Characterization of Dihydrochalcones in Apple: Identifying Differentially Expressed Genes in Apple Interspecific Hybrids

Abraham Porschet[1], Ben Gutierrez[2], Tori Meakem[2]

[1]Departments of Mathematics and Computer Science, Swarthmore College, Swarthmore, Pennsylvania
[2]United States Department of Agriculture-Agricultural Research Service, Plant Genetics Resource Unit,Geneva, NY 14456

Introduction

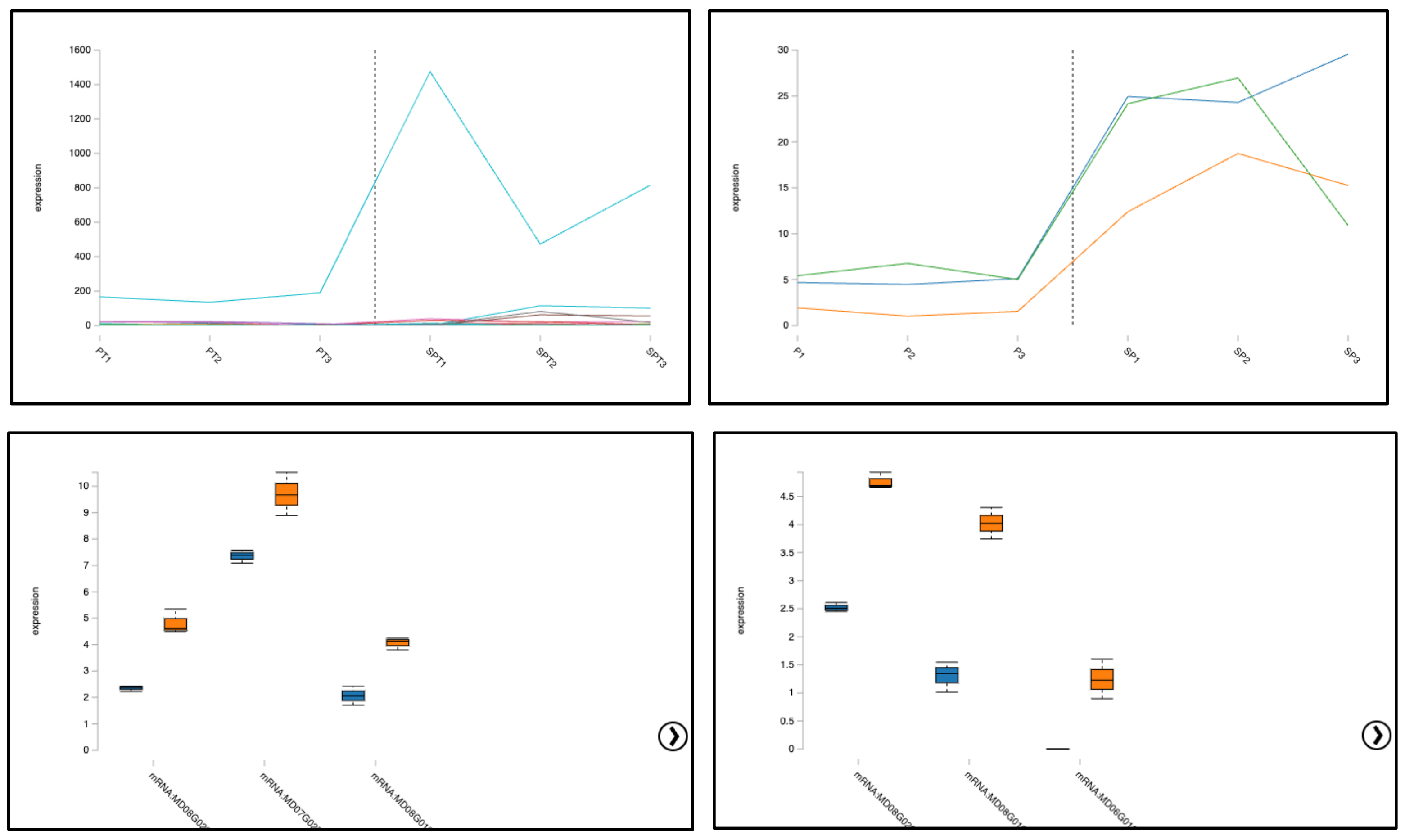
Some species of apples contain chemicals known as Dihydrochalcones (DHCs), which are known for their potential health effects: helping fight diabetes, cancer, inflammation, and herpes among other positives (Stompor et al 2019). Exploring the genomes of these apples can give researchers more information for future breeding efforts in order to increase the health benefits of various breeds of apples. This study explores if there are genes that predominantly control the expression of three DHCs: sieboldin (S), phloridzin (P), and trilobatin (T). Through RNA sequencing and differential expression, pairing sets of samples that express one of or combinations of the three chosen DHCs, we can see which which pairs of samples have statistically significant differences in the counts of specific genes.

Materials and Methods

Leaves were picked with four different phenotypes (P, PT, (S)P, and SPT) where (S)P contained both sieboldin and phloridzin but expressed as if it was just phloridzin. Each phenotype had four replicates and each replicate was pooled from four trees. Apple leaves were sequenced using RNA sequencing technology and then RNA-seq data was processed using R, bash scripting, and python, with the packages, limma, edgeR, DESeq2, and their various Bioconductor dependency packages in addition to the fastqc packages for data quality control. Inspiration was also taken from the dashboard displays of RaNA-seq.

Results

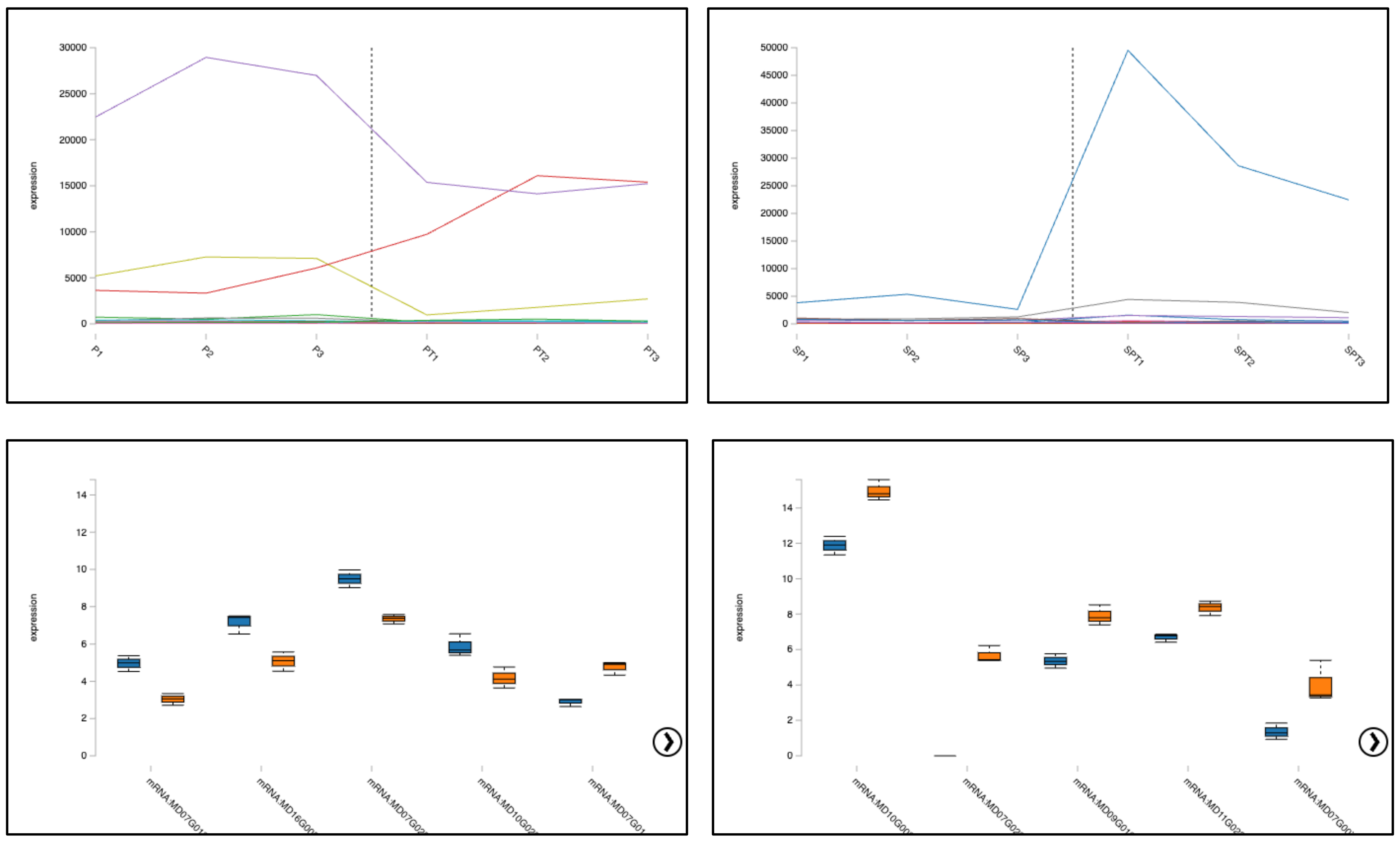
Using differential gene expression (DGE or DG), a method of finding which gene(s) are expressed with statistically significant numbers by comparing gene counts between transcriptomes (usually after some sort of statistical correction, in this case a Bonferroni correction to reduce probability of type 1 error) we compared the RNA-seq data of samples that included the following combinations of sieboldin, phloridzin, and trilobatin: P, SP, PT, and SPT.



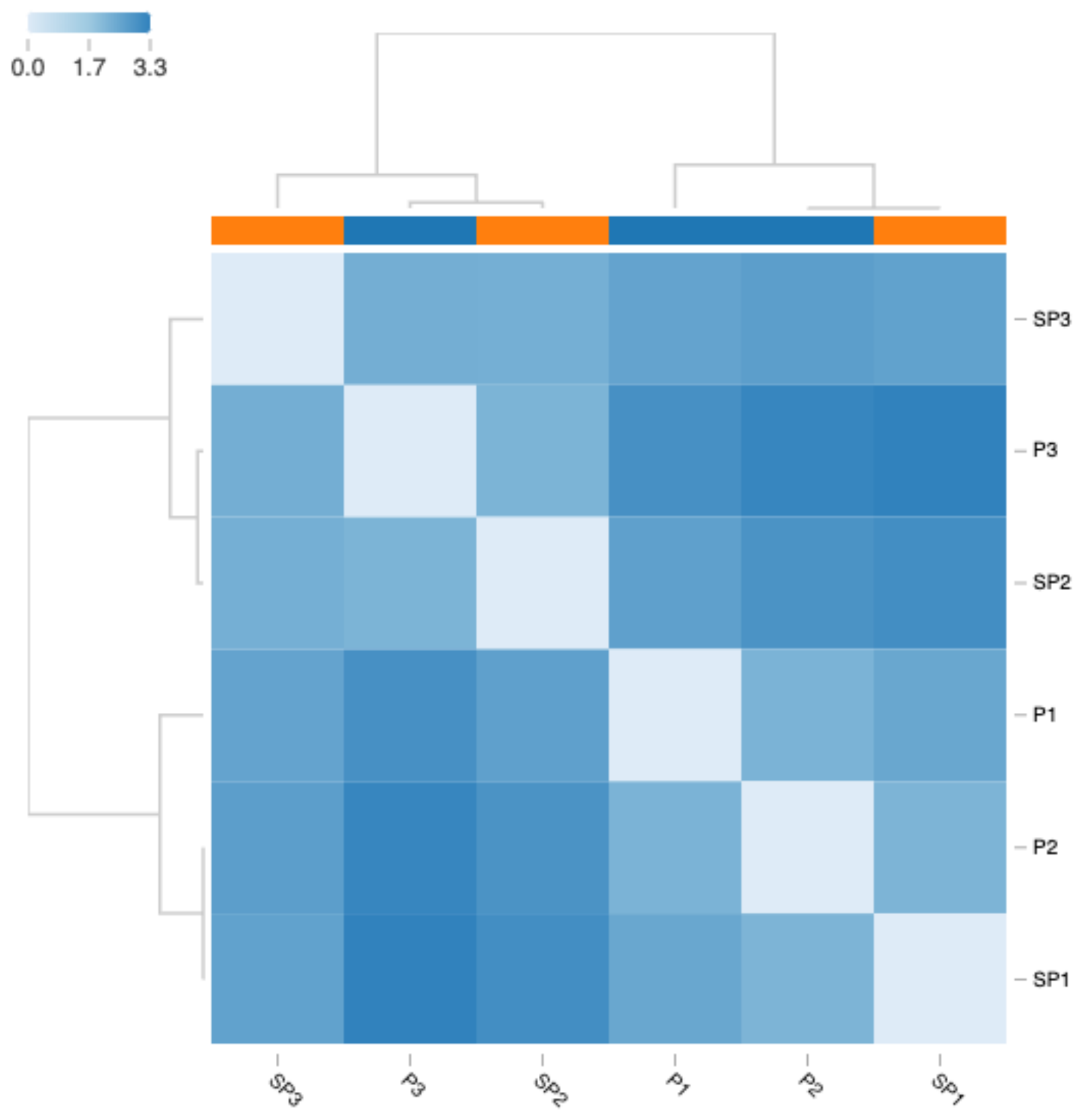
Figs 1-4 Comparing samples that express sieboldin vs those that don't

These top two plots show the normalized counts of the genes that tested as statistically significant. The genes are not labeled since their names are quite long, but the most significant gene for both pairings is in blue and is named MD08G0202400 and has evidence to be one of the key genes for sieboldin expression. However, both pairs of samples share counts of multiple genes that are statistically significant that could be dependent on each other for expression.

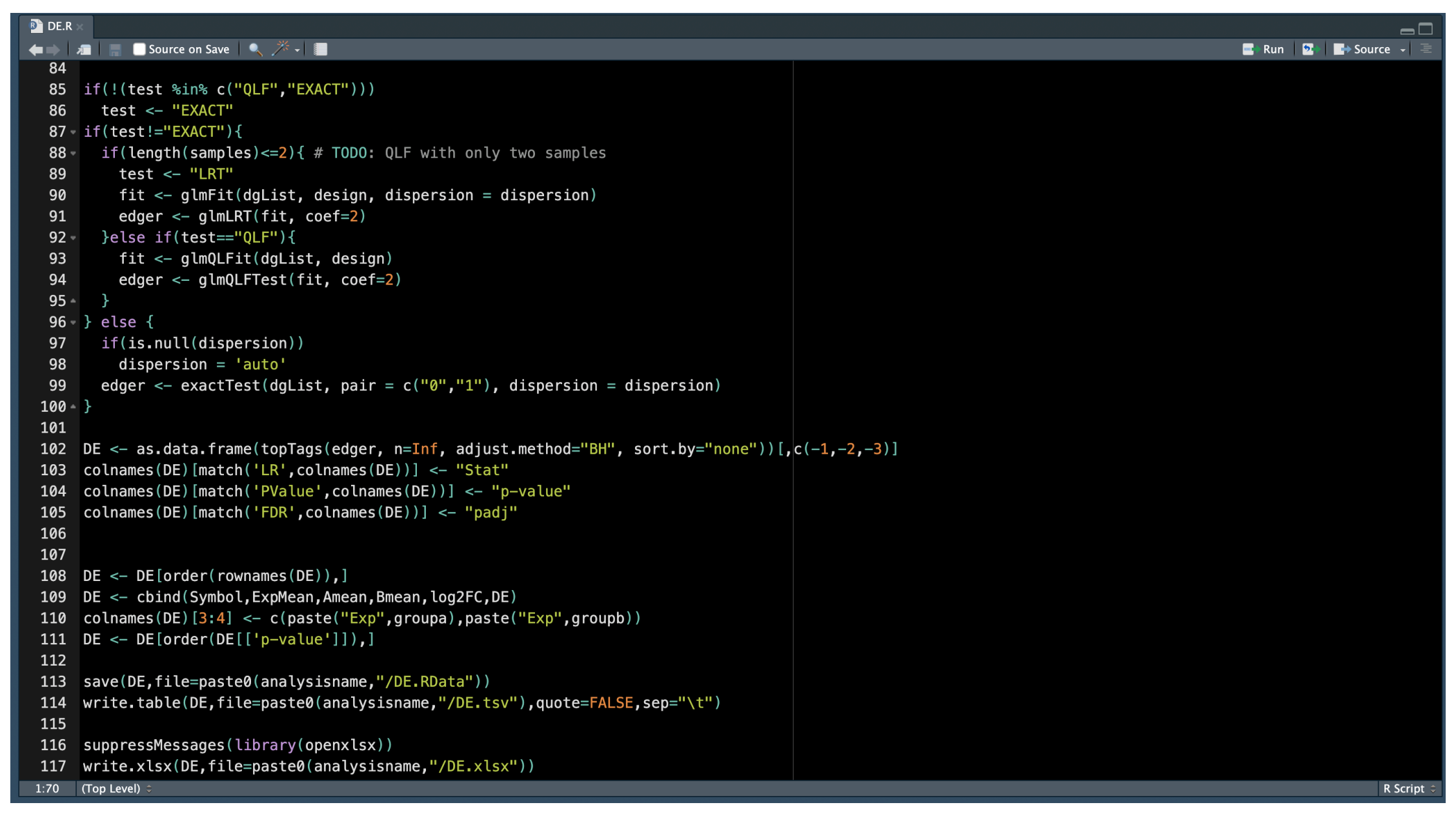
These bottom two plots show the same two pairs as above, but show the data for the counts for each of the significant genes. Plots on the left correspond with each other and the plots on right correspond as well



These plots compare the expression of genes between groups with trilobatin and groups without (**Figs. 5-8**). While both pairs of groups have multiple statistically significant differences in gene expression, they only share one gene in their top 10 genes with the lowest adjusted p-values (MD07G0233400) indicating that the gene shared in both could have great importance in the expression and production of trilobatin. This difference in DG results is probably due, at least in part, to the fact that SP also doesn't present sieboldin meaning that some of the genes given in the SP vs SPT pairing contribute to the expression of sieboldin as well as trilobatin



This heat map (**Fig. 9**) is a good expression of the ambiguity between the P and SP phenotypes since there is not much distinction between the groups as a whole, where within the same group there will be the similarly large differences as between the groups. The darker the box, the greater the difference between the two samples.



```r
if(!(test %in% c("QLF","EXACT")))
  test <- "EXACT"
if(test!="EXACT"){
  if(length(samples)<=2){ # TODO: QLF with only two samples
    test <- "LRT"
    fit <- glmFit(dgList, design, dispersion = dispersion)
    edger <- glmLRT(fit, coef=2)
  }else if(test=="QLF"){
    fit <- glmQLFit(dgList, design)
    edger <- glmQLFTest(fit, coef=2)
  }
} else {
  if(is.null(dispersion))
    dispersion = 'auto'
  edger <- exactTest(dgList, pair = c("0","1"), dispersion = dispersion)
}

DE <- as.data.frame(topTags(edger, n=Inf, adjust.method="BH", sort.by="none"))[,c(-1,-2,-3)]
colnames(DE)[match('LR',colnames(DE))] <- "Stat"
colnames(DE)[match('PValue',colnames(DE))] <- "p-value"
colnames(DE)[match('FDR',colnames(DE))] <- "padj"


DE <- DE[order(rownames(DE)),]
DE <- cbind(Symbol,ExpMean,Amean,Bmean,log2FC,DE)
colnames(DE)[3:4] <- c(paste("Exp",groupa),paste("Exp",groupb))
DE <- DE[order(DE[['p-value']]),]

save(DE,file=paste0(analysisname,"/DE.RData"))
write.table(DE,file=paste0(analysisname,"/DE.tsv"),quote=FALSE,sep="\t")

suppressMessages(library(openxlsx))
write.xlsx(DE,file=paste0(analysisname,"/DE.xlsx"))
```

Fig. 10, a snippet of R code to Calculate differential expression

Note: These results were taken using the edgeR package and testing method since limma's linear testing method is meant for more continuous values and DESeq2 brought on too much type 1 error (Proved through wilcoxon rank-sum and edgeR results).

Conclusions

- There are a few genes that strongly correlate to the expression of sieboldin, most notably MD08G0202400.
- However, finding genes that correlate to the expression of trilobatin is more complicated because of the nature of the SP phenotype, but MD07G0233400 is promising.
- For the statistically more complicated question of what causes the expression of trilobatin, machine learning clustering methods could help bring more clarity to the dependency problems mentioned above.

References

- Gutierrez, Benjamin & Arro, Jie & Zhong, Gan-Yuan & Brown, Susan. (2018). Linkage and association analysis of dihydrochalcones phloridzin, sieboldin, and trilobatin in Malus. Tree Genetics & Genomes. 14. 10.1007/s11295-018-1304-7.
- Stompor M, Broda D, Bajek-Bil A. Dihydrochalcones: Methods of Acquisition and Pharmacological Properties—A First Systematic Review. *Molecules*. 2019; 24(24):4468.
- Carlos Prieto, David Barrios, RaNA-Seq: interactive RNA-Seq analysis from FASTQ files to functional analysis, *Bioinformatics*, Volume 36, Issue 6, 15 March 2020, Pages 1955–1956

Acknowledgments

I am grateful to Ben Gutierrez and Victoria Meakem for their scientific support and mentoring as well as their hand in the RNA sequencing of the leaves. I am also thankful for Samantha Page, Anna Cohen, and Marcella Venettozzi

United States Department of Agriculture
National Institute of Food and Agriculture