Gene enrichment analysis in the supergene for haplo v pleo metrotic gene expression differences

Carlos Martinez Ruiz

18 June 2019

```
## [1] "Package ggplot2 version 3.1.1"
## [1] "Package GenomicFeatures version 1.36.1"
## [1] "Package AnnotationDbi version 1.46.0"
## [1] "Package Biobase version 2.44.0"
## [1] "Package GenomicRanges version 1.36.0"
## [1] "Package GenomeInfoDb version 1.20.0"
## [1] "Package IRanges version 2.18.1"
## [1] "Package S4Vectors version 0.22.0"
## [1] "Package BiocGenerics version 0.30.0"
## [1] "Package parallel version 3.6.0"
## [1] "Package stats4 version 3.6.0"
## [1] "Package readxl version 1.3.1"
## [1] "Package stats version 3.6.0"
## [1] "Package graphics version 3.6.0"
## [1] "Package grDevices version 3.6.0"
## [1] "Package utils version 3.6.0"
## [1] "Package datasets version 3.6.0"
## [1] "Package methods version 3.6.0"
## [1] "Package base version 3.6.0"
```

Load all the xls sheets containing the gene names of each module and merge them into a single data frame

Get the genome locations for each gene from the gnG annotation file (REFseq version, as the names of the scaffolds in the regions file are in this notation)

Load the genomic locations of the supergene and merge with the locations of the genes to obtain a table with gene name and position in the supergene

Once the information about gene position within the supergene has been generated, we can use it to answer questions about supergene loci enrichment in the modules, or simpler questions such as, are all gp9 genes in the supergene?

They are all in the supergene region, LOC105194481 is OBP-3 from (Pracana et al., 2017)https://onlinelibrary.wiley.com/doi/full/10.1002/evl3.22.

Enrichment per module

Run enrichment analyses, first based on expected vs observed numbers of supergene genes within each module.

In total, there are 640 genes in the supergene and 13973 outside. In green module: 6 genes in the supergene, 16 outside. Grey module: 20 genes in the supergene, 218 outside.

Plot these results against expected values for each module. The expectations of how many loci from either position in the genome there should be in any module are based in the proportion of supergene/non-supergene loci across the whole genome.

Supergene enrichment in DEGs

Use the p value per gene from all comparisons to check whether there is an enrichment of lower p values from genes in the supergene compared to genes in the rest of the genome.

Load the data for the significance level per gene

Assign a position within or otside the supergene to each gene in every comparison

Run a KS test comparing the distribution of p values (uncorrected!) between supergene and non-supergene loci for all comparisons. Store the results of the enrichment tests in a dataframe for all comparisons

There is enrichment for significance in supergene loci in SFQ vs GFQ at 25 days (independently of whether large or small)