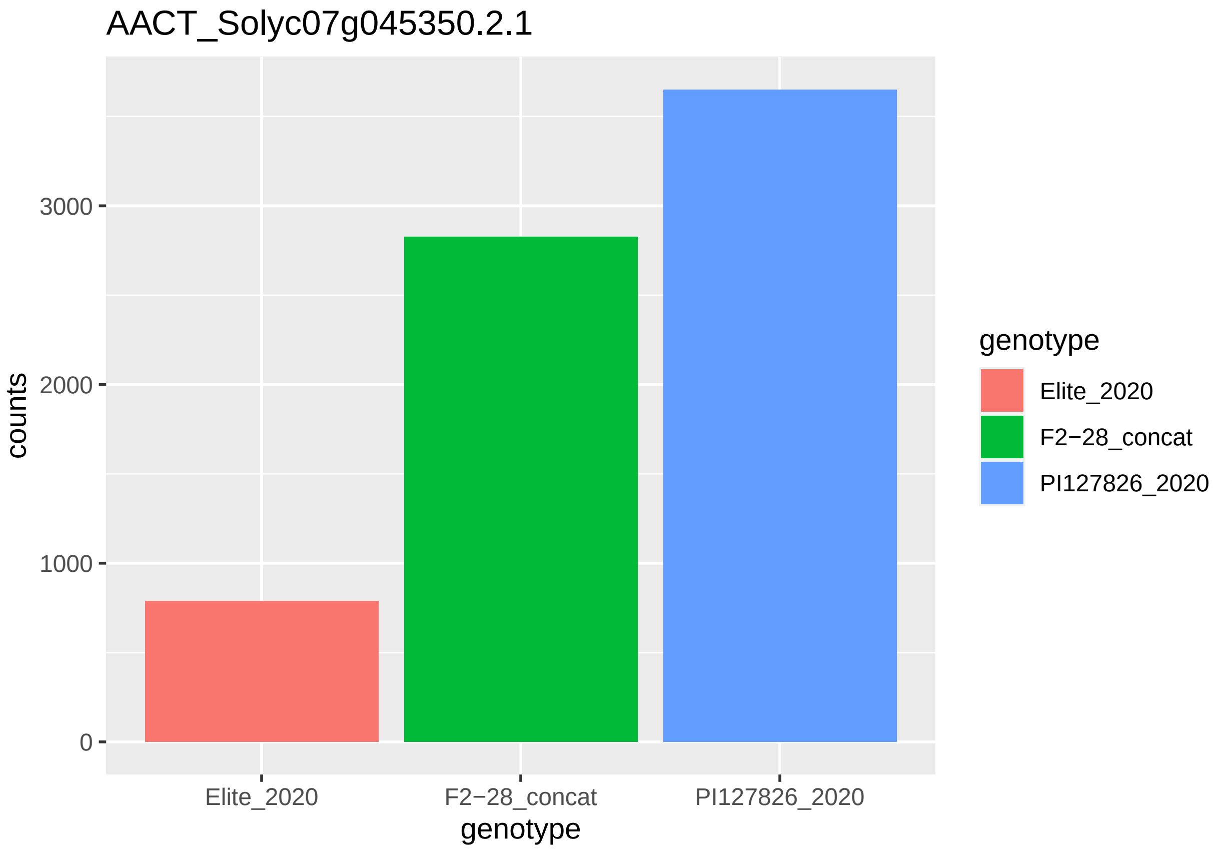

AACT_Solyc04g015100.2.1
counts
2500
2000
1500
1000
500
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

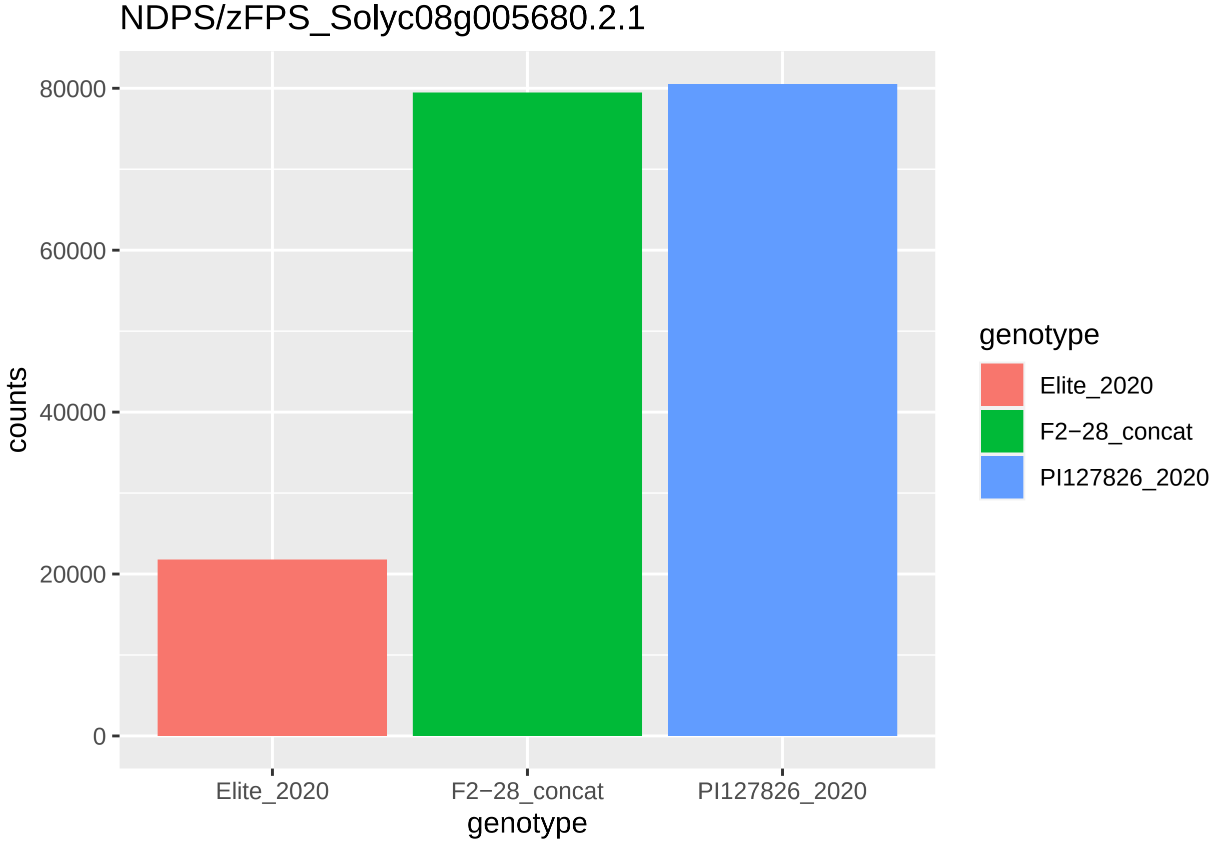

IDI1_Solyc04g056390.2.1
counts
15000
10000
5000
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
🟥 Elite_2020
🟩 F2−28_concat
🟦 PI127826_2020

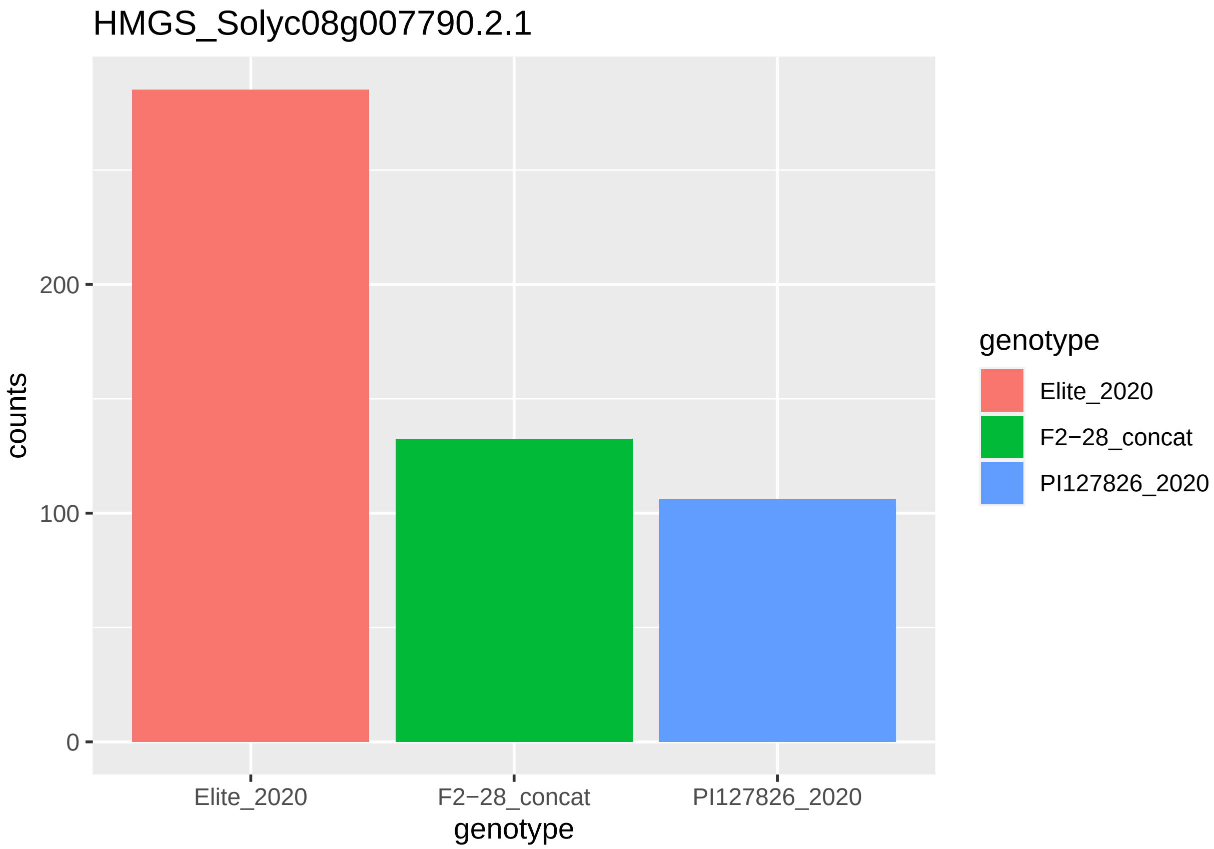

AACT_Solyc05g017760.2.1
counts
2500
2000
1500
1000
500
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
🟥 Elite_2020
🟩 F2−28_concat
🟦 PI127826_2020

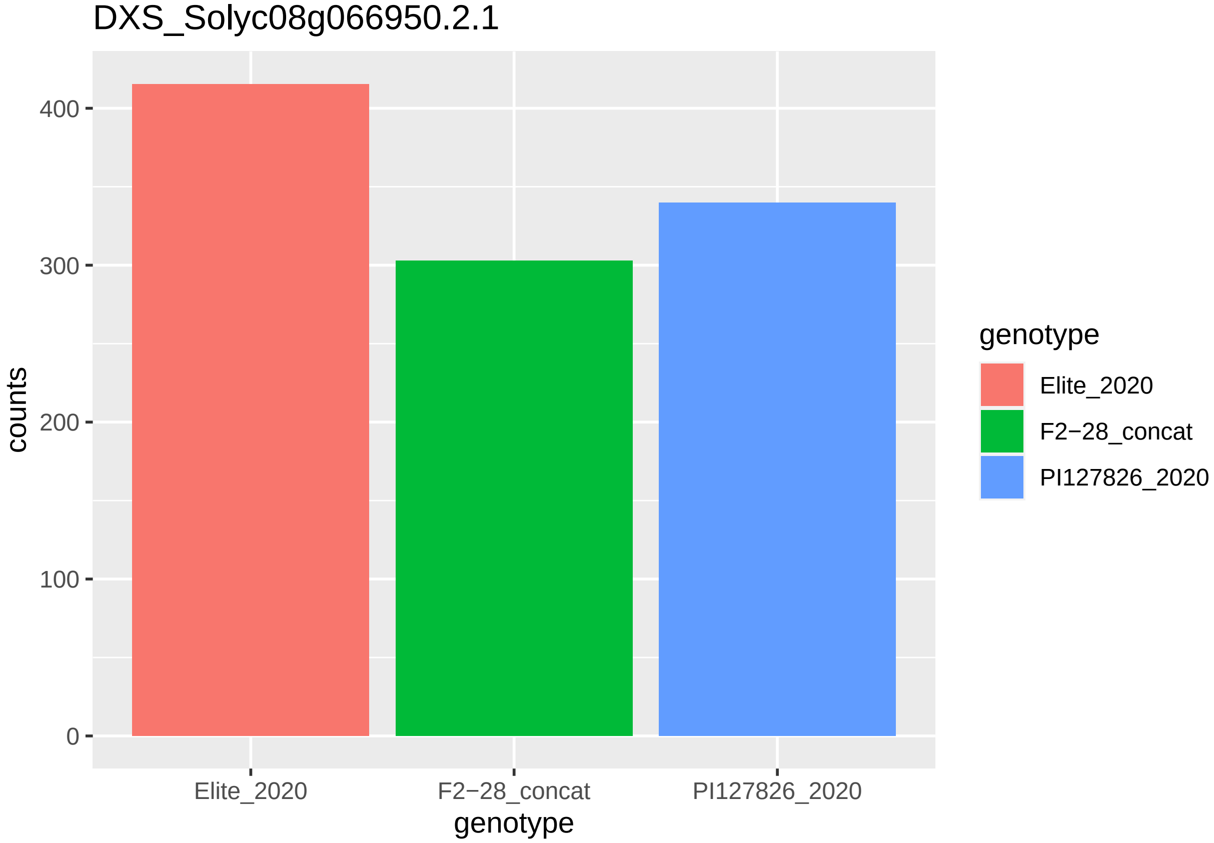

IDI2_Solyc05g055760.2.1
counts
0
300
600
900
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020


pMVK_Solyc06g066310.2.1
counts
0
250
500
750
1000
1250
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

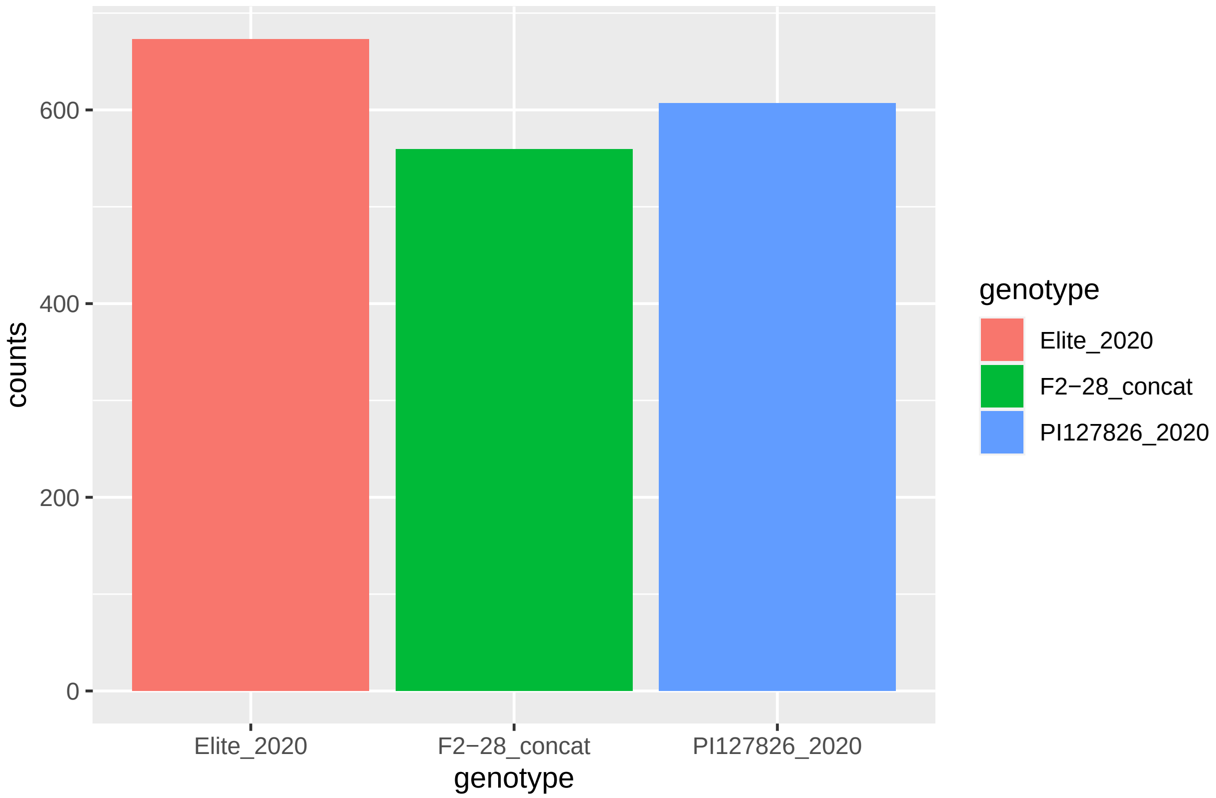

AACT_Solyc07g045350.2.1


NDPS/zFPS_Solyc08g005680.2.1

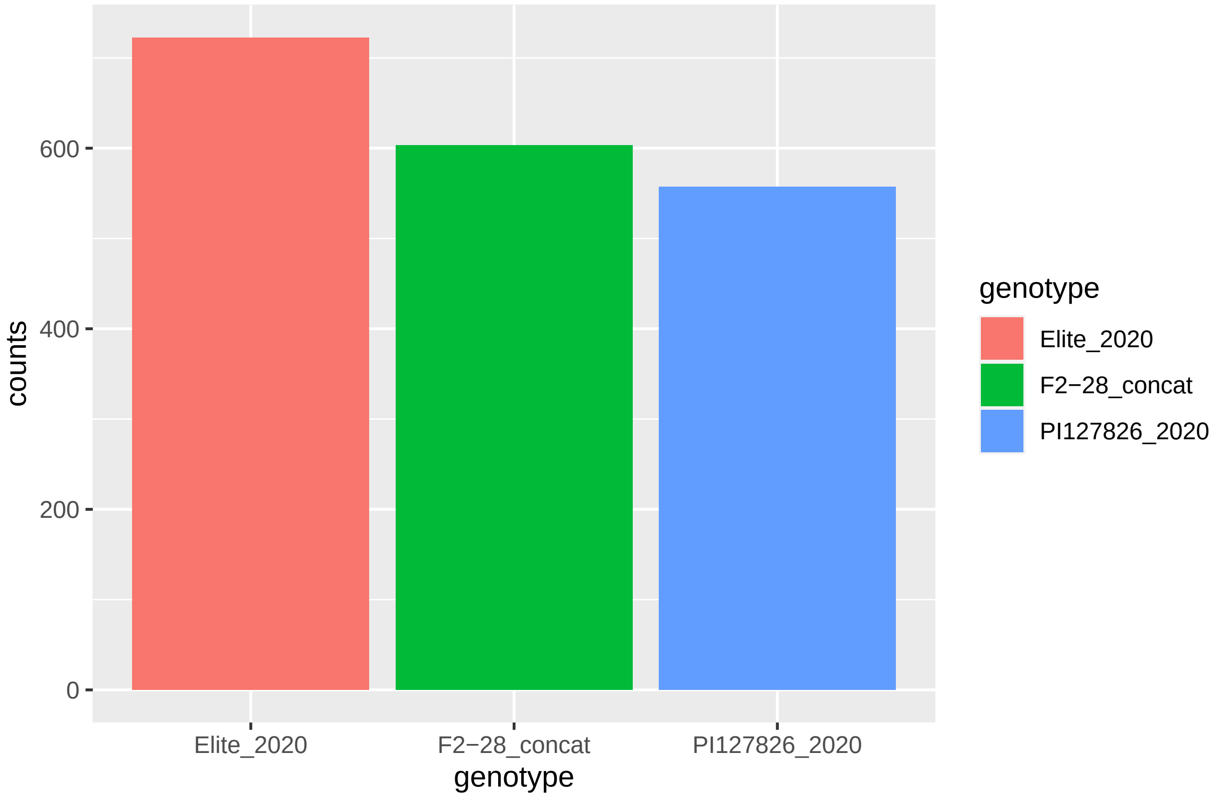

HMGS_Solyc08g007790.2.1
counts
200
100
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

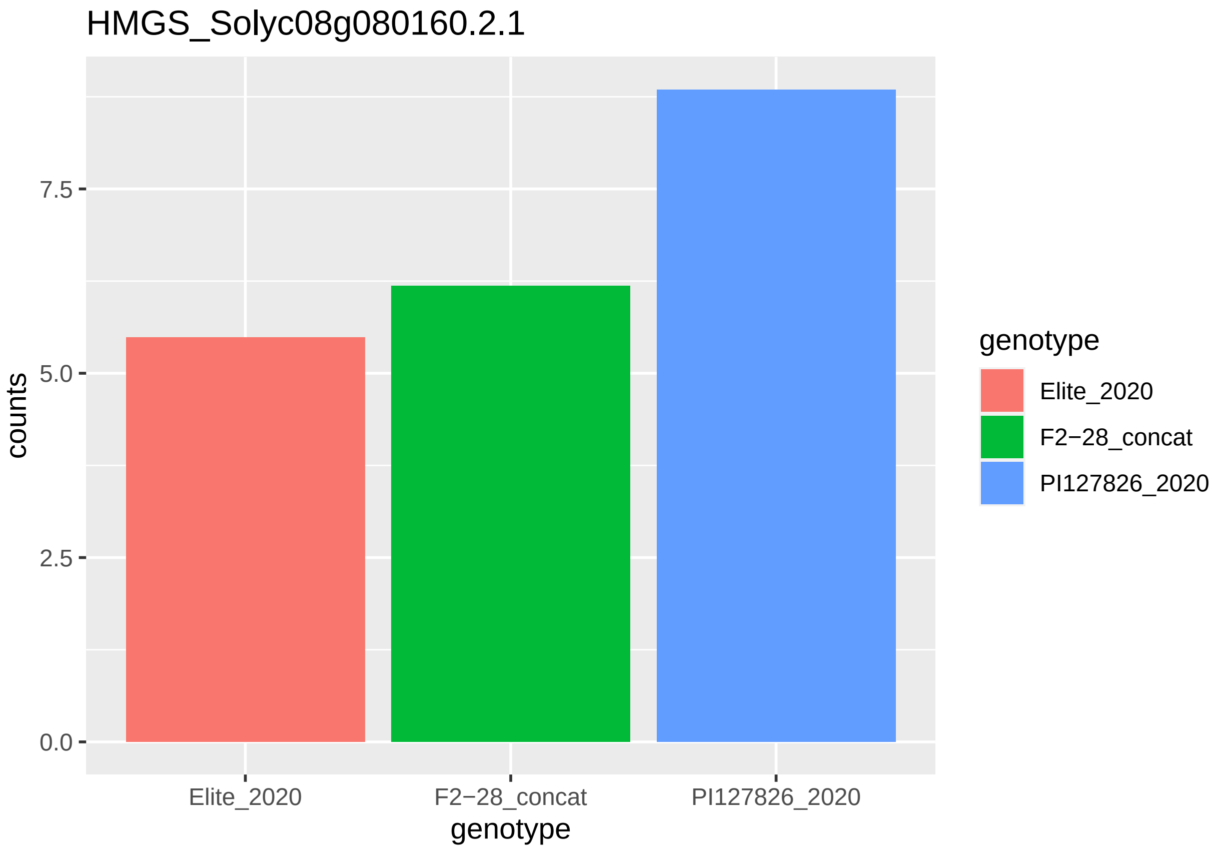

DXS_Solyc08g066950.2.1
counts
400
300
200
100
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

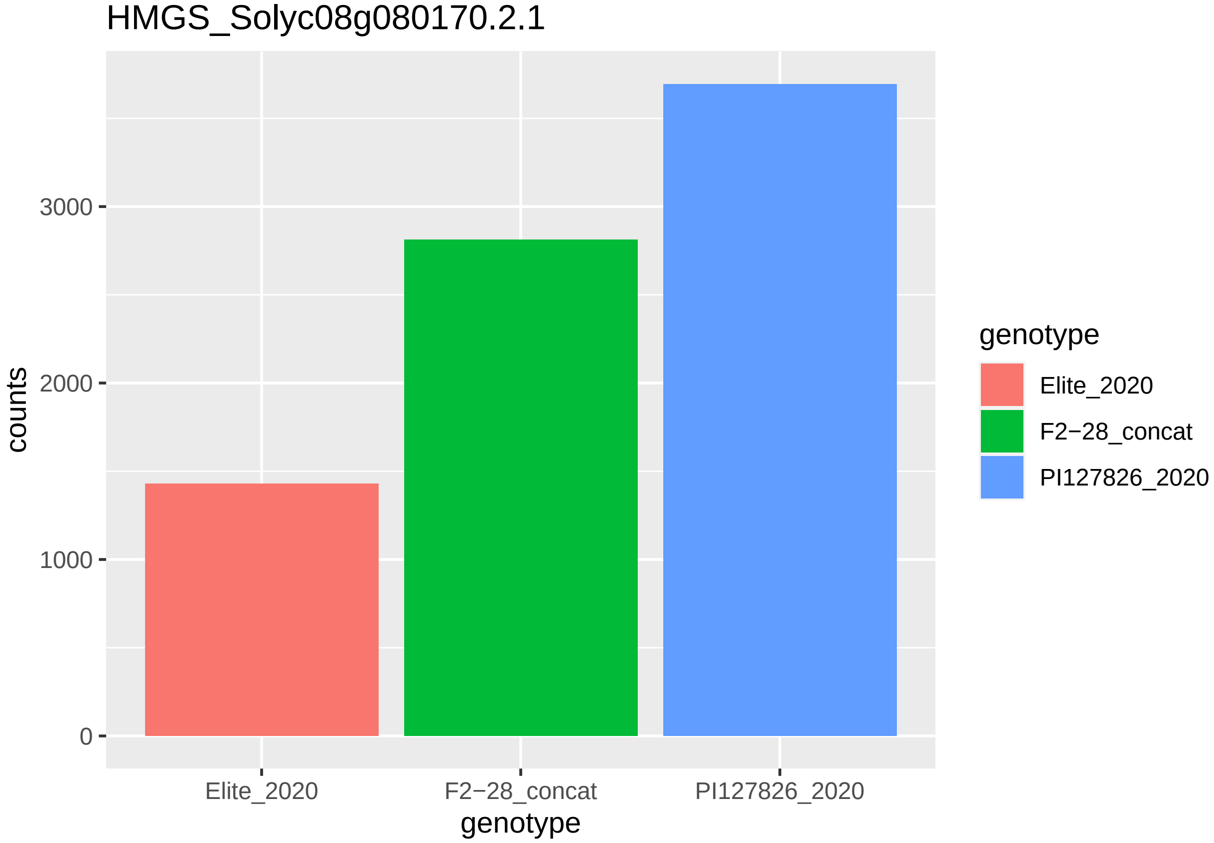
Nudix_Solyc08g075390.2.1

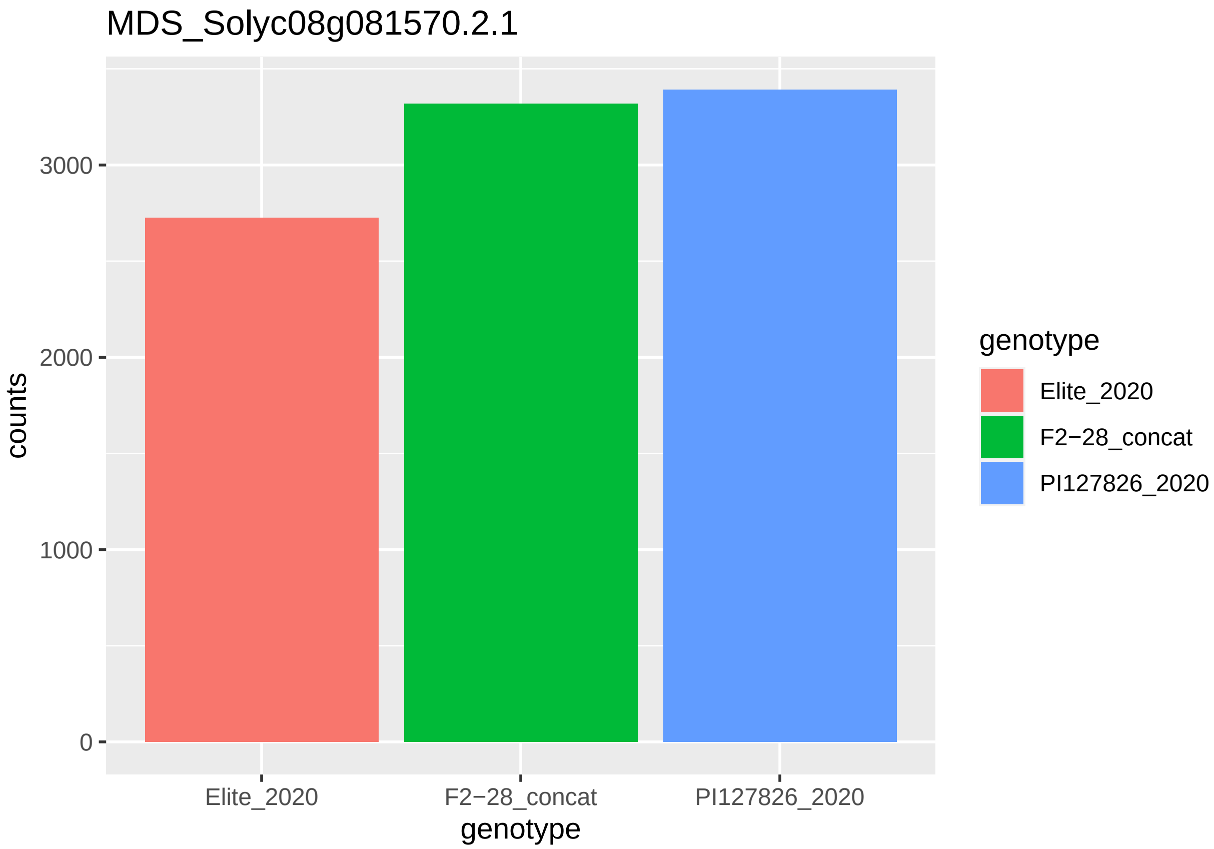


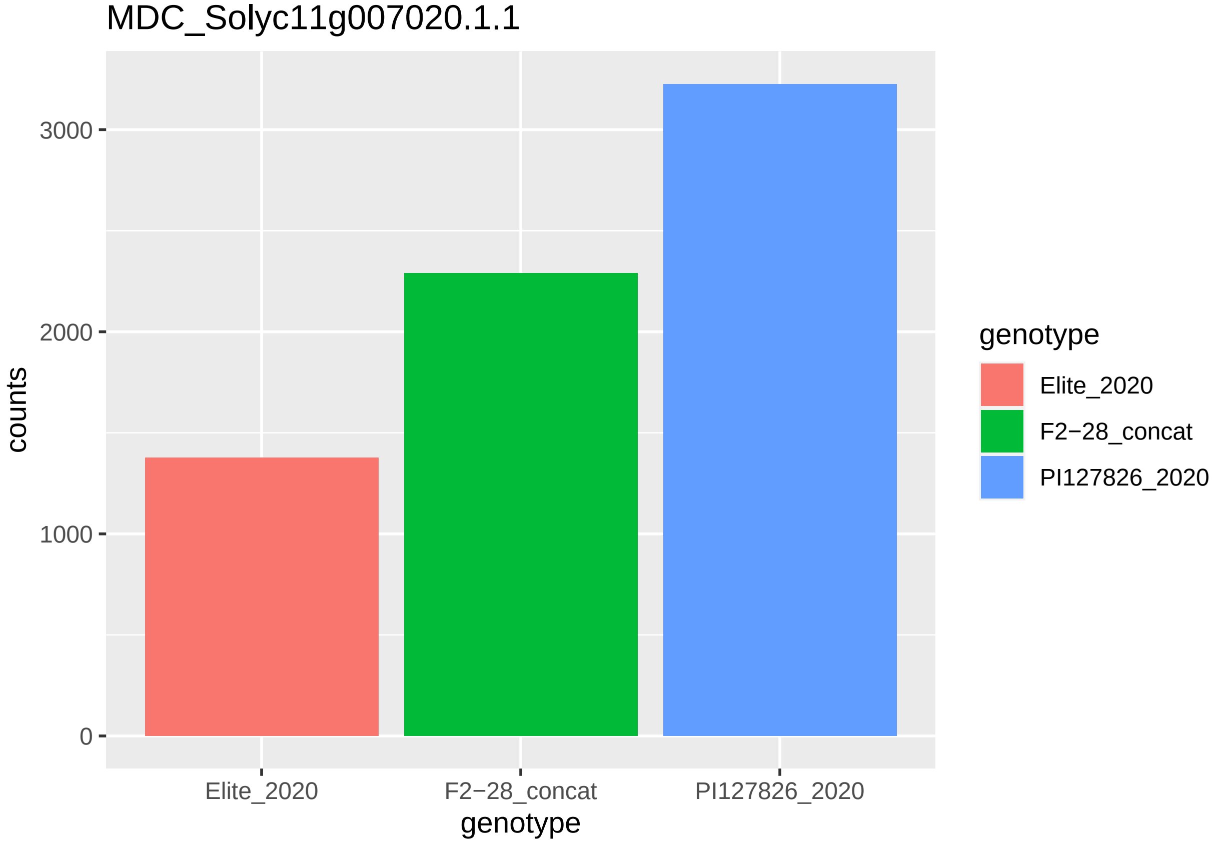
pMVK_Solyc08g076140.2.1

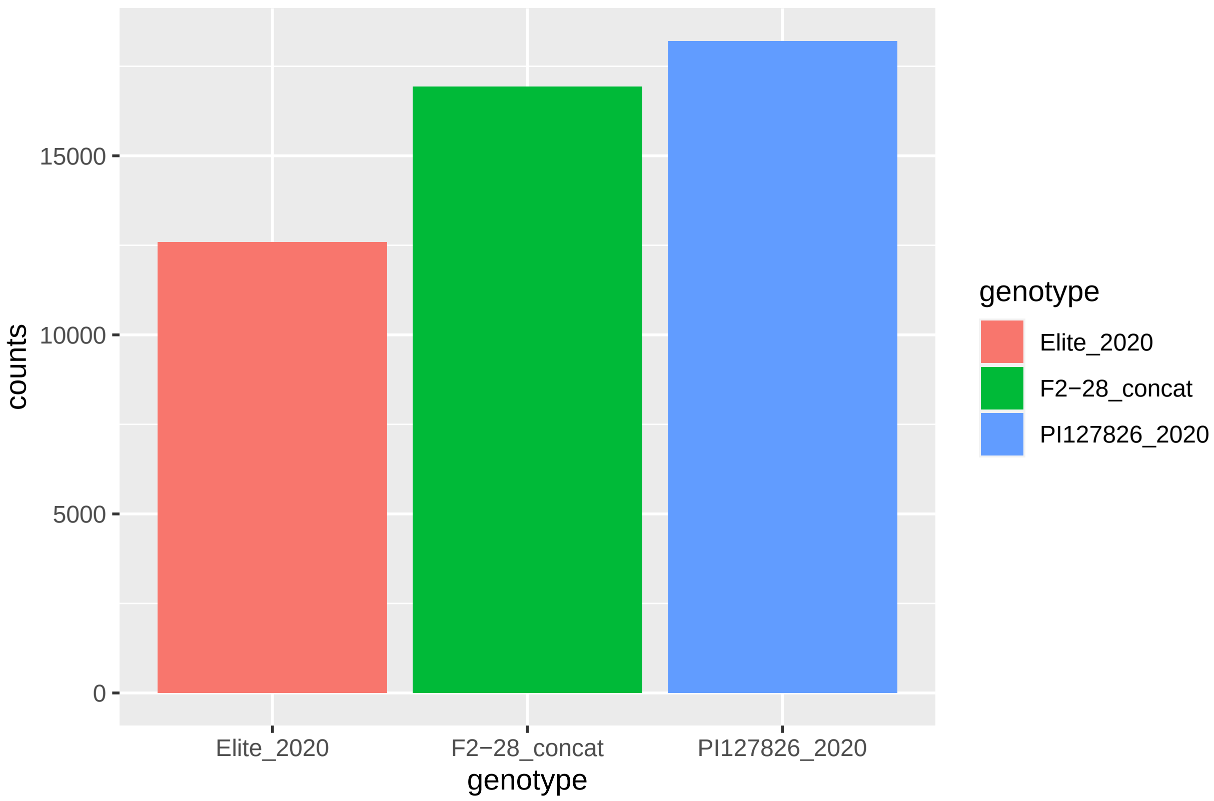


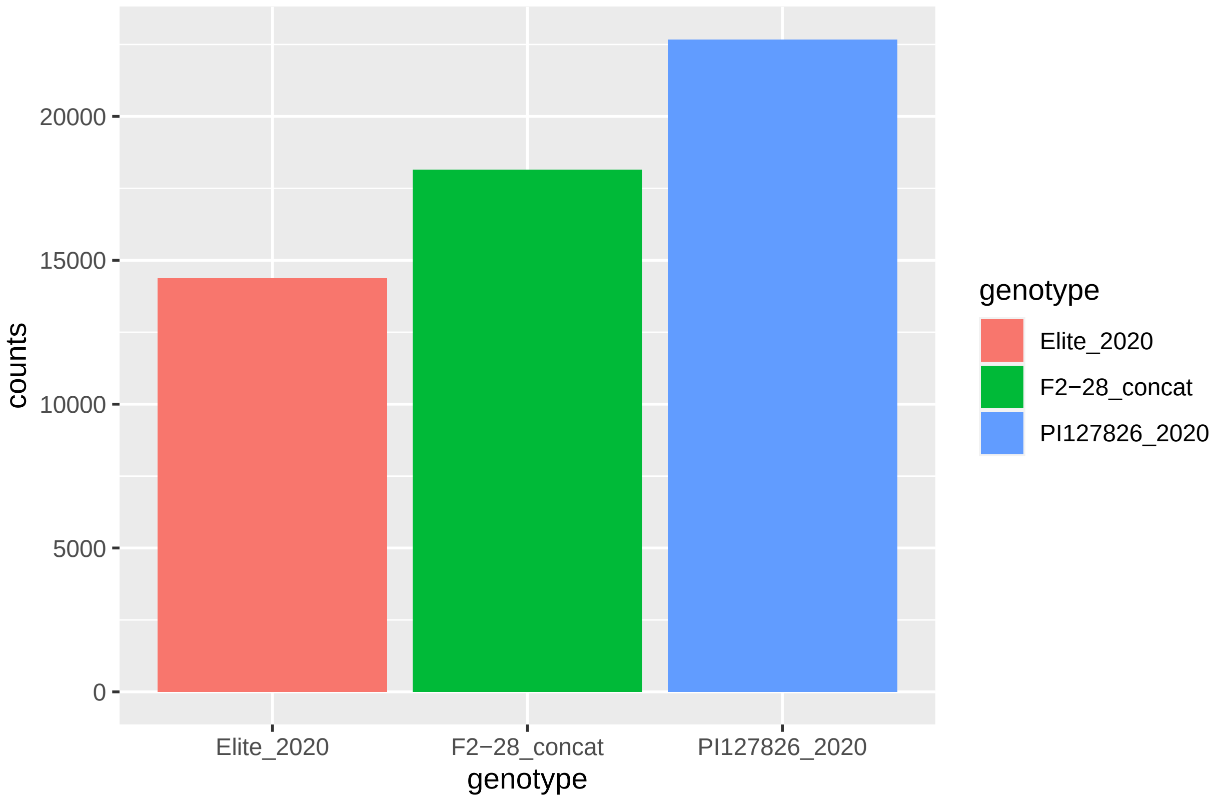

HMGS_Solyc08g080160.2.1
counts
0.0
2.5
5.0
7.5
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020

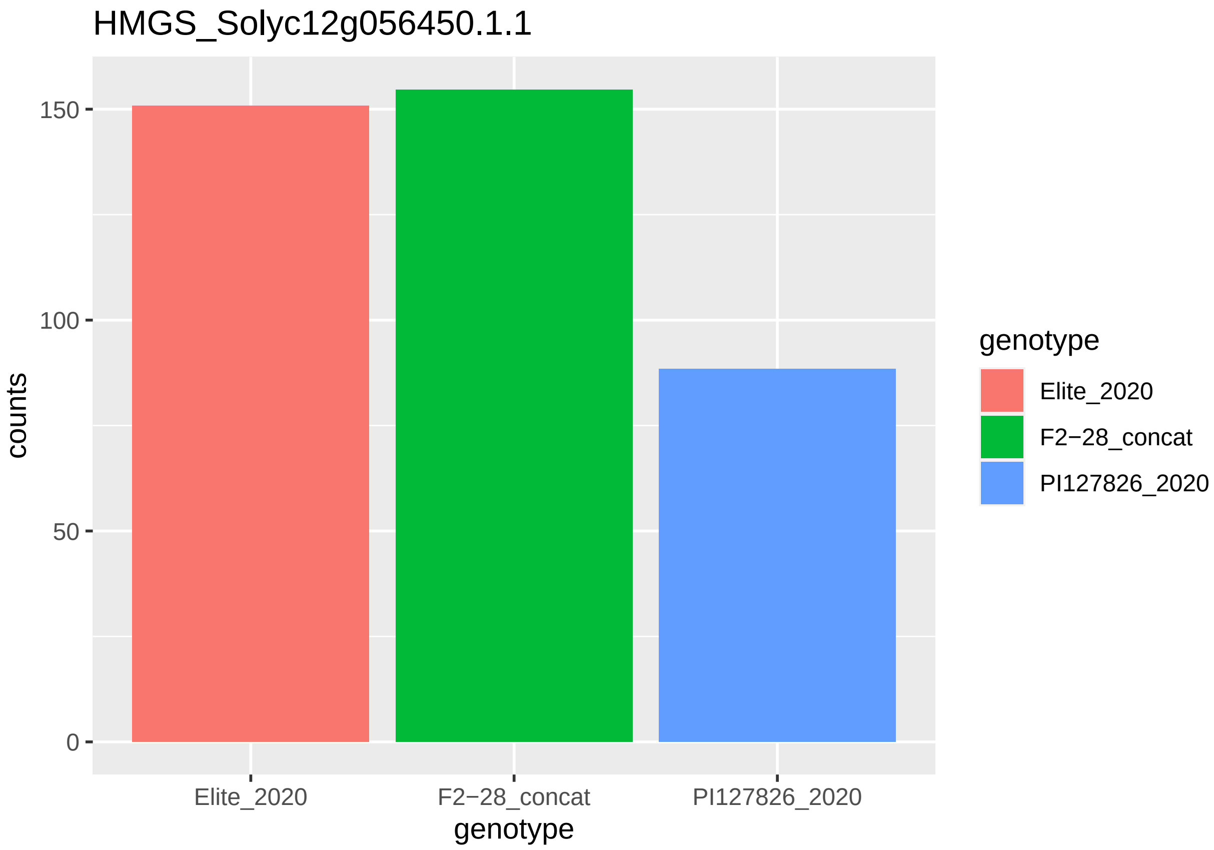

HMGS_Solyc08g080170.2.1
counts
0
1000
2000
3000
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020


MDS_Solyc08g081570.2.1
counts
3000
2000
1000
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020


MDC_Solyc11g007020.1.1
counts
3000
2000
1000
0
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020


DXS_Solyc11g010850.1.1


HDS_Solyc11g069380.1.1

HMGS_Solyc12g056450.1.1
150
100
50
0
counts
Elite_2020
F2−28_concat
PI127826_2020
genotype
genotype
Elite_2020
F2−28_concat
PI127826_2020