

Differential expression (F2.28, 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 F2.28 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

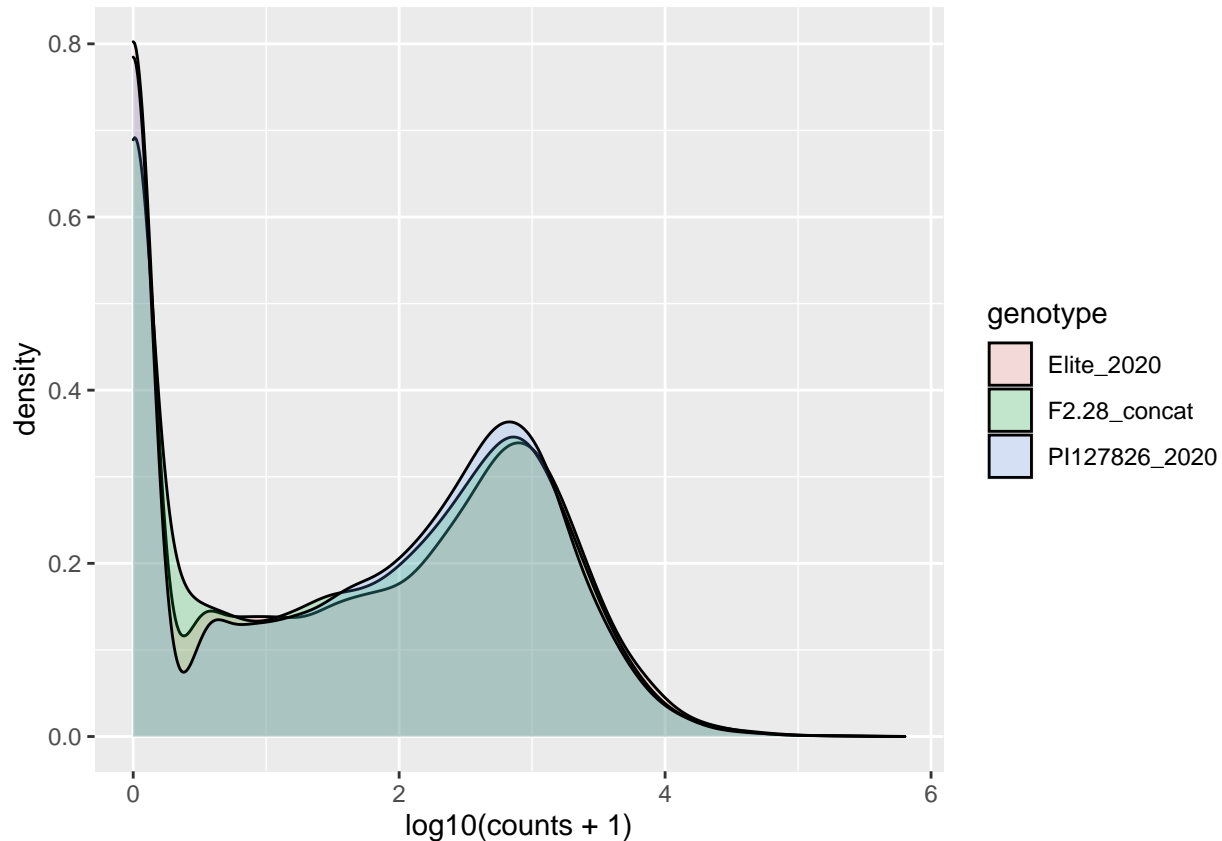
1.1 Import scaled counts

```
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", "F2.28_concat", "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 with low count values ($\log_{10} = 0.3$) is lower for F2.28 but seems to be comparable for other genes.

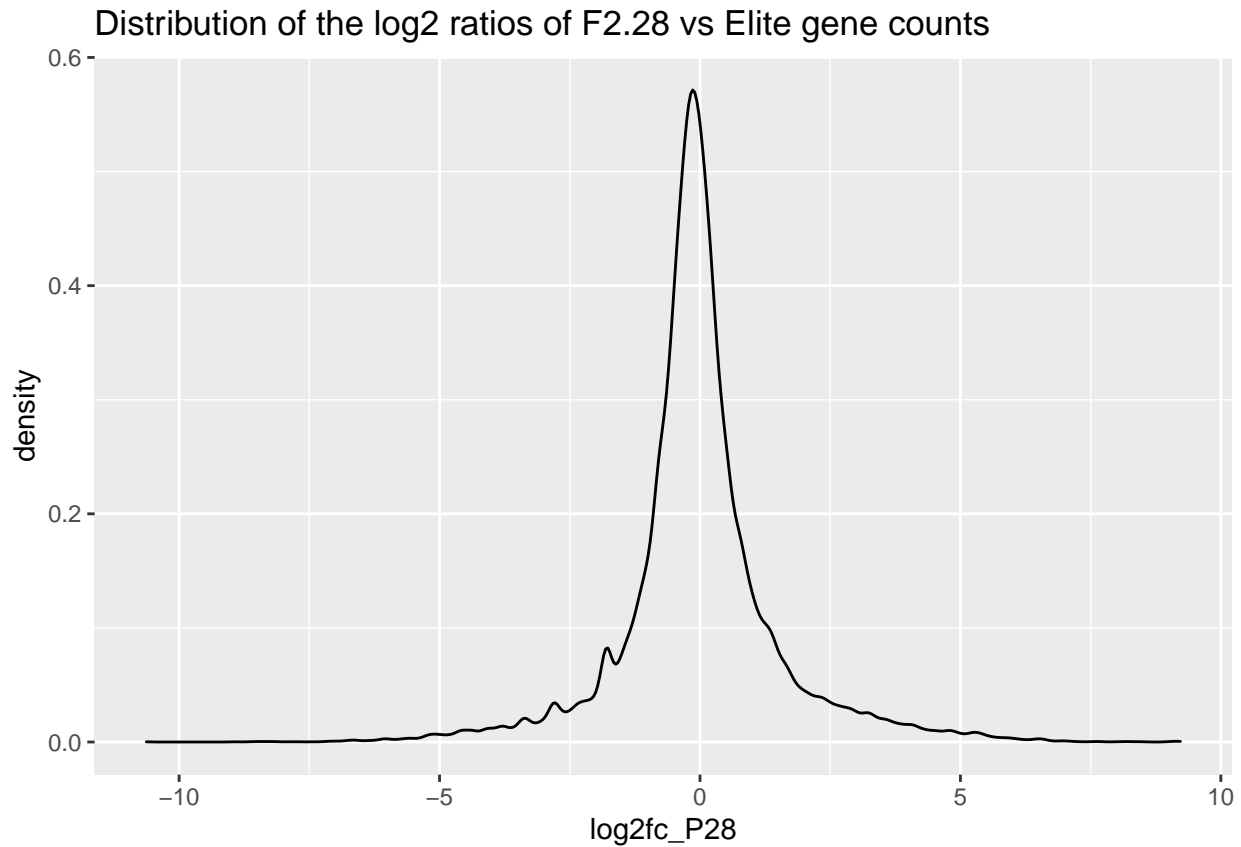
2.2 Plot log2ratio F2.28 vs Elite

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).
```

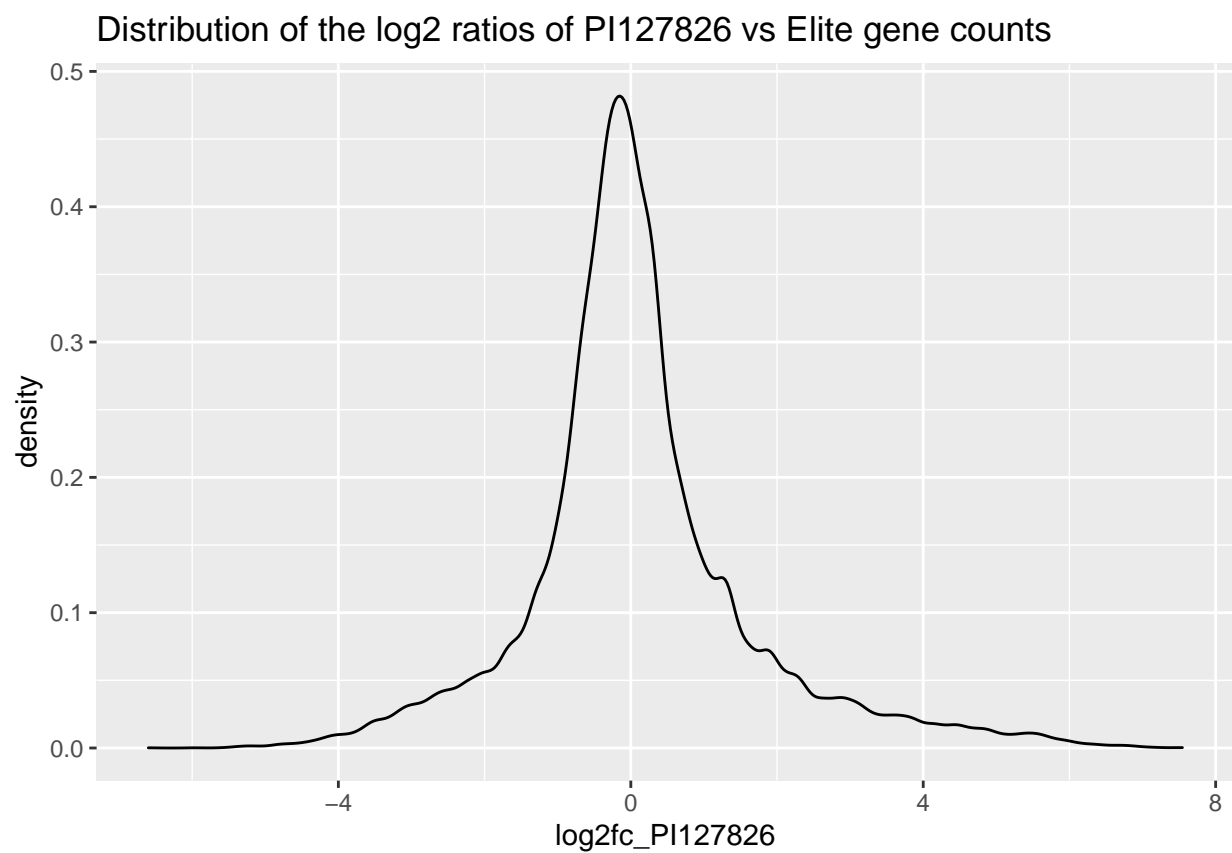


The log2ratio distribution looks OK.

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).
```



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   -1.00166699          -1.8945217
## 2 Solyc00g005050.2.1    0.05994873           0.3874636
## 3 Solyc00g005080.1.1    0.52189497           0.3071122
## 4 Solyc00g005150.1.1    1.75455573           2.3945750
## 5 Solyc00g005840.2.1    1.05504701           0.6485176
## 6 Solyc00g005860.1.1   -1.31460630          -3.0148159
```

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=FALSE)) %>%
  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~         9.22           6.67           6.18           4.23
## 2 Soly~         9.15           5.96           6.13           3.77
## 3 Soly~         9.14           6.80           6.12           4.31
## 4 Soly~         9.12           3.31           6.10           2.06
## 5 Soly~         8.99           5.15           6.02           3.25
## 6 Soly~         8.49           3.82           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

```
dir.create(path = "./tables_F2-28/", showWarnings = F, recursive = T)  
write.csv(final,  
  file = "tables_F2-28/diff_res_F2-28_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

Should be significant ($p < 0.05$) in F2.28 vs Elite.

```
signif_genes <- filter(final, pval_P28 < 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_P28, log2ratio_P127826, pval_P28, pval_P127826) %>%  
  knitr::kable()
```

name	gene	pathway	log2ratio_P28	log2ratio_P127826	pval_P28	pval_P127826
pMVK	Solyc06g066310.2.1	MVA	3.451938	3.474898	0.01070946	0.01499932

```
write.csv(mep_mva_gene_signif,  
          file = "tables_F2-28/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, `F2.28_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)
}
```

