Transcriptomics of symptomatic hosts, potato and mint, and asymptomatic host, mustard, during infection with host-adapted isolates of *Verticillim dahliae*

Authors: David Linnard Wheeler | Jeness Scott | Jeremiah Kam Sung Dung | Dennis Johnson

Table of Contents:

Objectives

Back to Table of Contents

- Characterize the differentially expressed genes involved in symptomatic (potato and mint) and asymptomatic interactions (mustard) between hosts and *Verticillium dahliae*

Hypotheses

Back to Table of Contents

- **Science H_o 1**: There are no differentially expressed genes (i) between symptomatic and asymptomatic hosts, (ii) between isolates within a host, and (iii) between hosts within an isolate.
- **Science H_o 2**: Symptomatic and asymptomatic hosts exhibit similar responses to $V.\ dahliae$ infection
- **Science H_o 3**: Gene expression of $V.\ dahliae$ does not differ accross fungal strains or between asymptomatic and symptomatic hosts
- **Statistical H_o**:
    - Observed variation in DEG across treatments represents random variation, not systematic effects of hosts or isolates Variation in the DEG is unrelated to variation in the hosts and isolates and is no greater than expected by chance or sampling error.
    - More formally:

      $$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

      where the counts, K_{ij} for each gene, i, and sample, j, follow a negative binomial with the mean,μ_{ij}, and dispersion parameter for each gene, α_i. The dispersion parameter,α_i, describes the relationship between variance of an observed count and its mean value- the expected distance of the observed count from its mean. The mean,μ_{ij}, can be decomposed into a sample-specific size factor, s_j, and a parameter,q_{ij}, that is proportional to the expected concentation of transcripts for sample j:

      $$\mu_{ij} = s_j q_{ij}$$

      Log2 fold changes for gene i in each column of the model matrix, X,are provided by the coefficients, β_i:

      $$log_2(q_{ij}) = x_{j.}\beta_i$$
    - In short, the effect sizes between the groups are 0.

Experimental design

Back to Table of Contents

- **Treatment structure**: 2 way factorial
    - **Independent variables**:
        - 3 cultivars:
            1. Potato
            2. Mint
            3. Mustard
        - 3 fungi:
            1. $Verticillium\ dahliae\ 653$
            2. $Verticillium\ dahliae\ 111$
            3. Non-inoculated control
        - 1 time point:
            - 10 days after inoculation
        - 3 replicates
    - **Dependent variables**:

- Constructs:
  - Gene expression
- Variables:
  - Counts of RNA transcripts

- **Design structure**: randomized complete block design
- **Observational unit**: plant
- **Experimental unit**: plant
- **Samples**: whole plants
- **Data**:
  - RNA quantity and quality
  - Counts of RNA transcripts
- **Analysis**:
  - Differential gene expression analyses

Materials and Methods

Back to Table of Contents

Inoculum preparation for root dips (3.5"pot):

- Grow Verticillium dahliae 653 and 111 on separate plates of PDA agar at room temperature/
- Harvest 0.5 cm cores from each plate.
- Add one core per one 200 ml flask filled with 125ml of PDA broth.
- Incubate cultures at room temp/22 C for 7-10 days in the dark.
  - Spin at 125 RPM.
- Filter inoculum through two layers of sterilized cheesecloth with vacuum filter.
- Quantify inoculum with hemocytometer.
- Dilute inoculum to 10^6 conidia/mL with sterilized diH20.
- Inoculate via root drench.
  - Pour 100 mL of 10^6 conidia/mL inoculum over the soil/turface surface.
  - I did this mostly in lieu of watering. For example, if you normally give the plants 100 mL of water/day, give them 0 mL of water and inoculate; if you normally give the plants 200 mL of water/day, give them 100mL of water and inoculate.

- Inoculum for isolates 653 and 111:
  - 180 plants/3 isolates = 60 plants/isolate * 100 ml/plant (Dung et al. 2010) = 6000 ml = 6 L inoculum
  - 6 L of 10^6 conidia/ml inoculum is needed
  - 6 L/200 ml/flask = 30 flasks
- Trial one planted: 5/1/2018
- Trial one inoculated: 5/19/2018
- First Harvest: potato, mint, and mustards harvested @ 10 dpi on 5/29/2018

Open data

Back to Table of Contents

- **Install and invoke packages**

In [1]:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("DEFormats", version = "3.8")
```

```
Bioconductor version 3.8 (BiocManager 1.30.4), R 3.5.2 (2018-12-20)
Installing package(s) 'DEFormats'


The downloaded binary packages are in
        /var/folders/8c/7fwkqlvd4ps_rj6zf9lr_xcw0000gn/T//RtmptF4nug/down
loaded_packages

Update old packages: 'annotate', 'assertthat', 'BiocInstaller', 'callr',
'cli',
  'colorspace', 'gtable', 'highr', 'knitr', 'lazyeval', 'openssl', 'pkgbu
ild',
  'processx', 'purrr', 'Rcpp', 'RcppArmadillo', 'readxl', 'rgdal', 'rlan
g',
  'rmarkdown', 'rstudioapi', 'spam', 'sys', 'tibble', 'tinytex', 'XML',
  'diffobj', 'e1071', 'fs', 'git2r', 'Matrix', 'mgcv', 'zoo'
```

In [2]:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("apeglm", version = "3.8")
```

```
Bioconductor version 3.8 (BiocManager 1.30.4), R 3.5.2 (2018-12-20)
Installing package(s) 'apeglm'


The downloaded binary packages are in
        /var/folders/8c/7fwkqlvd4ps_rj6zf9lr_xcw0000gn/T//RtmptF4nug/down
loaded_packages

Update old packages: 'annotate', 'assertthat', 'BiocInstaller', 'callr',
'cli',
  'colorspace', 'gtable', 'highr', 'knitr', 'lazyeval', 'openssl', 'pkgbu
ild',
  'processx', 'purrr', 'Rcpp', 'RcppArmadillo', 'readxl', 'rgdal', 'rlan
g',
  'rmarkdown', 'rstudioapi', 'spam', 'sys', 'tibble', 'tinytex', 'XML',
  'diffobj', 'e1071', 'fs', 'git2r', 'Matrix', 'mgcv', 'zoo'
```

In [10]:
```
library("data.table")
library("tibble")
library("eulerr")
library("apeglm")
library("DESeq2")
library("edgeR")
library("DEFormats")
library("dplyr")
library("ggplot2")
library("reshape2")
library("pheatmap")
library("RColorBrewer")
library("PoiClaClu")
library("ggbeeswarm")
library("EnhancedVolcano")
library("devtools")
library("gridExtra")
library("grid")
library("cowplot")
library("genefilter")
library("viridis")
library("VennDiagram")
library("prob")
library("seqinr")
library("stringr")
```

```
Loading required package: S4Vectors
Loading required package: stats4
Loading required package: BiocGenerics
Loading required package: parallel

Attaching package: ‘BiocGenerics’

The following objects are masked from ‘package:parallel’:

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, parApply, parCapply, parLapply,
    parLapplyLB, parRapply, parSapply, parSapplyLB

The following objects are masked from ‘package:stats’:

    IQR, mad, sd, var, xtabs

The following objects are masked from ‘package:base’:
```

- Grab working directory

In [11]:
```
getwd()
```

'/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/SCRIPTS'

- Set working directory

In [12]:
```
setwd("/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/DATA/R_FIL
```

- **Open data**

Fragments

In [13]:
```
DF = read.csv("Mentha_reads.csv",header=T)
```

Gene names

In [14]:
```
gnDF = read.csv("/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/
                header=T)
```

In [15]:
```
names(gnDF)[2]<-"id"
names(gnDF)[3]<-"Comparison"
names(gnDF)[6]<-"KO.Name"
names(gnDF)[10]<-"Hit1_acc"
```

Sequences

In [16]:
```
fastafile <- read.fasta(file = "/Users/davidwheeler/Desktop/RESEARCH/Data/T
                        seqtype = "AA",as.string = TRUE, set.attributes = FA
```

Gene ontology

In [17]:
```
GO = read.csv("/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/DA
                header=T)
```

Curate data

Back to Table of Contents

- **Set first column to index**

In [25]:
```
DF_1 <- data.frame(DF[,-1], row.names = DF[,1])
```

In [26]:
```
head(DF_1)
```

	S2_3_2_1	S2_3_2_9	S2_3_2_4	S2_1_2_1	S2_1_2_4	S2_1_2_5	S2_2_2_2	S2_2_2_5	S
Cluster-67248.142691	4.03	0.00	0.00	3.72	2.16	1.92	0.00	8.79	
Cluster-67248.107952	106.65	67.52	77.12	161.88	114.64	188.30	176.11	144.88	
Cluster-58782.0	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	
Cluster-67248.152869	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Cluster-67248.17374	143.91	112.02	101.26	253.87	162.03	236.01	226.94	263.46	
Cluster-67248.56631	2.31	7.86	6.27	11.81	3.72	0.00	4.92	0.00	

- **Rounds floats/decimals to integer counts: since these data were generated *de novo* decimals are abound**

In [27]:
```
DF = round(DF_1, digits = 0)
```

- **Create DGEList Object**

> - Convert dataframe to matrix

In [28]:
```
df = data.matrix(DF)
```

- **Vector for column/treatment names**

In [29]:
```
group = rep(c("Control", "653", "111"), each = 3)
```

In [30]:
```
dge = DGEList(df, group = group)
```

```
In [31]: dge
```

$counts

	S2_3_2_1	S2_3_2_9	S2_3_2_4	S2_1_2_1	S2_1_2_4	S2_1_2_5	S2_2_2_2	S2_2_2_5	S2
Cluster-67248.142691	4	0	0	4	2	2	0	9	
Cluster-67248.107952	107	68	77	162	115	188	176	145	
Cluster-58782.0	0	1	0	0	0	0	0	0	
Cluster-67248.152869	0	0	0	0	0	0	0	0	
Cluster-67248.17374	144	112	101	254	162	236	227	263	
Cluster-67248.56631	2	8	6	12	4	0	5	0	
Cluster-73865.0	0	3	0	7	6	6	0	3	

- **Coerce DGElist to DESeqDataSet**

```
In [32]: dds = as.DESeqDataSet(dge)
```

```
converting counts to integer mode
  it appears that the last variable in the design formula, 'group',
  has a factor level, 'Control', which is not the reference level. we rec
ommend
  to use factor(...,levels=...) or relevel() to set this as the reference
level
  before proceeding. for more information, please see the 'Note on factor
levels'
  in vignette('DESeq2').
```

```
In [33]: dds
```

```
class: DESeqDataSet
dim: 266009 9
metadata(1): version
assays(1): counts
rownames(266009): Cluster-67248.142691 Cluster-67248.107952 ...
  Cluster-67248.27096 Cluster-67248.132887
rowData names(0):
colnames(9): S2_3_2_1 S2_3_2_9 ... S2_2_2_5 S2_2_2_9
colData names(3): group lib.size norm.factors
```

Gene names data

Subset data

In [34]:
```
gnDF = gnDF[,c("id","Comparison","KO.Name","Hit1_acc")]
```

Summary Statistics

Back to Table of Contents

Tabulate DEG data (Comparison - basemean -direction of regulation - $\log_2$-fold change - p-value - adjusted p-value - gene name - function)

DEGs from 653 vs Control

In [514]:
```
Cv653_gn = subset(Cv653_GeneNames, Comparison == "Cv653")
```

- Order genes by fold change

In [515]:
```
table_cv6 <- (Cv653_gn[order((-Cv653_gn$log2FoldChange)),c(8,1,2,3,6,7,10)]
```

- Round digits

In [516]:
```
table_cv6$baseMean <- round(table_cv6$baseMean, 1)
table_cv6$log2FoldChange <- round(table_cv6$log2FoldChange, 1)
table_cv6$pvalue <- round(table_cv6$pvalue, 7)
table_cv6$padj <- round(table_cv6$padj, 7)
```

- Format columns

> - Comparison

In [517]:
```
table_cv6$Comparison = "control vs 653"
```

> - Host

In [518]:
```
table_cv6$Host = "Mentha x piperita"
```

> - Gene function

In [519]:

```
table_cv6$Function = NA
```

- Sequence

In [520]:

```
table_cv6$Sequence = NA
```

- Reorder columns

In [521]:

```
table_cv6 = table_cv6[,c(8,1,3:7,2,9:10)]
```

- Grab top 5 genes

In [522]:

```
Up_cv6 = head(table_cv6,n=5)
```

- Add column for regulation status

In [523]:

```
Up_cv6$Regulation = "up"
```

- Grab botton 5 genes

In [524]:

```
Down_cv6 = tail(table_cv6,n=5)
```

- Add column for regulation status

In [525]:

```
Down_cv6$Regulation = "down"
```

- Table for control vs 653

In [526]:

```
Table_cv6 = rbind(Up_cv6,Down_cv6)
```

In [527]:

```
Table_cv6
```

	Host	Comparison	baseMean	log2FoldChange	pvalue	padj	Hit1_acc	
1006	Mentha x piperita	control vs 653	124.3	3.5	0e+00	0.0000000	FB30_ARATH	C 67248.
1411	Mentha x piperita	control vs 653	41.0	3.5	0e+00	0.0000001	PMTK_ARATH	C 67248.
738	Mentha x piperita	control vs 653	121.8	3.4	0e+00	0.0000000	CO1A1_HUMAN	C 67248.
764	Mentha x piperita	control vs 653	20.3	2.8	3e-07	0.0002652	-	C 67248.1
1821	Mentha x piperita	control vs 653	21.3	2.3	0e+00	0.0000221	-	C 71
1747	Mentha x piperita	control vs 653	19.3	-3.2	0e+00	0.0000004	CB21_SINAL	C 67248.
544	Mentha x piperita	control vs 653	90.7	-3.6	0e+00	0.0000000	ARP3_ARATH	C 67248.1
49	Mentha x piperita	control vs 653	28.4	-3.8	0e+00	0.0000000	RBS2_BRANA	C 6
1017	Mentha x piperita	control vs 653	114.4	-6.1	0e+00	0.0000000	RBS2_BRANA	C 6724
1174	Mentha x piperita	control vs 653	203.0	-6.3	0e+00	0.0000000	CNGC5_ARATH	C 67248.

- Rearrange columns

In [528]:

```
Table_cv6 = Table_cv6[,c(1:2,11,3:10)]
```

In [529]:
```
Table_cv6
```

	Host	Comparison	Regulation	baseMean	log2FoldChange	pvalue	padj	Hit1_
1006	Mentha x piperita	control vs 653	up	124.3	3.5	0e+00	0.0000000	FB30_AR
1411	Mentha x piperita	control vs 653	up	41.0	3.5	0e+00	0.0000001	PMTK_AR
738	Mentha x piperita	control vs 653	up	121.8	3.4	0e+00	0.0000000	CO1A1_HUM
764	Mentha x piperita	control vs 653	up	20.3	2.8	3e-07	0.0002652	
1821	Mentha x piperita	control vs 653	up	21.3	2.3	0e+00	0.0000221	
1747	Mentha x piperita	control vs 653	down	19.3	-3.2	0e+00	0.0000004	CB21_SI
544	Mentha x piperita	control vs 653	down	90.7	-3.6	0e+00	0.0000000	ARP3_AR
49	Mentha x piperita	control vs 653	down	28.4	-3.8	0e+00	0.0000000	RBS2_BR
1017	Mentha x piperita	control vs 653	down	114.4	-6.1	0e+00	0.0000000	RBS2_BR
1174	Mentha x piperita	control vs 653	down	203.0	-6.3	0e+00	0.0000000	CNGC5_AR

- Rename column headers

In [530]:
```
names(Table_cv6)[4] = "Base mean"
names(Table_cv6)[5] = "Log2 fold change"
names(Table_cv6)[6] = "p-value"
names(Table_cv6)[7] = "adjusted p-value"
names(Table_cv6)[8] = "Gene name"
names(Table_cv6)[9] = "Gene ID"
names(Table_cv6)[10] = "Function"
names(Table_cv6)[11] = "Sequence"
```

In [531]:
```
Table_cv6
```

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Gen
1006	Mentha x piperita	control vs 653	up	124.3	3.5	0e+00	0.0000000	FB30_ARATH	Clu 67248.4
1411	Mentha x piperita	control vs 653	up	41.0	3.5	0e+00	0.0000001	PMTK_ARATH	Clu 67248.8
738	Mentha x piperita	control vs 653	up	121.8	3.4	0e+00	0.0000000	CO1A1_HUMAN	Clu 67248.1
764	Mentha x piperita	control vs 653	up	20.3	2.8	3e-07	0.0002652	-	Clu 67248.14
1821	Mentha x piperita	control vs 653	up	21.3	2.3	0e+00	0.0000221	-	Clu 719
1747	Mentha x piperita	control vs 653	down	19.3	-3.2	0e+00	0.0000004	CB21_SINAL	Clu 67248.9
544	Mentha x piperita	control vs 653	down	90.7	-3.6	0e+00	0.0000000	ARP3_ARATH	Clu 67248.12
49	Mentha x piperita	control vs 653	down	28.4	-3.8	0e+00	0.0000000	RBS2_BRANA	Clu 62
1017	Mentha x piperita	control vs 653	down	114.4	-6.1	0e+00	0.0000000	RBS2_BRANA	Clu 67248.
1174	Mentha x piperita	control vs 653	down	203.0	-6.3	0e+00	0.0000000	CNGC5_ARATH	Clu 67248.6

DEGs from 111 vs Control

In [532]:
```
Cv111_gn = subset(Cv111_GeneNames, Comparison == "Cv111")
```

- Order genes by fold change

In [533]:
```
table_cv1 <- (Cv111_gn[order((-Cv111_gn$log2FoldChange)),c(8,1,2,3,6,7,10)]
```

- Round digits

In [534]:
```
table_cv1$baseMean <- round(table_cv1$baseMean, 1)
table_cv1$log2FoldChange <- round(table_cv1$log2FoldChange, 1)
table_cv1$pvalue <- round(table_cv1$pvalue, 7)
table_cv1$padj <- round(table_cv1$padj, 7)
```

- Format columns

- Comparisons

In [535]:
```
table_cv1$Comparison = "control vs 111"
```

- Host

In [536]:
```
table_cv1$Host = "Mentha x piperita"
```

- Gene function

In [537]:
```
table_cv1$Function = NA
```

- Sequence

In [538]:
```
table_cv1$Sequence = NA
```

- Reorder columns

In [539]:
```
table_cv1 = table_cv1[,c(8,1,3:7,2,9:10)]
```

- Grab top 5 DEGs

In [540]:
```
Up_cv1 = head(table_cv1,n=5)
```

- Column for regulation status

```
In [541]: Up_cv1$Regulation = "up"
```

- Grab bottom 5 DEGs

```
In [542]: Down_cv1 = tail(table_cv1,n=5)
```

- Column for regulation status

```
In [543]: Down_cv1$Regulation = "down"
```

- Table for control vs 111

```
In [544]: Table_cv1 = rbind(Up_cv1,Down_cv1)
```

In [545]:
```
Table_cv1
```

	Host	Comparison	baseMean	log2FoldChange	pvalue	padj	Hit1_acc	
1005	Mentha x piperita	control vs 111	124.3	3.4	0.0e+00	0.0000000	FB30_ARATH	(6724ε
739	Mentha x piperita	control vs 111	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUMAN	(6724ε
862	Mentha x piperita	control vs 111	51.3	2.5	0.0e+00	0.0000000	IFRH_ARATH	(67248.
395	Mentha x piperita	control vs 111	18.6	2.5	1.5e-06	0.0001704	EGL1_ARATH	(67248.
447	Mentha x piperita	control vs 111	79.7	2.4	0.0e+00	0.0000000	PER45_ARATH	(67248.
1746	Mentha x piperita	control vs 111	19.3	-3.2	0.0e+00	0.0000000	CB21_SINAL	(6724ε
1046	Mentha x piperita	control vs 111	55.2	-3.5	0.0e+00	0.0000000	PNSB3_ARATH	(6724ε
50	Mentha x piperita	control vs 111	28.4	-3.8	0.0e+00	0.0000001	RBS2_BRANA	(
1016	Mentha x piperita	control vs 111	114.4	-5.9	0.0e+00	0.0000000	RBS2_BRANA	(6724
1175	Mentha x piperita	control vs 111	203.0	-6.6	0.0e+00	0.0000000	CNGC5_ARATH	(6724ε

- Rearrange columns

In [546]:
```
Table_cv1 = Table_cv1[,c(1:2,11,3:10)]
```

In [547]:

```
Table_cv1
```

	Host	Comparison	Regulation	baseMean	log2FoldChange	pvalue	padj	Hit1_
1005	Mentha x piperita	control vs 111	up	124.3	3.4	0.0e+00	0.0000000	FB30_AF
739	Mentha x piperita	control vs 111	up	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUI
862	Mentha x piperita	control vs 111	up	51.3	2.5	0.0e+00	0.0000000	IFRH_AF
395	Mentha x piperita	control vs 111	up	18.6	2.5	1.5e-06	0.0001704	EGL1_AF
447	Mentha x piperita	control vs 111	up	79.7	2.4	0.0e+00	0.0000000	PER45_AF
1746	Mentha x piperita	control vs 111	down	19.3	-3.2	0.0e+00	0.0000000	CB21_SI
1046	Mentha x piperita	control vs 111	down	55.2	-3.5	0.0e+00	0.0000000	PNSB3_AF
50	Mentha x piperita	control vs 111	down	28.4	-3.8	0.0e+00	0.0000001	RBS2_BR
1016	Mentha x piperita	control vs 111	down	114.4	-5.9	0.0e+00	0.0000000	RBS2_BR
1175	Mentha x piperita	control vs 111	down	203.0	-6.6	0.0e+00	0.0000000	CNGC5_AF

- Rename column headers

In [548]:

```
names(Table_cv1)[4] = "Base mean"
names(Table_cv1)[5] = "Log2 fold change"
names(Table_cv1)[6] = "p-value"
names(Table_cv1)[7] = "adjusted p-value"
names(Table_cv1)[8] = "Gene name"
names(Table_cv1)[9] = "Gene ID"
names(Table_cv1)[10] = "Function"
names(Table_cv1)[11] = "Sequence"
```

In [549]:

```
Table_cv1
```

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Ge
1005	Mentha x piperita	control vs 111	up	124.3	3.4	0.0e+00	0.0000000	FB30_ARATH	C 67248.
739	Mentha x piperita	control vs 111	up	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUMAN	C 67248.
862	Mentha x piperita	control vs 111	up	51.3	2.5	0.0e+00	0.0000000	IFRH_ARATH	C 67248.1
395	Mentha x piperita	control vs 111	up	18.6	2.5	1.5e-06	0.0001704	EGL1_ARATH	C 67248.1
447	Mentha x piperita	control vs 111	up	79.7	2.4	0.0e+00	0.0000000	PER45_ARATH	C 67248.1
1746	Mentha x piperita	control vs 111	down	19.3	-3.2	0.0e+00	0.0000000	CB21_SINAL	C 67248.
1046	Mentha x piperita	control vs 111	down	55.2	-3.5	0.0e+00	0.0000000	PNSB3_ARATH	C 67248.
50	Mentha x piperita	control vs 111	down	28.4	-3.8	0.0e+00	0.0000001	RBS2_BRANA	C (
1016	Mentha x piperita	control vs 111	down	114.4	-5.9	0.0e+00	0.0000000	RBS2_BRANA	C 67248
1175	Mentha x piperita	control vs 111	down	203.0	-6.6	0.0e+00	0.0000000	CNGC5_ARATH	C 67248.

DEGs from 653 vs 111

In [550]:

```
i653v111_gn = subset(i653v111_GeneNames, Comparison == "653v111")
```

- Order genes by fold change

In [551]:

```
table_6v1 <- (i653v111_gn[order((-i653v111_gn$log2FoldChange)),c(8,1,2,3,6,
```

- Round Digits

In [552]:

```
table_6v1$baseMean <- round(table_6v1$baseMean, 1)
table_6v1$log2FoldChange <- round(table_6v1$log2FoldChange, 1)
table_6v1$pvalue <- round(table_6v1$pvalue, 7)
table_6v1$padj <- round(table_6v1$padj, 7)
```

- Format columns

- By Columns

In [553]:

```
table_6v1$Comparison = "653 vs 111"
```

- Host

In [554]:

```
table_6v1$Host = "Mentha x piperita"
```

- Gene function

In [555]:

```
table_6v1$Function = NA
```

- Sequence

In [556]:

```
table_6v1$Sequence = NA
```

- Reorder columns

In [557]:

```
table_6v1 = table_6v1[,c(8,1,3:7,2,9:10)]
```

- Grab top 5 genes

In [558]:

```
Up_6v1 = head(table_6v1,n=5)
```

- Column for regulation

In [559]:
```
Up_6v1$Regulation = "up"
```

- Grab bottom 5 DEGs

In [560]:
```
Down_6v1 = tail(table_6v1,n=5)
```

- Column for regulation

In [561]:
```
Down_6v1$Regulation = "down"
```

- Table 653 vs 111

In [562]:
```
Table_6v1 = rbind(Up_6v1,Down_6v1)
```

- Rearrange columns

In [563]:
```
Table_6v1 = Table_6v1[,c(1:2,11,3:10)]
```

In [564]:
```
Table_6v1
```

	Host	Comparison	Regulation	baseMean	log2FoldChange	pvalue	padj	Hit1_acc
1412	Mentha x piperita	653 vs 111	up	41.0	3.7	0	0.00e+00	PMTK_ARATH
765	Mentha x piperita	653 vs 111	up	20.3	3.1	0	3.00e-07	-
1460	Mentha x piperita	653 vs 111	up	23.3	2.4	0	4.00e-06	P2C14_ARATH
694	Mentha x piperita	653 vs 111	up	24.2	2.3	0	1.59e-05	PSL4_ARATH
1484	Mentha x piperita	653 vs 111	up	30.3	2.0	0	3.46e-05	-
1263	Mentha x	653 vs 111	down	15.0	-2.2	0	4.92e-05	SBT16_ARATH

- Rename column headers

In [565]:
```
names(Table_6v1)[4] = "Base mean"
names(Table_6v1)[5] = "Log2 fold change"
names(Table_6v1)[6] = "p-value"
names(Table_6v1)[7] = "adjusted p-value"
names(Table_6v1)[8] = "Gene name"
names(Table_6v1)[9] = "Gene ID"
names(Table_6v1)[10] = "Function"
names(Table_6v1)[11] = "Sequence"
```

In [566]:
```
Table_6v1
```

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Gene I
1412	Mentha x piperita	653 vs 111	up	41.0	3.7	0	0.00e+00	PMTK_ARATH	Cluste 67248.8424
765	Mentha x piperita	653 vs 111	up	20.3	3.1	0	3.00e-07	-	Cluste 67248.14209
1460	Mentha x piperita	653 vs 111	up	23.3	2.4	0	4.00e-06	P2C14_ARATH	Cluste 67248.8757
694	Mentha x piperita	653 vs 111	up	24.2	2.3	0	1.59e-05	PSL4_ARATH	Cluste 67248.13295
1484	Mentha x piperita	653 vs 111	up	30.3	2.0	0	3.46e-05	-	Cluste 67248.8852
1263	Mentha x piperita	653 vs 111	down	15.0	-2.2	0	4.92e-05	SBT16_ARATH	Cluste 67248.7534
823	Mentha x piperita	653 vs 111	down	27.8	-2.4	0	2.60e-06	PIF1_XENLA	Cluste 67248.14950
1289	Mentha x piperita	653 vs 111	down	24.4	-2.5	0	2.80e-06	C3H53_ORYSJ	Cluste 67248.7685
394	Mentha x piperita	653 vs 111	down	18.6	-2.7	0	6.65e-05	EGL1_ARATH	Cluste 67248.11220
545	Mentha x piperita	653 vs 111	down	90.7	-4.0	0	0.00e+00	ARP3_ARATH	Cluste 67248.12197

Combine all tables

In [567]:
```
table_all = rbind(Table_cv6,Table_cv1,Table_6v1)
```

In [568]:

```
table_all
```

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Ge
1006	Mentha x piperita	control vs 653	up	124.3	3.5	0.0e+00	0.0000000	FB30_ARATH	C 67248.
1411	Mentha x piperita	control vs 653	up	41.0	3.5	0.0e+00	0.0000001	PMTK_ARATH	C 67248.
738	Mentha x piperita	control vs 653	up	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUMAN	C 67248.
764	Mentha x piperita	control vs 653	up	20.3	2.8	3.0e-07	0.0002652	-	C 67248.1
1821	Mentha x piperita	control vs 653	up	21.3	2.3	0.0e+00	0.0000221	-	C 7
1747	Mentha x piperita	control vs 653	down	19.3	-3.2	0.0e+00	0.0000004	CB21_SINAL	C 67248.
544	Mentha x piperita	control vs 653	down	90.7	-3.6	0.0e+00	0.0000000	ARP3_ARATH	C 67248.1
49	Mentha x piperita	control vs 653	down	28.4	-3.8	0.0e+00	0.0000000	RBS2_BRANA	C 6
1017	Mentha x piperita	control vs 653	down	114.4	-6.1	0.0e+00	0.0000000	RBS2_BRANA	C 6724
1174	Mentha x piperita	control vs 653	down	203.0	-6.3	0.0e+00	0.0000000	CNGC5_ARATH	C 67248.
1005	Mentha x piperita	control vs 111	up	124.3	3.4	0.0e+00	0.0000000	FB30_ARATH	C 67248.
739	Mentha x piperita	control vs 111	up	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUMAN	C 67248.
862	Mentha x piperita	control vs 111	up	51.3	2.5	0.0e+00	0.0000000	IFRH_ARATH	C 67248.1
395	Mentha x piperita	control vs 111	up	18.6	2.5	1.5e-06	0.0001704	EGL1_ARATH	C 67248.1
447	Mentha x piperita	control vs 111	up	79.7	2.4	0.0e+00	0.0000000	PER45_ARATH	C 67248.1
1746	Mentha x piperita	control vs 111	down	19.3	-3.2	0.0e+00	0.0000000	CB21_SINAL	C 67248.

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Ge
1046	Mentha x piperita	control vs 111	down	55.2	-3.5	0.0e+00	0.0000000	PNSB3_ARATH	C 67248.
50	Mentha x piperita	control vs 111	down	28.4	-3.8	0.0e+00	0.0000001	RBS2_BRANA	C 6
1016	Mentha x piperita	control vs 111	down	114.4	-5.9	0.0e+00	0.0000000	RBS2_BRANA	C 6724
1175	Mentha x piperita	control vs 111	down	203.0	-6.6	0.0e+00	0.0000000	CNGC5_ARATH	C 67248.
1412	Mentha x piperita	653 vs 111	up	41.0	3.7	0.0e+00	0.0000000	PMTK_ARATH	C 67248.
765	Mentha x piperita	653 vs 111	up	20.3	3.1	0.0e+00	0.0000003	-	C 67248.1
1460	Mentha x piperita	653 vs 111	up	23.3	2.4	0.0e+00	0.0000040	P2C14_ARATH	C 67248.
694	Mentha x piperita	653 vs 111	up	24.2	2.3	0.0e+00	0.0000159	PSL4_ARATH	C 67248.1
1484	Mentha x piperita	653 vs 111	up	30.3	2.0	0.0e+00	0.0000346	-	C 67248.
1263	Mentha x piperita	653 vs 111	down	15.0	-2.2	0.0e+00	0.0000492	SBT16_ARATH	C 67248.
823	Mentha x piperita	653 vs 111	down	27.8	-2.4	0.0e+00	0.0000026	PIF1_XENLA	C 67248.1
1289	Mentha x piperita	653 vs 111	down	24.4	-2.5	0.0e+00	0.0000028	C3H53_ORYSJ	C 67248.
394	Mentha x piperita	653 vs 111	down	18.6	-2.7	0.0e+00	0.0000665	EGL1_ARATH	C 67248.1
545	Mentha x piperita	653 vs 111	down	90.7	-4.0	0.0e+00	0.0000000	ARP3_ARATH	C 67248.1

Insert sequences

- First, subset the target sequences from the fasta file

```
In [569]: seqs = fastafile[names(fastafile) %in% table_all[,9]]
```

- Sanity checks

```
In [570]: length(seqs)
```

```
20
```

```
In [571]: length(seqs) == length(table_all[,9])
```

```
FALSE
```

```
In [572]: sum(abs(length(seqs) - length(table_all[,9])))
```

```
10
```

- There are 10 duplicates

```
In [573]: sum(duplicated(table_all[,9]))
```

```
10
```

- Extract gene IDs

```
In [574]: names(seqs)
```

```
'Cluster-67248.98511'   'Cluster-67248.87571'   'Cluster-67248.4354'   'Cluster-67248.65881'
'Cluster-67248.112206'   'Cluster-67248.149503'   'Cluster-67248.84245'   'Cluster-67248.50623'
'Cluster-67248.75344'   'Cluster-71973.0'   'Cluster-67248.155958'   'Cluster-67248.76854'
'Cluster-67248.41609'   'Cluster-67248.132953'   'Cluster-67248.13909'   'Cluster-6227.0'
'Cluster-67248.115536'   'Cluster-67248.142094'   'Cluster-67248.121974'
'Cluster-67248.88523'
```

- Match names of gene IDs from fasta file with column in table

In [575]:

```
# For every row in the "Gene ID" column
for (i in seq(table_all[,9])){

# For every entry in the list of sequences
    for (j in seq(names(seqs))){

        # If the row matches the sequence
        if (table_all[i,9] == names(seqs[j])){

            # Grab the gene ID
            y = names(seqs[j])

            # Grab the sequence
            x = (seqs[j]) # unlist

            #
            # x = (str_split(x, "''"))

            # Add gene ID to table
            table_all[i,12] = y

            # Add sequence to table
            table_all[i,11] = x
     }

 }

 }
```

- Table

```
In [576]: table_all
```

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Ge
1006	Mentha x piperita	control vs 653	up	124.3	3.5	0.0e+00	0.0000000	FB30_ARATH	C 67248.
1411	Mentha x piperita	control vs 653	up	41.0	3.5	0.0e+00	0.0000001	PMTK_ARATH	C 67248.
738	Mentha x piperita	control vs 653	up	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUMAN	C 67248.
764	Mentha x piperita	control vs 653	up	20.3	2.8	3.0e-07	0.0002652	-	C 67248.1
1821	Mentha x piperita	control vs 653	up	21.3	2.3	0.0e+00	0.0000221	-	C 7
1747	Mentha x piperita	control vs 653	down	19.3	-3.2	0.0e+00	0.0000004	CB21_SINAL	C 67248.
544	Mentha x piperita	control vs 653	down	90.7	-3.6	0.0e+00	0.0000000	ARP3_ARATH	C 67248.1
49	Mentha x piperita	control vs 653	down	28.4	-3.8	0.0e+00	0.0000000	RBS2_BRANA	C 6
1017	Mentha x piperita	control vs 653	down	114.4	-6.1	0.0e+00	0.0000000	RBS2_BRANA	C 6724
1174	Mentha x piperita	control vs 653	down	203.0	-6.3	0.0e+00	0.0000000	CNGC5_ARATH	C 67248.
1005	Mentha x piperita	control vs 111	up	124.3	3.4	0.0e+00	0.0000000	FB30_ARATH	C 67248.
739	Mentha x piperita	control vs 111	up	121.8	3.4	0.0e+00	0.0000000	CO1A1_HUMAN	C 67248.
862	Mentha x piperita	control vs 111	up	51.3	2.5	0.0e+00	0.0000000	IFRH_ARATH	C 67248.1
395	Mentha x piperita	control vs 111	up	18.6	2.5	1.5e-06	0.0001704	EGL1_ARATH	C 67248.1
447	Mentha x piperita	control vs 111	up	79.7	2.4	0.0e+00	0.0000000	PER45_ARATH	C 67248.1
1746	Mentha x piperita	control vs 111	down	19.3	-3.2	0.0e+00	0.0000000	CB21_SINAL	C 67248.

	Host	Comparison	Regulation	Base mean	Log2 fold change	p-value	adjusted p-value	Gene name	Ge
1046	Mentha x piperita	control vs 111	down	55.2	-3.5	0.0e+00	0.0000000	PNSB3_ARATH	C 67248.
50	Mentha x piperita	control vs 111	down	28.4	-3.8	0.0e+00	0.0000001	RBS2_BRANA	C 6
1016	Mentha x piperita	control vs 111	down	114.4	-5.9	0.0e+00	0.0000000	RBS2_BRANA	C 67248
1175	Mentha x piperita	control vs 111	down	203.0	-6.6	0.0e+00	0.0000000	CNGC5_ARATH	C 67248.
1412	Mentha x piperita	653 vs 111	up	41.0	3.7	0.0e+00	0.0000000	PMTK_ARATH	C 67248.
765	Mentha x piperita	653 vs 111	up	20.3	3.1	0.0e+00	0.0000003	-	C 67248.1
1460	Mentha x piperita	653 vs 111	up	23.3	2.4	0.0e+00	0.0000040	P2C14_ARATH	C 67248.
694	Mentha x piperita	653 vs 111	up	24.2	2.3	0.0e+00	0.0000159	PSL4_ARATH	C 67248.1
1484	Mentha x piperita	653 vs 111	up	30.3	2.0	0.0e+00	0.0000346	-	C 67248.
1263	Mentha x piperita	653 vs 111	down	15.0	-2.2	0.0e+00	0.0000492	SBT16_ARATH	C 67248.
823	Mentha x piperita	653 vs 111	down	27.8	-2.4	0.0e+00	0.0000026	PIF1_XENLA	C 67248.1
1289	Mentha x piperita	653 vs 111	down	24.4	-2.5	0.0e+00	0.0000028	C3H53_ORYSJ	C 67248.
394	Mentha x piperita	653 vs 111	down	18.6	-2.7	0.0e+00	0.0000665	EGL1_ARATH	C 67248.1
545	Mentha x piperita	653 vs 111	down	90.7	-4.0	0.0e+00	0.0000000	ARP3_ARATH	C 67248.1

- Sanity checks

> - Do the names match?

In [577]:
```
table_all[,9] == table_all[,12]
```

```
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

- Do the sequence lengths match?

- Reorder sequence names first

In [578]:
```
ordered_seqs = seqs[order(match(names(seqs), table_all[,9]))]
```

In [579]:
```
nchar(ordered_seqs) == nchar(unique(table_all[,11]))
```

```
 Cluster-67248.41609   TRUE
 Cluster-67248.84245   TRUE
 Cluster-67248.13909   TRUE
Cluster-67248.142094   TRUE
       Cluster-71973.0   TRUE
 Cluster-67248.98511   TRUE
Cluster-67248.121974   TRUE
        Cluster-6227.0   TRUE
  Cluster-67248.4354   TRUE
 Cluster-67248.65881   TRUE
Cluster-67248.155958   TRUE
Cluster-67248.112206   TRUE
Cluster-67248.115536   TRUE
 Cluster-67248.50623   TRUE
 Cluster-67248.87571   TRUE
Cluster-67248.132953   TRUE
 Cluster-67248.88523   TRUE
 Cluster-67248.75344   TRUE
Cluster-67248.149503   TRUE
 Cluster-67248.76854   TRUE
```

- Export file

In [580]:
```
write.csv(table_all[,c(1:5,8,11)], file = "Mentha_table.csv", row.names=FAL
```

Add column for Gene Ontology

- Grab only the GO data for the gene IDs of interest

> - Subset method 1

In [581]:
```
GO_df = subset(GO, as.character(GO$Gene.ID) %in% table_all[,9])
```

> - Subset method 2

In [582]:
```
GO_DF = setDT(GO)[as.character(GO$Gene.ID) %chin% table_all[,9]]
```

- Sanity check- are the data the same?

In [583]:
```
GO_df == GO_DF
```

Gene.ID	Gene.Ontology.Biological.Pathway	BP.Description	Gene.Ontology.Molecular.Function	MF.Des
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	TRUE	

- Which DEGs are in GO?

In [584]:
```
unique(table_all[,9][table_all[,9] %in% as.character(GO_df$Gene.ID)])
```

```
'Cluster-67248.84245'   'Cluster-67248.13909'   'Cluster-67248.142094'   'Cluster-67248.121974'
'Cluster-67248.65881'   'Cluster-67248.155958'   'Cluster-67248.112206'
'Cluster-67248.115536'   'Cluster-67248.50623'   'Cluster-67248.87571'   'Cluster-67248.132953'
'Cluster-67248.149503'   'Cluster-67248.76854'
```

- Which DEGs are not in GO?

In [585]:
```
unique(table_all[,9][!(table_all[,9] %in% GO_df$Gene.ID)])
```

```
'Cluster-67248.41609'   'Cluster-71973.0'   'Cluster-67248.98511'   'Cluster-6227.0'
'Cluster-67248.4354'   'Cluster-67248.88523'   'Cluster-67248.75344'
```

In [586]:
```
length(unique(table_all[,9][table_all[,9] %in% as.character(GO_df$Gene.ID)]
```

```
13
```

In [587]:
```
length(unique(table_all[,9][!(table_all[,9] %in% GO_df$Gene.ID)]))
```

```
7
```

- How many total DEGs?

In [588]:
```
length(unique(table_all[,9][table_all[,9] %in% as.character(GO_df$Gene.ID)]
```

```
20
```

- Add GO to dataframe

In [589]:

```
# For every row in the "Gene ID" column
for (i in seq(table_all[,9])){

# For every entry in the list of sequences
    for (j in seq(as.character(GO_DF$Gene.ID))){

        # If the row matches the sequence
        if (table_all[i,9] == (as.character(GO_DF$Gene.ID[j]))){

            # Grab the gene ID
            y = (as.character(GO_DF$Gene.ID[j]))

            # Grab the molecular function column
            x = (as.character(GO_DF$MF.Description[j])) # unlist

            # Add gene ID to table
            table_all[i,12] = y

            # Add molecular function to table
            table_all[i,10] = x
    }
 }

 }
```

- Sanity checks

> - Do the names match?

In [590]:

```
sum((table_all[,9] == table_all[,12]),na.rm=T)
```

```
30
```

In [591]:

```
table_all[,9] == table_all[,12]
```

```
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

- Add column for source of gene ontology information

In [592]:

```
table_all$Source = ifelse(table_all$Function == 'NA', 'NA', 'GO')
```

- Remove excessive columns

In [593]:
```
names(table_all)
```

```
'Host'   'Comparison'   'Regulation'   'Base mean'   'Log2 fold change'   'p-value'
'adjusted p-value'   'Gene name'   'Gene ID'   'Function'   'Sequence'   'V12'   'Source'
```

In [594]:
```
table_all = table_all[,c(1:5,8:10,13,11)]
```

- Add functional data from other sources

In [621]:
```
table_all$Function = ifelse(table_all[,7] == 'Cluster-67248.41609',
                            'F-box and associated interaction domains-conta
                            ,table_all$Function)
table_all$Source = ifelse(table_all[,7] == 'Cluster-67248.41609',
                          'Blast_NR', table_all$Source)
```

In [623]:
```
table_all$Function = ifelse(table_all[,7] == 'Cluster-71973.0',
                            'PREDICTED: Solanum tuberosum uncharacterized I
                            table_all$Function)
table_all$Source = ifelse(table_all[,7] == 'Cluster-71973.0',
                          'Blast_NT', table_all$Source)
```

In [626]:
```
table_all$Function = ifelse(table_all[,7] == 'Cluster-67248.98511',
                            'light-harvesting complex II chlorophyll a/b bi
                            table_all$Function)
table_all$Source = ifelse(table_all[,7] == 'Cluster-67248.98511',
                          'Blast_NT, KO', table_all$Source)
```

In [628]:
```
table_all$Function = ifelse(table_all[,7] == 'Cluster-6227.0',
                            'Carbon fixation in photosynthetic organisms: r
                            table_all$Function)
table_all$Source = ifelse(table_all[,7] == 'Cluster-6227.0',
                          'Blast_NT, KO', table_all$Source)
```

In [630]:
```
table_all$Function = ifelse(table_all[,7] == 'Cluster-67248.4354',
                            'Carbon metabolism: ribulose-bisphosphate carbo
                            table_all$Function)
table_all$Source = ifelse(table_all[,7] == 'Cluster-67248.4354',
                          'Blast_NT, KO', table_all$Source)
```

In [632]:
```
table_all$Function = ifelse(table_all[,7] == 'Cluster-67248.75344',
                            'hypothetical protein PHAVU_011G034700g [Phasec
                            table_all$Function)
table_all$Source = ifelse(table_all[,7] == 'Cluster-67248.75344',
                          '', table_all$Source)
```

```
In [633]: table_all
```

	Host	Comparison	Regulation	Base mean	Log2 fold change	Gene name	Gene ID	
1006	Mentha x piperita	control vs 653	up	124.3	3.5	FB30_ARATH	Cluster-67248.41609	F-bo domains cacao]>gi|5905 F-bo domains cacao]>gi|50 box and ass containin cacao]>gi|50 box and ass containin
1411	Mentha x piperita	control vs 653	up	41.0	3.5	PMTK_ARATH	Cluster-67248.84245	
738	Mentha x piperita	control vs 653	up	121.8	3.4	CO1A1_HUMAN	Cluster-67248.13909	protein
764	Mentha x piperita	control vs 653	up	20.3	2.8	-	Cluster-67248.142094	
1821	Mentha x piperita	control vs 653	up	21.3	2.3	-	Cluster-71973.0	PRED unch
1747	Mentha x piperita	control vs 653	down	19.3	-3.2	CB21_SINAL	Cluster-67248.98511	light-harvestir
544	Mentha x piperita	control vs 653	down	90.7	-3.6	ARP3_ARATH	Cluster-67248.121974	
49	Mentha x piperita	control vs 653	down	28.4	-3.8	RBS2_BRANA	Cluster-6227.0	Carb organi
1017	Mentha x piperita	control vs 653	down	114.4	-6.1	RBS2_BRANA	Cluster-67248.4354	C bisphosph
1174	Mentha x piperita	control vs 653	down	203.0	-6.3	CNGC5_ARATH	Cluster-67248.65881	ion char

	Host	Comparison	Regulation	Base mean	Log2 fold change	Gene name	Gene ID	
1005	Mentha x piperita	control vs 111	up	124.3	3.4	FB30_ARATH	Cluster-67248.41609	F-bo domains cacao]>gi|5905 F-bo domains cacao]>gi|50 box and ass containin cacao]>gi|50 box and ass containin
739	Mentha x piperita	control vs 111	up	121.8	3.4	CO1A1_HUMAN	Cluster-67248.13909	protein
862	Mentha x piperita	control vs 111	up	51.3	2.5	IFRH_ARATH	Cluster-67248.155958	oxid
395	Mentha x piperita	control vs 111	up	18.6	2.5	EGL1_ARATH	Cluster-67248.112206	
447	Mentha x piperita	control vs 111	up	79.7	2.4	PER45_ARATH	Cluster-67248.115536	heme
1746	Mentha x piperita	control vs 111	down	19.3	-3.2	CB21_SINAL	Cluster-67248.98511	light-harvesti
1046	Mentha x piperita	control vs 111	down	55.2	-3.5	PNSB3_ARATH	Cluster-67248.50623	iron-sulfur clu
50	Mentha x piperita	control vs 111	down	28.4	-3.8	RBS2_BRANA	Cluster-6227.0	Carb organi
1016	Mentha x piperita	control vs 111	down	114.4	-5.9	RBS2_BRANA	Cluster-67248.4354	C bisphosp
1175	Mentha x piperita	control vs 111	down	203.0	-6.6	CNGC5_ARATH	Cluster-67248.65881	ion cha
1412	Mentha x piperita	653 vs 111	up	41.0	3.7	PMTK_ARATH	Cluster-67248.84245	
765	Mentha x piperita	653 vs 111	up	20.3	3.1	-	Cluster-67248.142094	
1460	Mentha x piperita	653 vs 111	up	23.3	2.4	P2C14_ARATH	Cluster-67248.87571	

	Host	Comparison	Regulation	Base mean	Log2 fold change	Gene name	Gene ID	
694	Mentha x piperita	653 vs 111	up	24.2	2.3	PSL4_ARATH	Cluster-67248.132953	hydr anhydride anhydrid binding//nu
1484	Mentha x piperita	653 vs 111	up	30.3	2.0	-	Cluster-67248.88523	
1263	Mentha x piperita	653 vs 111	down	15.0	-2.2	SBT16_ARATH	Cluster-67248.75344	PHAVU vulgaris]>gi|5 PHAVU vulgaris]
823	Mentha x piperita	653 vs 111	down	27.8	-2.4	PIF1_XENLA	Cluster-67248.149503	nucleotide ph fre binding//n acti bin binding/
1289	Mentha x piperita	653 vs 111	down	24.4	-2.5	C3H53_ORYSJ	Cluster-67248.76854	binding//p
394	Mentha x piperita	653 vs 111	down	18.6	-2.7	EGL1_ARATH	Cluster-67248.112206	
545	Mentha x piperita	653 vs 111	down	90.7	-4.0	ARP3_ARATH	Cluster-67248.121974	

- Write file

In [636]:

```
setwd('/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/DATA/R_FIL
```

In [637]:

```
write.csv(table_all, file = "Mentha_table.csv", row.names=FALSE)
```

Exploratory data analyses

Back to Table of Contents

- **Transform data for pattern recognition**
- **Raw, untransformed data are used for inference downstream**

- **Filter out counts <1 to reduce dataset dimensions & expedite analysis**

In [35]:

```
dds <- dds[rowSums(counts(dds)) > 1, ]
nrow(dds)
```

246300

In [36]:

```
dds$group <- relevel(dds$group, ref = "Control")
```

- **Stabilize variance (since it is related to the mean) with variance stabilizing transformation (VST)**

In [37]:

```
vsd <- vst(dds, blind = FALSE)
```

```
-- note: fitType='parametric', but the dispersion trend was not well capt
ured by the
   function: y = a/x + b, and a local regression fit was automatically su
bstituted.
   specify fitType='local' or 'mean' to avoid this message next time.
```

- **Stabilize variance with regularized-logarithm transformation (rlog)**

In [38]:

```
rld <- rlog(dds, blind = FALSE)
```

- **Inspect the transformed data**

> VST

In [39]:

```
head(assay(vsd), 3)
```

	S2_3_2_1	S2_3_2_9	S2_3_2_4	S2_1_2_1	S2_1_2_4	S2_1_2_5	S2_2_2_2	S2_2_2_5	S
Cluster-67248.142691	5.635508	5.433058	5.433058	5.561595	5.535468	5.520594	5.433058	5.689746	5
Cluster-67248.107952	7.265288	6.946230	7.119010	7.095979	7.161691	7.474105	7.370777	7.078834	7
Cluster-58782.0	5.433058	5.506769	5.433058	5.433058	5.433058	5.433058	5.433058	5.433058	5

> rlog

In [40]:
```
head(assay(rld), 3)
```

	S2_3_2_1	S2_3_2_9	S2_3_2_4	S2_1_2_1	S2_1_2_4	S2_1_2_5	S2_2_2_2	S2
Cluster-67248.142691	1.771567	1.6816257	1.682194	1.7318188	1.7194922	1.7120304	1.6777742	1.8
Cluster-67248.107952	7.039000	6.7659611	6.915980	6.8921934	6.9509049	7.2132424	7.1280055	6.8
Cluster-58782.0	-0.914685	-0.8935841	-0.912875	-0.9164012	-0.9160391	-0.9162232	-0.9162492	-0.9

- **Visualize effect of transformation on data**

In [41]:

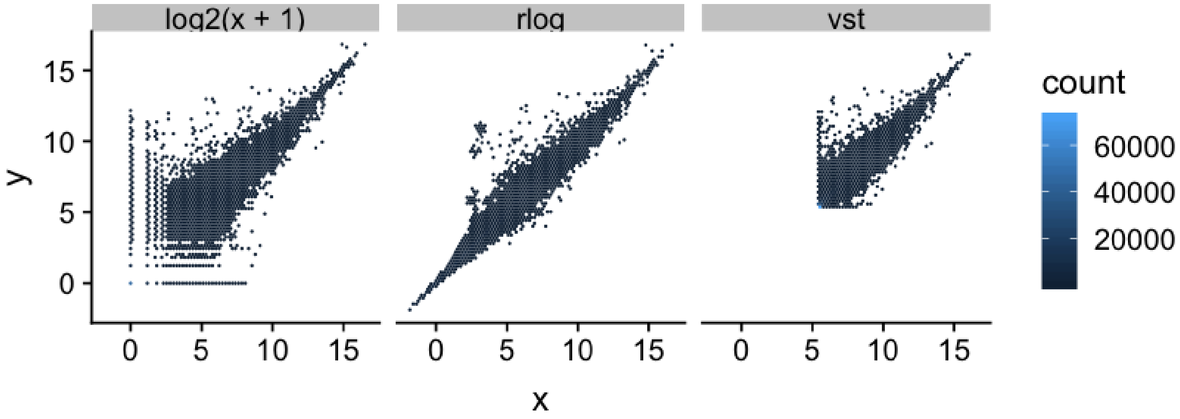
```
dds <- estimateSizeFactors(dds)

df <- bind_rows(
  as_data_frame(log2(counts(dds, normalized=TRUE)[, 1:2]+1)) %>%
         mutate(transformation = "log2(x + 1)"),
  as_data_frame(assay(vsd)[, 1:2]) %>% mutate(transformation = "vst"),
  as_data_frame(assay(rld)[, 1:2]) %>% mutate(transformation = "rlog"))

colnames(df)[1:2] <- c("x", "y")

ggplot(df, aes(x = x, y = y)) + geom_hex(bins = 80) +
  coord_fixed() + facet_grid( . ~ transformation)
```

```
Warning message:
“`as_data_frame()` is deprecated, use `as_tibble()` (but mind the new sem
antics).
This warning is displayed once per session.”
```



- **Compute Euclidean sample distances**

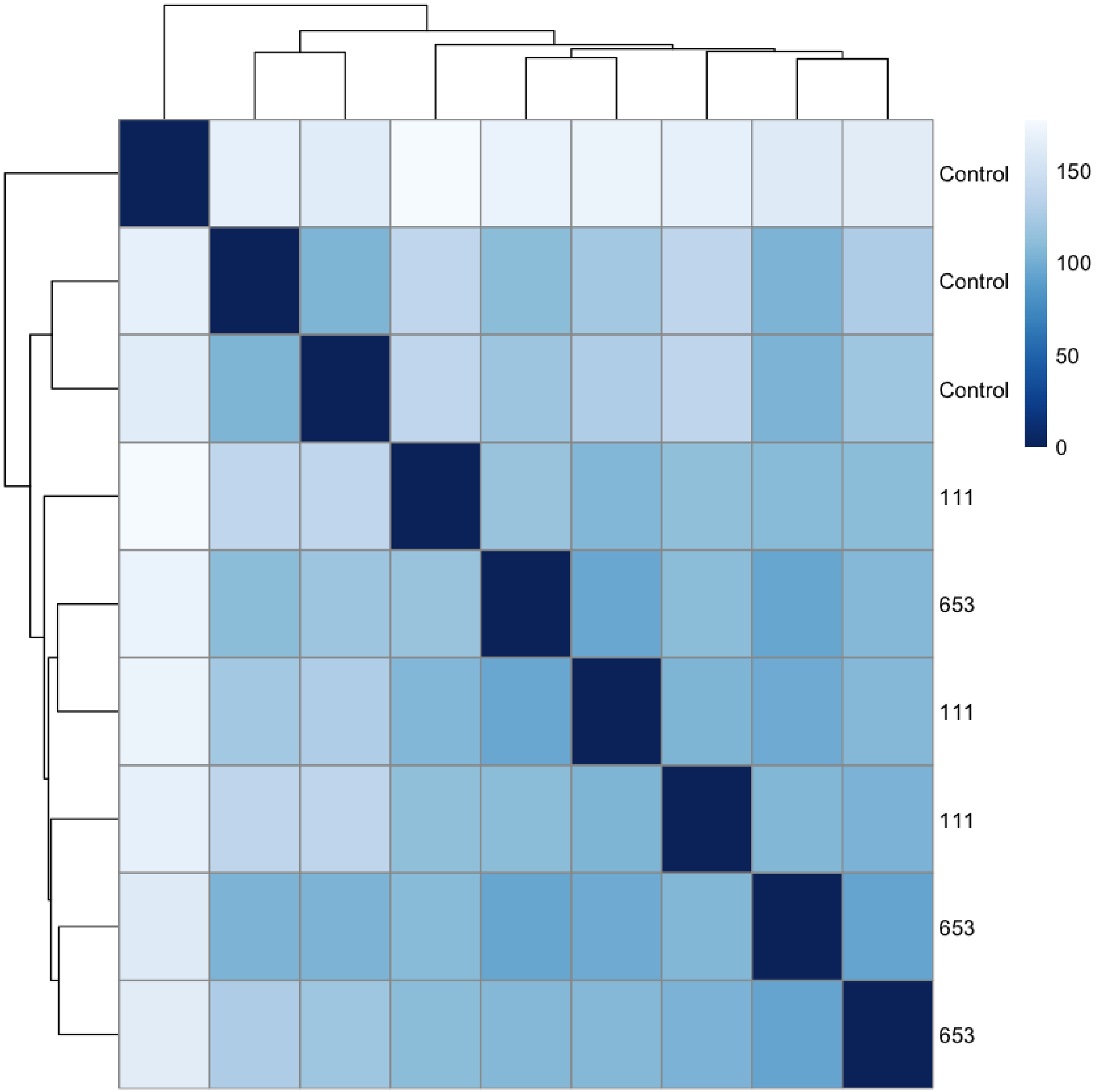
  - Transpose dataset & coerce into matrix

```
In [42]: sampleDists <- dist(t(assay(vsd)))
         sampleDists
```

```
          S2_3_2_1  S2_3_2_9  S2_3_2_4  S2_1_2_1  S2_1_2_4  S2_1_2_5  S2_
2_2_2
S2_3_2_9 167.28496
S2_3_2_4 104.57038 162.75034
S2_1_2_1 104.13654 160.54175 104.33672
S2_1_2_4 126.75408 163.34956 120.17469  93.91661
S2_1_2_5 110.18749 169.81285 118.84469  94.13099 107.49950
S2_2_2_2 136.08553 167.08227 136.59572 105.99619 103.36629 109.92072
S2_2_2_5 121.31476 171.43671 128.65036  97.71853 107.58663  95.84291 105.
08904
S2_2_2_9 138.30702 177.54195 138.01094 108.78442 110.23610 116.27587 112.
14574
          S2_2_2_5
S2_3_2_9
S2_3_2_4
S2_1_2_1
S2_1_2_4
S2_1_2_5
S2_2_2_2
S2_2_2_5
S2_2_2_9 107.05215
```

- **Visualize sample distances with heatmap using Euclidean distances**

```
In [43]: sampleDistMatrix <- as.matrix( sampleDists )
         rownames(sampleDistMatrix) <- paste( vsd$group)
         colnames(sampleDistMatrix) <- NULL
         colors <- colorRampPalette(rev(brewer.pal(9, "Blues")))(255)
         pheatmap(sampleDistMatrix,
                  clustering_distance_rows = sampleDists,
                  clustering_distance_cols = sampleDists,
                  cluster_rows=T, cluster_cols=T,
                  col = colors)
```



- **Compute Poisson distances (Witten 2011)**

- Distances

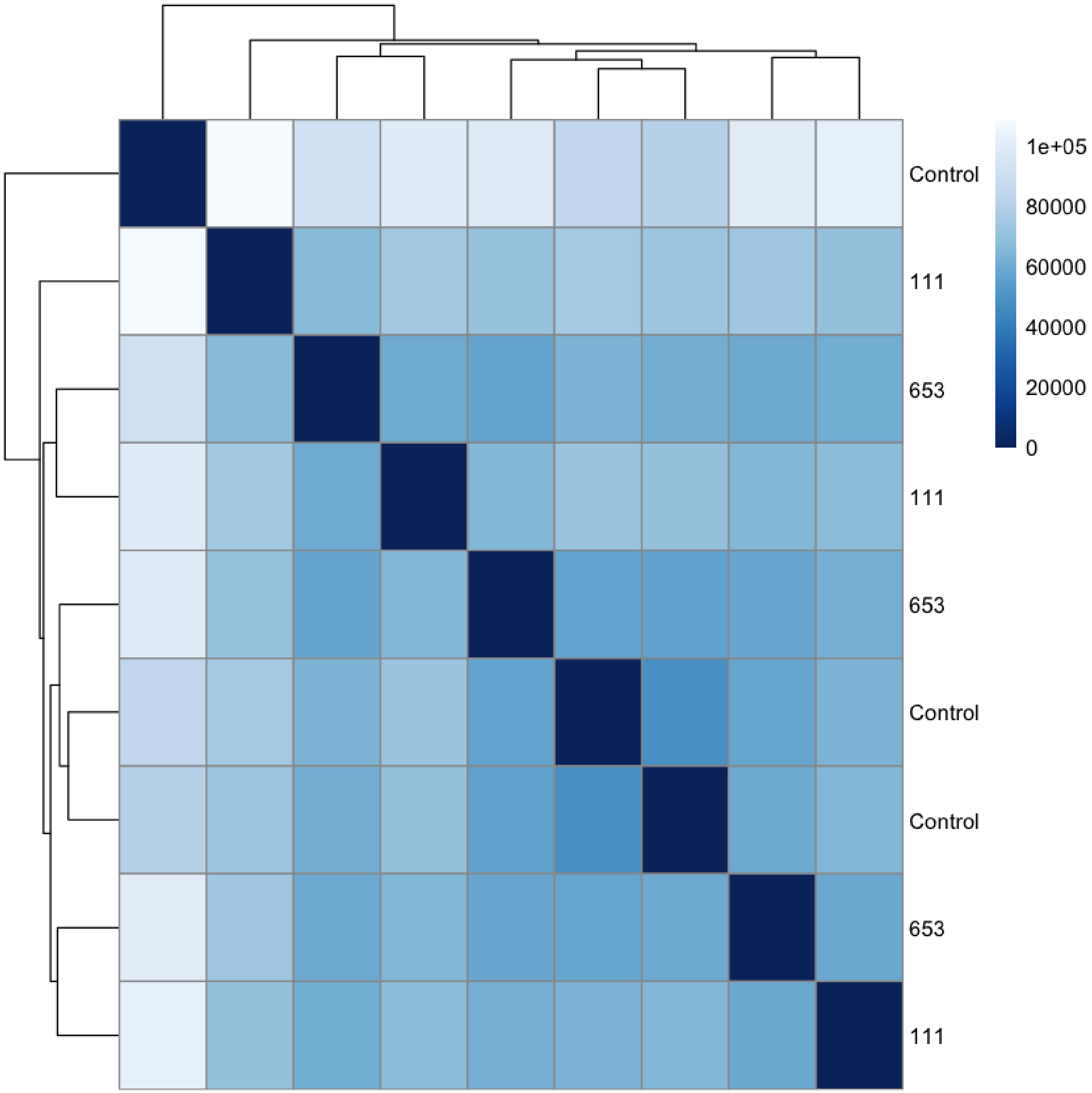
In [44]:

```
poisd <- PoissonDistance(t(counts(dds)))
```

- **Visualize sample distances with heatmap using Poisson distances**

In [45]:

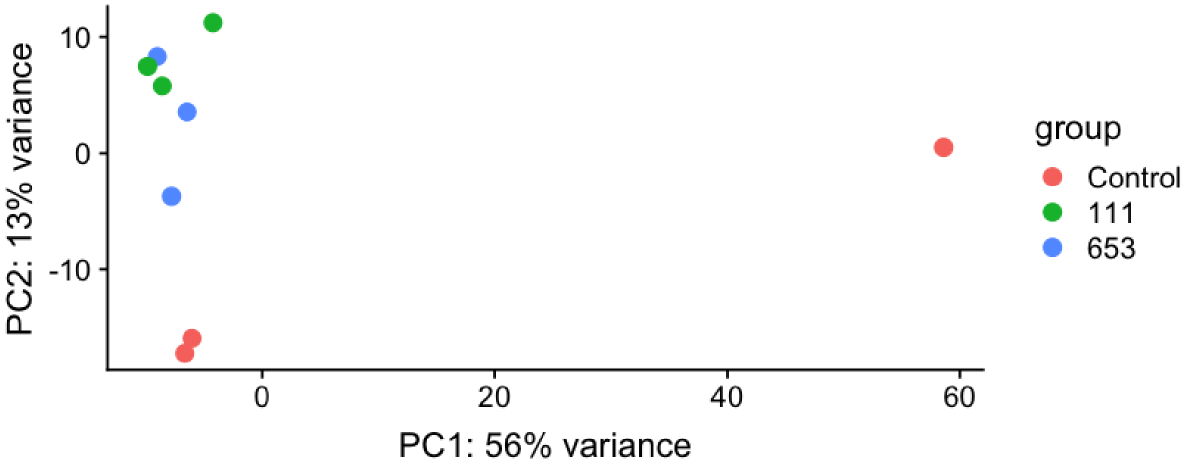
```
samplePoisDistMatrix <- as.matrix( poisd$dd )
rownames(samplePoisDistMatrix) <- paste( dds$group, sep=" - " )
colnames(samplePoisDistMatrix) <- NULL
pheatmap(samplePoisDistMatrix,
         clustering_distance_rows = poisd$dd,
         clustering_distance_cols = poisd$dd,
         col = colors)
```



- **Visualize sample-to-sample differences with pricipal components analysis (PCA)**

- PCA with DESeq2

In [46]:
```
plotPCA(vsd, intgroup = c("group"))
```



- PCA with ggplot2

- View PCs

In [47]:
```
pcaData <- plotPCA(vsd, intgroup = c( "group"), returnData = TRUE)
pcaData
```

	PC1	PC2	group	group.1	name
S2_3_2_1	-6.646067	-17.2333967	Control	Control	S2_3_2_1
S2_3_2_9	58.586935	0.4998174	Control	Control	S2_3_2_9
S2_3_2_4	-6.029665	-15.9366397	Control	Control	S2_3_2_4
S2_1_2_1	-7.783224	-3.7150135	653	653	S2_1_2_1
S2_1_2_4	-6.452647	3.5560928	653	653	S2_1_2_4
S2_1_2_5	-9.011399	8.3241647	653	653	S2_1_2_5
S2_2_2_2	-4.222302	11.2345655	111	111	S2_2_2_2
S2_2_2_5	-8.588481	5.7921773	111	111	S2_2_2_5
S2_2_2_9	-9.853150	7.4782322	111	111	S2_2_2_9

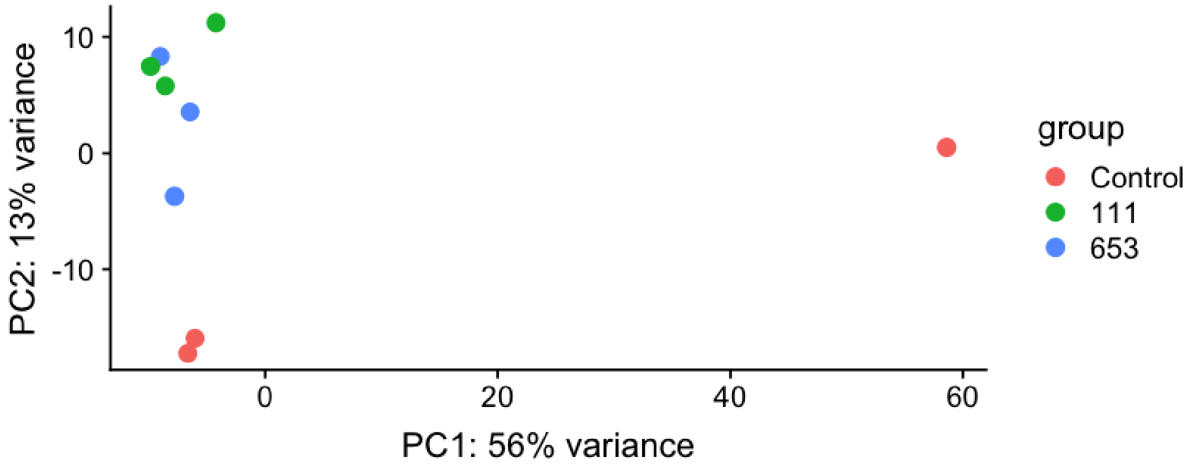
- Express variation explained to percentage

In [48]:
```
percentVar <- round(100 * attr(pcaData, "percentVar"))
```

- Plot data

In [49]:

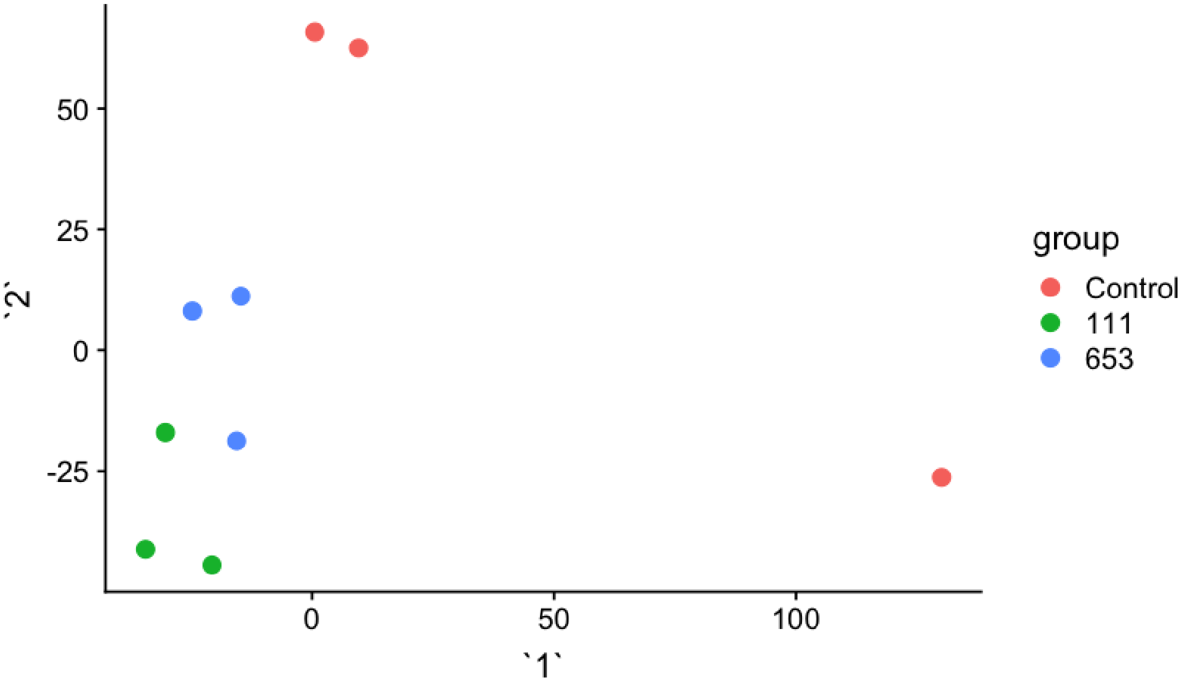
```
ggplot(pcaData, aes(x = PC1, y = PC2, color = group)) +
  geom_point(size =3) +
  xlab(paste0("PC1: ", percentVar[1], "% variance")) +
  ylab(paste0("PC2: ", percentVar[2], "% variance")) +
  coord_fixed()
```



- **Visualize sample-to-sample differences with multidimensional scaling (MDS)**

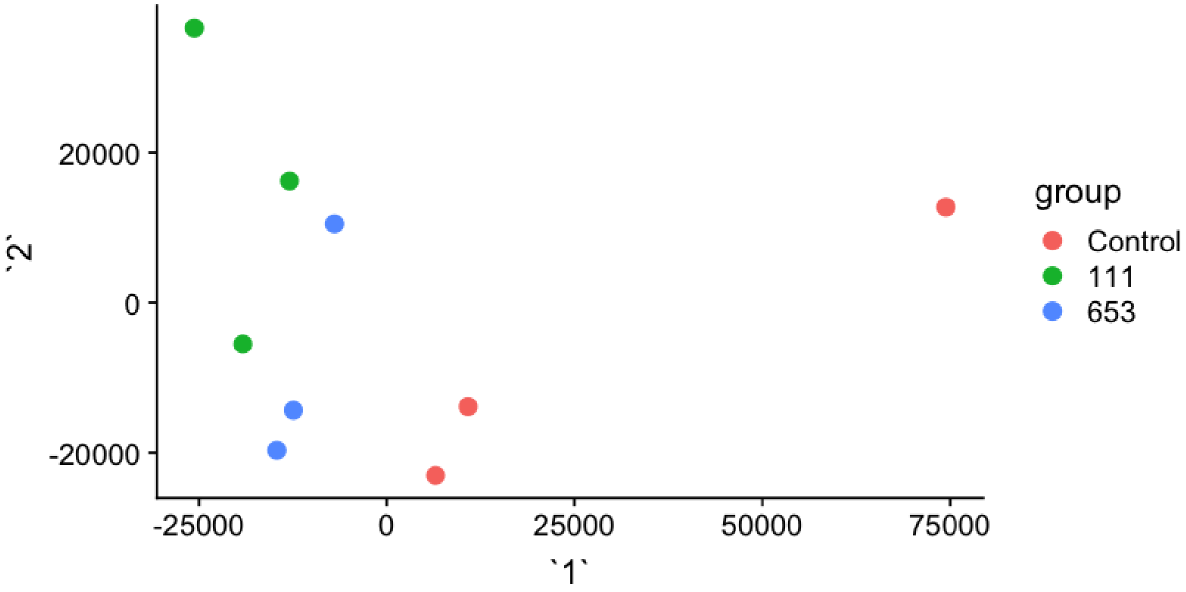
- VSD data

```
In [50]: mds <- as.data.frame(colData(vsd))  %>%
                  cbind(cmdscale(sampleDistMatrix))
         ggplot(mds, aes(x = `1`, y = `2`, color = group)) +
           geom_point(size = 3) + coord_fixed()
```



- Poisson data

In [51]:
```
mdsPois <- as.data.frame(colData(dds)) %>%
   cbind(cmdscale(samplePoisDistMatrix))
ggplot(mdsPois, aes(x = `1`, y = `2`, color = group)) +
  geom_point(size = 3) + coord_fixed()
```



Diagnostics

Back to Table of Contents

In [ ]:

Parametric analysis: differential expression analysis

Back to Table of Contents

- **Identify differentially expressed genes with raw count data**

In [52]:
```
dds <- DESeq(dds)
```

```
using pre-existing size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

- **Tabulate results, set $\alpha = 0.05$**
- **Adjust p-values with Benjamini & Hochberg (1995) to account for false discoveries**
- **Shrink/deflate effect sizes (Log fold change estimates)**

- Contrast control vs 653

In [53]:
```
Cv653 = results(dds, contrast=c("group","653","Control"),
                independentFiltering=TRUE, alpha=0.001, pAdjustMethod="BH",
Cv653 = lfcShrink(dds, contrast=c("group","653","Control"), res=Cv653)
```

```
using 'normal' for LFC shrinkage, the Normal prior from Love et al (201
4).
additional priors are available via the 'type' argument, see ?lfcShrink f
or details
```

- Summary

In [54]:
```
summary(Cv653)
```

```

out of 246300 with nonzero total read count
adjusted p-value < 0.001
LFC > 0 (up)       : 73, 0.03%
LFC < 0 (down)     : 111, 0.045%
outliers [1]       : 2543, 1%
low counts [2]     : 138421, 56%
(mean count < 6)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

- Contrast control vs 111

In [55]:
```
Cv111 = results(dds, contrast=c("group","111","Control"),
                independentFiltering=TRUE, alpha=0.001, pAdjustMethod="BH",
Cv111 = lfcShrink(dds, contrast=c("group","111","Control"), res=Cv111)
```

```
using 'normal' for LFC shrinkage, the Normal prior from Love et al (201
4).
additional priors are available via the 'type' argument, see ?lfcShrink f
or details
```

- Summary

In [56]:
```
summary(Cv111)
```

```
out of 246300 with nonzero total read count
adjusted p-value < 0.001
LFC > 0 (up)       : 378, 0.15%
LFC < 0 (down)     : 1173, 0.48%
outliers [1]       : 2543, 1%
low counts [2]     : 133688, 54%
(mean count < 5)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

- Contrast 653 vs 111

In [57]:
```
i653v111 = results(dds, contrast=c("group","653","111"),
                independentFiltering=TRUE, alpha=0.001, pAdjustMethod="BH",
i653v111 = lfcShrink(dds, contrast=c("group","653","111"), res=i653v111)
```

```
using 'normal' for LFC shrinkage, the Normal prior from Love et al (201
4).
additional priors are available via the 'type' argument, see ?lfcShrink f
or details
```

- Summary

In [58]:

```
summary(i653v111)
```

```
out of 246300 with nonzero total read count
adjusted p-value < 0.001
LFC > 0 (up)       : 84, 0.034%
LFC < 0 (down)     : 14, 0.0057%
outliers [1]       : 2543, 1%
low counts [2]     : 152475, 62%
(mean count < 8)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

- **Subset gene with > or < log 2 fold change and q value < 0.05 (p value post FDR adjustment)**

> - First define the cutoffs for log2 fold differences and the q value

In [59]:

```
log2cutoff = 2
qvaluecutoff = 0.001
```

> - Concatenate results that are differentially expressed (>log2) and adjusted p-values < $q = 0.001$

In [60]:

```
diffXGenes <- unique(c(
  rownames(subset(Cv653, padj<=qvaluecutoff & abs(log2FoldChange)>=log2cutc
  rownames(subset(Cv111, padj<=qvaluecutoff & abs(log2FoldChange)>=log2cutc
  rownames(subset(i653v111, padj<=qvaluecutoff & abs(log2FoldChange)>=log2c
```

> - Build assay object

In [61]:

```
heat <- assay(rld)[diffXGenes,]
```

Check for unique genes between two datasets

In [85]:

```
diffXGenes[!(diffXGenes %in% DEGS$Gene_ID)]
```

- **Isolate genes for each comparison and sort by the log2 fold change estimates**

- Down-regulated genes

In [63]:
```
resSig_Cv653 = subset(Cv653, padj < 0.001)# control vs 653
```

In [64]:
```
head(resSig_Cv653[ order(resSig_Cv653$log2FoldChange), ])
```

```
log2 fold change (MAP): group 653 vs Control
Wald test p-value: group 653 vs Control
DataFrame with 6 rows and 6 columns
                             baseMean    log2FoldChange             lfcSE
                            <numeric>         <numeric>         <numeric>
Cluster-67248.65881  202.951197945512 -6.28646721243566 0.321043460493498
Cluster-67248.4354   114.361793332159 -6.10139058349401 0.324703217984243
Cluster-6227.0        28.413370682172 -3.81566588608497 0.364050345593477
Cluster-67248.121974 90.6587676924926 -3.64539680421656 0.338107941310253
Cluster-67248.98511  19.2627686051595 -3.18336169460425 0.366931856249029
Cluster-7595.0       18.0599056187062 -3.13211673429244 0.366926109301981
                                  stat               pvalue
                             <numeric>            <numeric>
Cluster-67248.65881  -11.6826157068244 1.56407165592058e-31
Cluster-67248.4354   -9.22015418935383 2.96673297884751e-20
Cluster-6227.0       -7.10607300633044 1.19391276197536e-12
Cluster-67248.121974 -7.68644909581806  1.5127483451757e-14
Cluster-67248.98511   -6.4482580684916 1.13142991747744e-10
Cluster-7595.0       -6.38722903844587 1.68918547510571e-10
                                     padj
                                <numeric>
Cluster-67248.65881   1.6475305194805e-26
Cluster-67248.4354   1.04167928353294e-15
Cluster-6227.0       1.04801662246197e-08
Cluster-67248.121974 2.65578099479047e-10
Cluster-67248.98511  4.25643934955012e-07
Cluster-7595.0       5.70727227013363e-07
```

In [65]:
```
summary(resSig_Cv653)
```

```

out of 184 with nonzero total read count
adjusted p-value < 0.001
LFC > 0 (up)       : 73, 40%
LFC < 0 (down)     : 111, 60%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 6)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results

```

In [66]:
```
resSig_Cv111 = subset(Cv111, padj < 0.001)# control vs 111
```

In [67]:
```
head(resSig_Cv111[ order(resSig_Cv111$log2FoldChange), ])
```

```
log2 fold change (MAP): group 111 vs Control
Wald test p-value: group 111 vs Control
DataFrame with 6 rows and 6 columns
                            baseMean    log2FoldChange             lfcSE
                           <numeric>         <numeric>         <numeric>
Cluster-67248.65881 202.951197945512 -6.55892034953777 0.326410632489976
Cluster-67248.4354  114.361793332159 -5.86374025346785 0.316705178166855
Cluster-6227.0       28.413370682172 -3.76118027552175 0.363723663894864
Cluster-67248.50623 55.2010260797926 -3.45561826424811 0.360606210708655
Cluster-67248.98511 19.2627686051595 -3.19922469856714 0.367038778951685
Cluster-7595.0      18.0599056187062 -3.07729077888003 0.367181471384888
                                 stat               pvalue
padj
                            <numeric>            <numeric>             <nu
meric>
Cluster-67248.65881 -9.87466391769537 5.36123751216411e-23 3.105821324875
74e-19
Cluster-67248.4354  -10.8839277483815 1.37507301949253e-27 1.681699024250
26e-23
Cluster-6227.0      -6.39886290937604 1.56538307394028e-10 7.491310850675
34e-08
Cluster-67248.50623  -7.5734168754873 3.63533356619918e-14 5.065032029088
33e-11
Cluster-67248.98511 -6.49184279953414 8.47927119486211e-11 4.423245977001
32e-08
Cluster-7595.0      -5.70215111106775 1.18304878101351e-08 2.925388058689
15e-06
```

In [68]:
```
resSig_653v111 = subset(i653v111, padj < 0.001)# 653 vs 111
```

```
In [69]: head(resSig_653v111[ order(resSig_653v111$log2FoldChange), ])
```

```
log2 fold change (MAP): group 653 vs 111
Wald test p-value: group 653 vs 111
DataFrame with 6 rows and 6 columns
                             baseMean    log2FoldChange             lfcSE
                            <numeric>         <numeric>         <numeric>
Cluster-67248.121974 90.6587676924926  -3.9589940802961 0.336368110261914
Cluster-67248.112206 18.5626354004593 -2.69648250849936 0.366845367519375
Cluster-67248.76854  24.3571393865261 -2.53083331116833 0.361370461597882
Cluster-67248.149503 27.8249871261726 -2.38866086700844 0.364384606107117
Cluster-67248.75344  15.0449546991255 -2.15398576339566 0.364284669558007
Cluster-67248.148461 9.72151122225922 -1.81687736592331 0.345862164163721
                                  stat              pvalue
                             <numeric>           <numeric>
Cluster-67248.121974 -7.99585298827209 1.28679849746251e-15
Cluster-67248.112206 -5.50382972112057 3.71628503530835e-08
Cluster-67248.76854  -6.18707077234617 6.12924522094512e-10
Cluster-67248.149503  -6.2069334740323 5.40284555384753e-10
Cluster-67248.75344  -5.56746651963085 2.58469719742434e-08
Cluster-67248.148461 -4.98768053367799 6.11085072523991e-07
                                     padj
                                <numeric>
Cluster-67248.121974 1.46826925556716e-11
Cluster-67248.112206 6.65156726652974e-05
Cluster-67248.76854  2.79744881129156e-06
Cluster-67248.149503 2.59569762024374e-06
Cluster-67248.75344  4.91534019948517e-05
Cluster-67248.148461 0.000633875768069715
```

- **Up-regulated genes**

```
In [70]: head(resSig_Cv653[ order(resSig_Cv653$log2FoldChange, decreasing = TRUE), ]
```

```
log2 fold change (MAP): group 653 vs Control
Wald test p-value: group 653 vs Control
DataFrame with 6 rows and 6 columns
                              baseMean   log2FoldChange             lfcSE
                             <numeric>        <numeric>         <numeric>
Cluster-67248.41609  124.279390389592 3.52725893385813 0.345570734128572
Cluster-67248.84245  40.9703792977569  3.4955170912277 0.366553535763028
Cluster-67248.13909   121.78368924237 3.42970727154396 0.299340298106147
Cluster-67248.142094 20.2956372731499  2.7908748658698 0.366388606928481
Cluster-71973.0      21.3301738242477 2.29032146110244 0.365199165612169
Cluster-67248.115536 79.6526611339483   2.288392884457 0.359349809555629
                                  stat               pvalue
padj
                             <numeric>            <numeric>            <nu
meric>
Cluster-67248.41609  7.64229899956177 2.13376951603116e-14 3.210896367723
69e-10
Cluster-67248.84245  6.77707740483589 1.22631195264817e-11 7.176377546897
06e-08
Cluster-67248.13909  10.0175098153321 1.27679313466229e-23 6.724614081639
37e-19
Cluster-67248.142094  5.1099591792728 3.22228426089691e-07 0.000265173855
395185
Cluster-71973.0      5.68570176100883  1.3027657593149e-08 2.213357000374
11e-05
Cluster-67248.115536 6.60449100950853  3.9888541638035e-11 2.000809248563
83e-07
```

```
In [71]: head(resSig_Cv111[ order(resSig_Cv111$log2FoldChange, decreasing = TRUE), ]
```

```
log2 fold change (MAP): group 111 vs Control
Wald test p-value: group 111 vs Control
DataFrame with 6 rows and 6 columns
                             baseMean   log2FoldChange            lfcSE
                            <numeric>        <numeric>        <numeric>
Cluster-67248.41609  124.279390389592 3.38418634239652 0.345620815354031
Cluster-67248.13909   121.78368924237 3.38307889682002 0.299295325657117
Cluster-67248.155958 51.2696932074041 2.48054331596675 0.367824731695813
Cluster-67248.112206 18.5626354004593 2.46892943178247 0.365128892872883
Cluster-67248.115536 79.6526611339483 2.43904926524883 0.359223108777017
Cluster-67248.76854  24.3571393865261 2.43072270277489 0.358678709538164
                                 stat               pvalue
padj
                            <numeric>            <numeric>            <nu
meric>
Cluster-67248.41609    7.496106389997 6.57414375079722e-14 8.824505225689
02e-11
Cluster-67248.13909  9.90931903031448  3.7922199806499e-23 2.608786631563
46e-19
Cluster-67248.155958 6.54049508941158 6.13155106650027e-11  3.29216436262
74e-08
Cluster-67248.112206 4.80749394072171 1.52834049779145e-06 0.000170438612
210139
Cluster-67248.115536 6.88865925610933 5.63206601016873e-12 4.132772491155
08e-09
Cluster-67248.76854  5.70729105119387 1.14788402827629e-08 2.858516902903
69e-06
```

In [72]:
```
head(resSig_653v111[ order(resSig_653v111$log2FoldChange, decreasing = TRUE
```

```
log2 fold change (MAP): group 653 vs 111
Wald test p-value: group 653 vs 111
DataFrame with 6 rows and 6 columns
                             baseMean   log2FoldChange             lfcSE
                            <numeric>        <numeric>         <numeric>
Cluster-67248.84245  40.9703792977569 3.70463793051371  0.36669338917001
Cluster-67248.142094 20.2956372731499 3.14473703586671 0.367133239478846
Cluster-67248.87571  23.3408647625934 2.36391076747395 0.364787539610653
Cluster-67248.132953 24.1723390703805 2.25356877582255 0.357912134362856
Cluster-67248.88523  30.3263718816141 2.03338159861437 0.362475626984758
Cluster-67248.50623  55.2010260797926 1.97704579017896 0.361559396776706
                                 stat               pvalue                
padj
                            <numeric>            <numeric>            <nu
meric>
Cluster-67248.84245    7.342528006312  2.0959664603159e-13 1.913240104305
56e-09
Cluster-67248.142094   6.552319395312 5.66502129690468e-11 2.872858189022
52e-07
Cluster-67248.87571  6.10019355209696 1.05940104359915e-09 4.029343585909
08e-06
Cluster-67248.132953 5.80984050003217 6.25323892095124e-09 1.585578208839
64e-05
Cluster-67248.88523  5.64312893998592 1.66987190358004e-08 3.464301070513
49e-05
Cluster-67248.50623  5.93363419791334 2.96301598160793e-09 9.015667494437
84e-06
```

Build Table of DEGs

Bind all DEGs from each comparisons, by rows

In [73]:
```
DEGs=rbind(c(resSig_Cv653@rownames,
             resSig_Cv111@rownames,
             resSig_653v111@rownames))
```

IDs

In [74]:
```
ID=rep(c("Cv653", "Cv111", "653v111"),
              c(length(resSig_Cv653@rownames),
               length(resSig_Cv111@rownames),
               length(resSig_653v111@rownames)))
```

Bind DEGs and IDs by column

In [75]:
```
DEGS=data.frame(as.character(DEGs),ID)
```

Rename Columns

```
In [76]: colnames(DEGS) = c("Gene_ID","Comparison")
```

Sanity checks

```
In [77]: length(DEGs)
```

```
1833
```

```
In [78]: length(ID)
```

```
1833
```

```
In [79]: length(resSig_Cv653@rownames)
```

```
184
```

```
In [80]: length(resSig_Cv111@rownames)
```

```
1551
```

```
In [81]: length(resSig_653v111@rownames)
```

```
98
```

```
In [82]: length(resSig_Cv653@rownames)+length(resSig_Cv111@rownames)+length(resSig_6
```

```
1833
```

write data into table

```
In [68]: write.csv(DEGS,
                   file="Mentha_DEGs.csv")
```

Merge Gene Names with DEGs

Create column with rownames

```
In [83]: Cv653$id <- rownames(Cv653)
         Cv111$id <- rownames(Cv111)
         i653v111$id <- rownames(i653v111)
```

Merge data tables

```
In [84]: Cv653_GeneNames <- merge(as(Cv653,"data.frame"), gnDF, by="id")
         Cv111_GeneNames <- merge(as(Cv111,"data.frame"), gnDF, by="id")
         i653v111_GeneNames <- merge(as(i653v111,"data.frame"), gnDF, by="id")
```

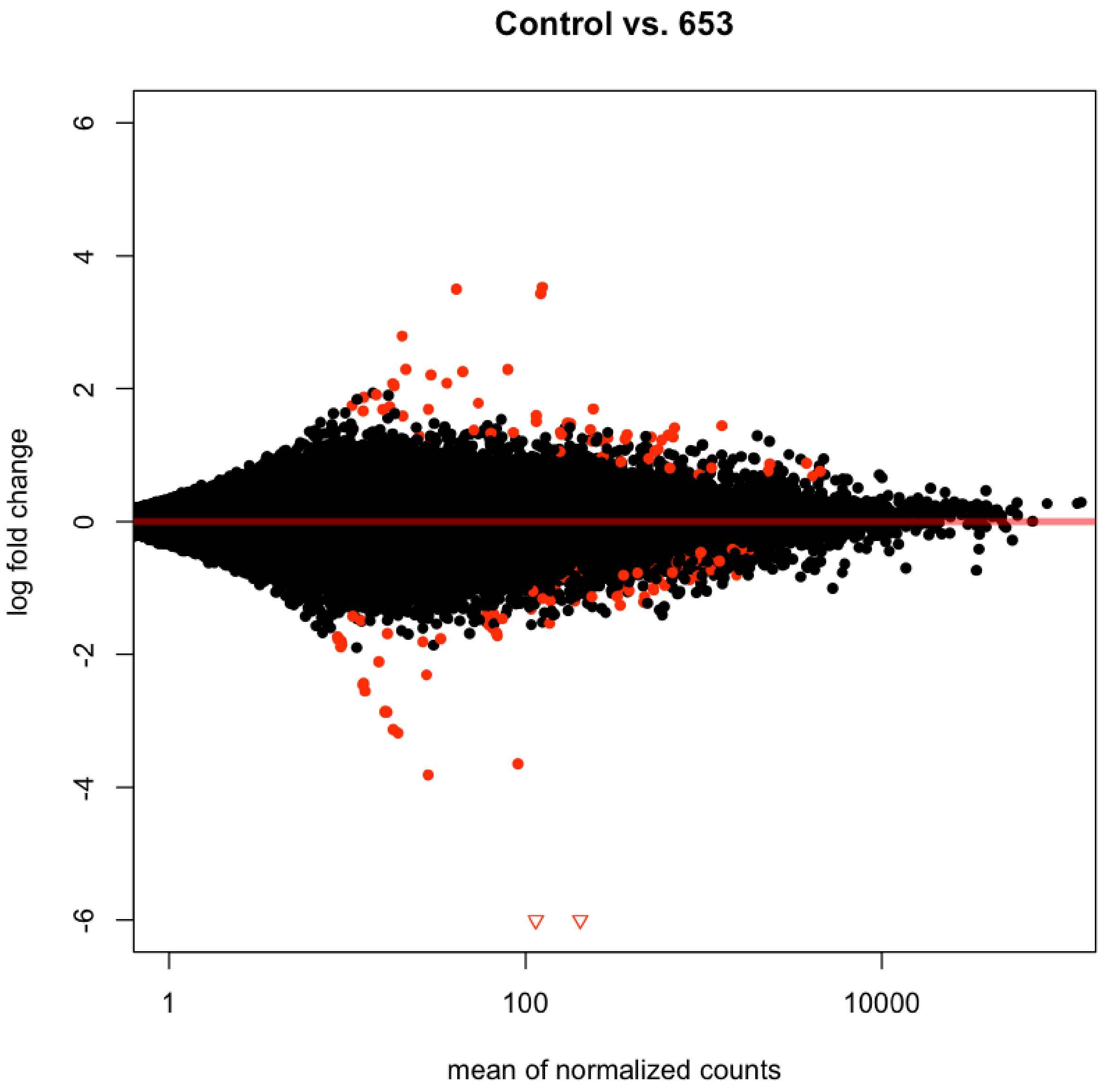
Visualization

Back to Table of Contents

MA plot

> - Control vs. 653

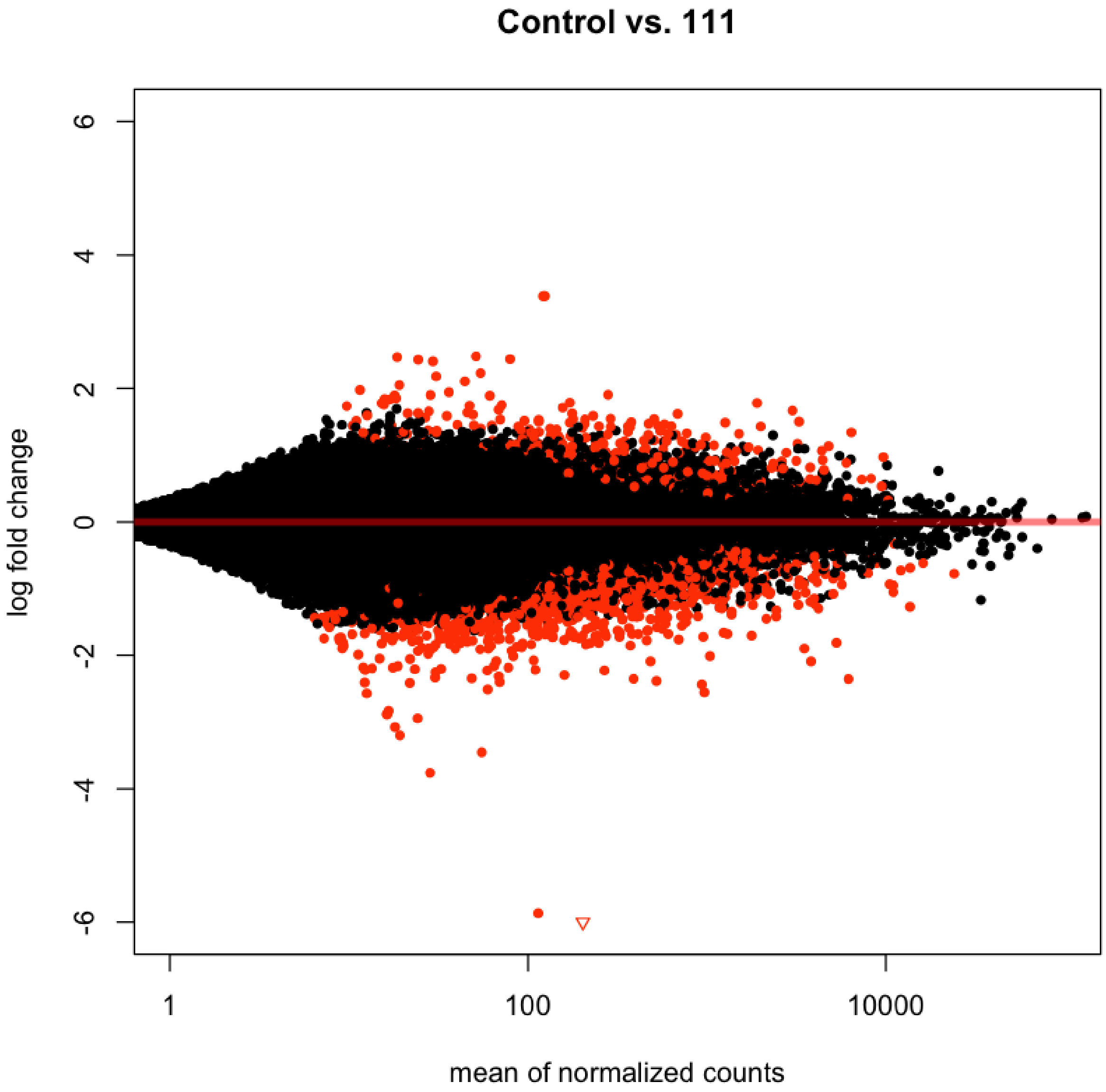
```
In [55]: setwd("/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/FIGURES")
In [56]: #tiff("Mint_MA-P_Cv653.tiff", width=5, height=5, units='in', res=300)
         xlim <- c(1,1e5); ylim <- c(-6,6)
         DESeq2::plotMA(Cv653,
                             xlim=xlim, ylim=ylim,
                             cex=0.9, alpha = 0.001,
                             main="Control vs. 653",
                             colNonSig="black",colSig="orangered1")
         #dev.off()
```



- 111 vs control

In [57]:

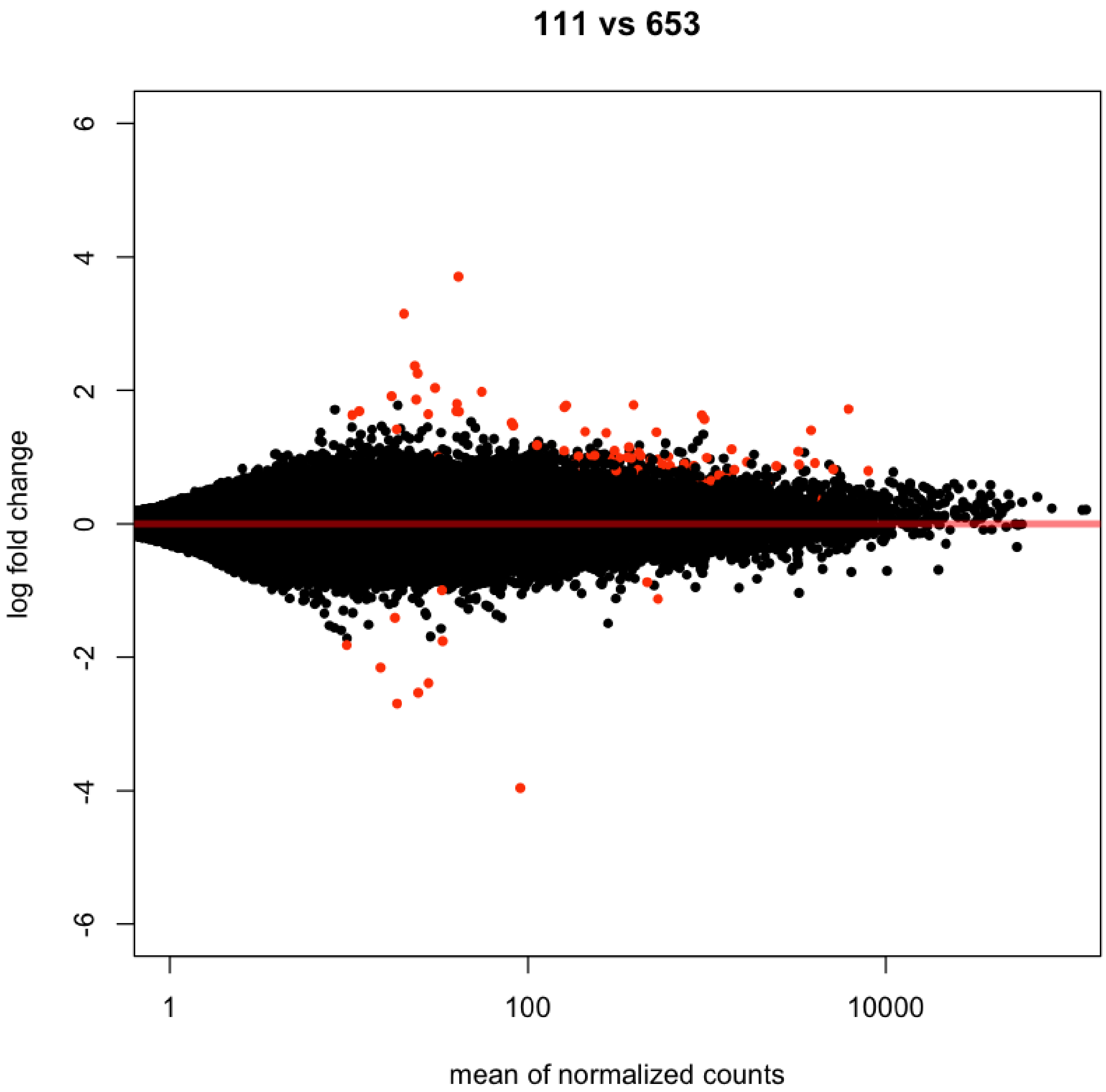
```
#tiff("Mint_MA-P_Cv111.tiff", width=5, height=5, units='in', res=300)
xlim <- c(1,1e5); ylim <- c(-6,6)
DESeq2::plotMA(Cv111,
                xlim=xlim, ylim=ylim,
                cex=0.8, alpha = 0.001,
                main="Control vs. 111",
                colNonSig="black",colSig="orangered1")
#dev.off()
```



- 111 vs. 653

In [58]:

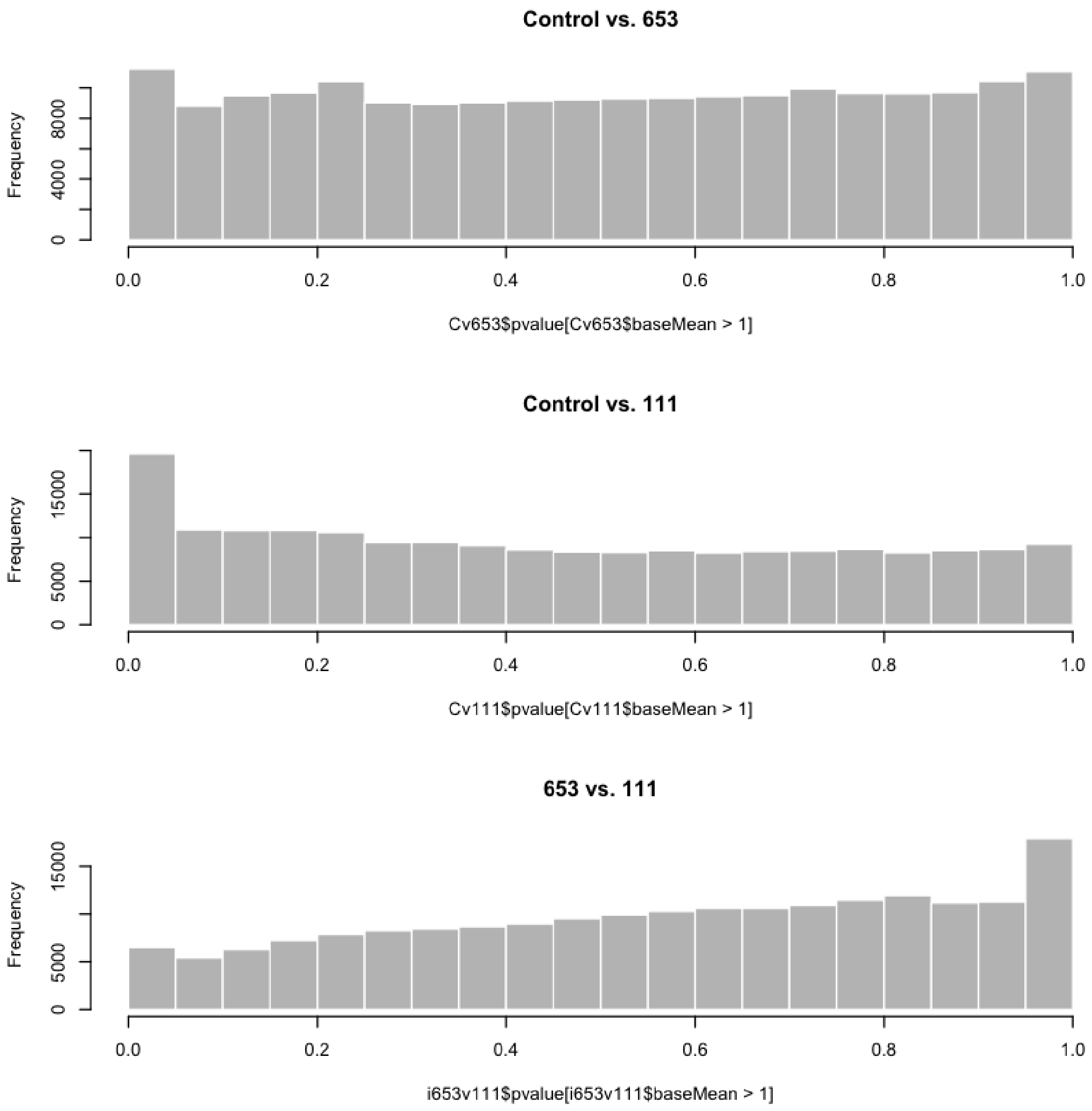
```
#tiff("Mint_MA-P_111v653.tiff", width=5, height=5, units='in', res=300)
xlim <- c(1,1e5); ylim <- c(-6,6)
DESeq2::plotMA(i653v111,
                xlim=xlim, ylim=ylim,
                cex=0.8, alpha = 0.001,
                main="111 vs 653",
                colNonSig="black",colSig="orangered1")
#dev.off()
```



Histograms of p-values

In [59]:

```
par(mfrow=c(3,1))
hist(Cv653$pvalue[Cv653$baseMean > 1], breaks = 0:20/20,
     col = "grey75", border = "white", main = "Control vs. 653")
hist(Cv111$pvalue[Cv111$baseMean > 1], breaks = 0:20/20,
     col = "grey75", border = "white", main = "Control vs. 111")
hist(i653v111$pvalue[i653v111$baseMean > 1], breaks = 0:20/20,
     col = "grey75", border = "white", main = "653 vs. 111")
```



Volcano plots

- 653 vs control

In [69]:

```
Cv653_gn = subset(Cv653_GeneNames, Comparison == "Cv653")
```

- Order genes by fold change values

In [76]:

```
Cv653_gn = (Cv653_gn[order(-abs(Cv653_gn$log2FoldChange)),])
```

Convert 10 DEG gene names from factors to vectors

In [119]:

```
labs = head(Cv653_gn$Hit1_acc, n=20)
lab = as.character(labs)
lab
labels = c('CNGC5_ARATH', 'RBS2_BRANA', 'RBS2_BRANA', 'ARP3_ARATH', 'FB30_A
           'PMTK_ARATH', 'CO1A1_HUMAN', 'CB21_SINAL', 'CB5_ARATH', 'TEX10_H
           'BCA1_ARATH', 'CB1C_ARATH', '-', 'CA4_ARATH', 'RCA_ARATH',
           'G3PA2_ARATH', 'PIF1_XENLA', '-','PER45_ARATH', 'PLST1_ARATH')
```

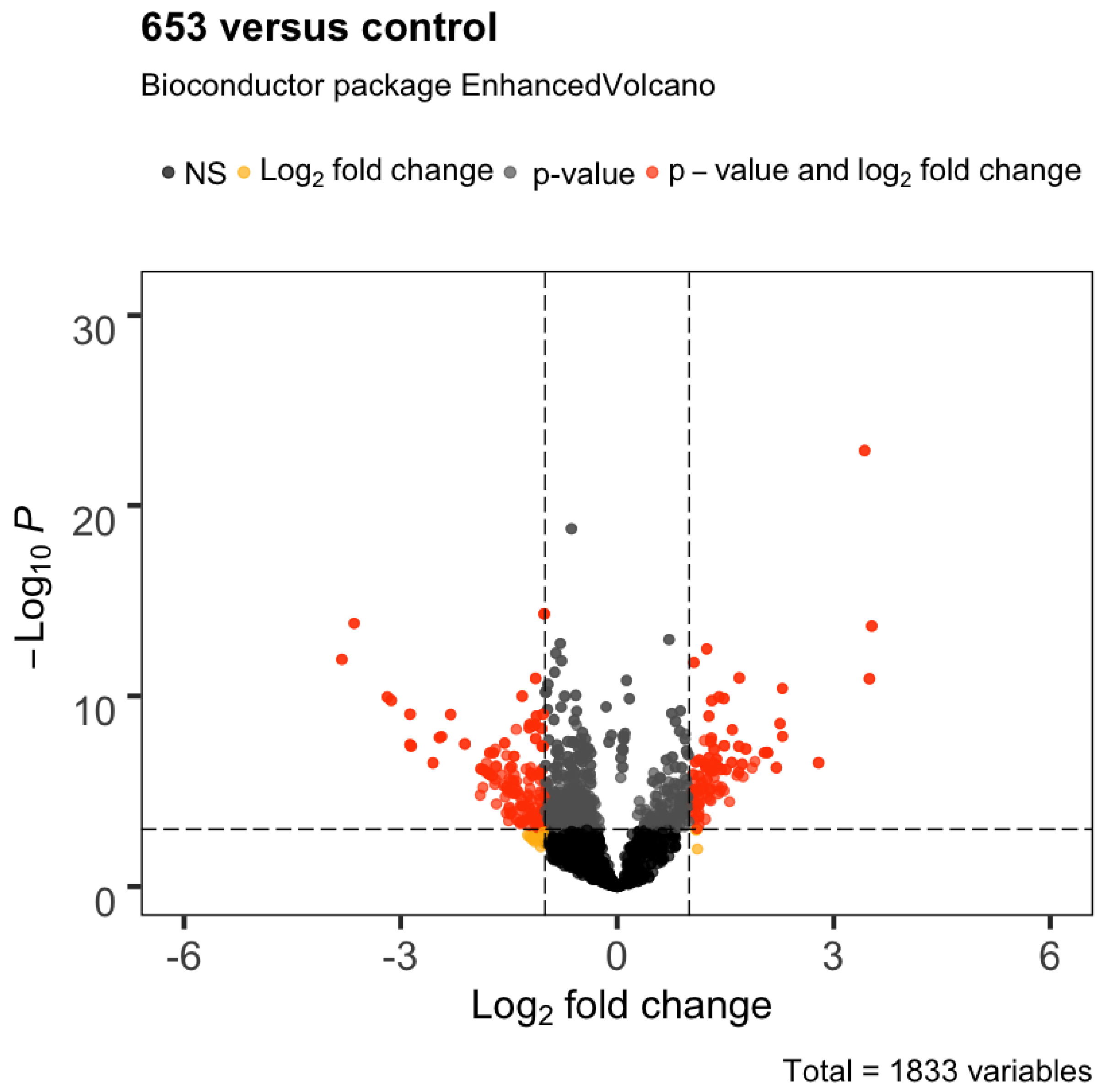
```
'CNGC5_ARATH'   'RBS2_BRANA'   'RBS2_BRANA'   'ARP3_ARATH'   'FB30_ARATH'
'PMTK_ARATH'   'CO1A1_HUMAN'   'CB21_SINAL'   'CB5_ARATH'   'TEX10_HUMAN'
'BCA1_ARATH'   'CB1C_ARATH'   '-'   'CA4_ARATH'   'RCA_ARATH'   'G3PA2_ARATH'
'PIF1_XENLA'   '-'   'PER45_ARATH'   'PLST1_ARATH'
```

In [123]:

```
Cv6=EnhancedVolcano(Cv653_GeneNames,
    lab = NA,
    x = 'log2FoldChange',
    y = 'pvalue',
    title = "653 versus control",
    legend=c("NS","Log2 fold-change","p-value",
    "p-value & Log2 fold-change"),
    legendPosition = "top",
    legendLabSize = 14,
    legendIconSize = 2.0,
    pCutoff = 0.001,
    FCcutoff = 1.0,
    transcriptPointSize = 1.75,
    transcriptLabSize = 3.0,
    colAlpha = 0.7,
    border = "full",
    gridlines.major = FALSE,
    gridlines.minor = FALSE,
    xlim = c(-6, 6),
    ylim = c(0, min(log10(Cv653$pvalue))),
    col=c("black", "darkgoldenrod1", "gray38", "orangered1"))

Cv6 + scale_color_manual(
  values=c(
    NS="black",
    FC="darkgoldenrod1",
    P="gray38",
    FC_P="orangered1"),
  labels=c(
    NS='NS',
    FC=expression(Log[2]~fold~change),
    P="p-value",
    FC_P=expression(p-value~and~log[2]~fold~change)))
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



In [123]:
```
setwd("/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/FIGURES")
```

In [124]:
```
tiff("Mint_VP_Cv653.tiff", width=10, height=10, units='in', res=300)
Cv6
dev.off()
```

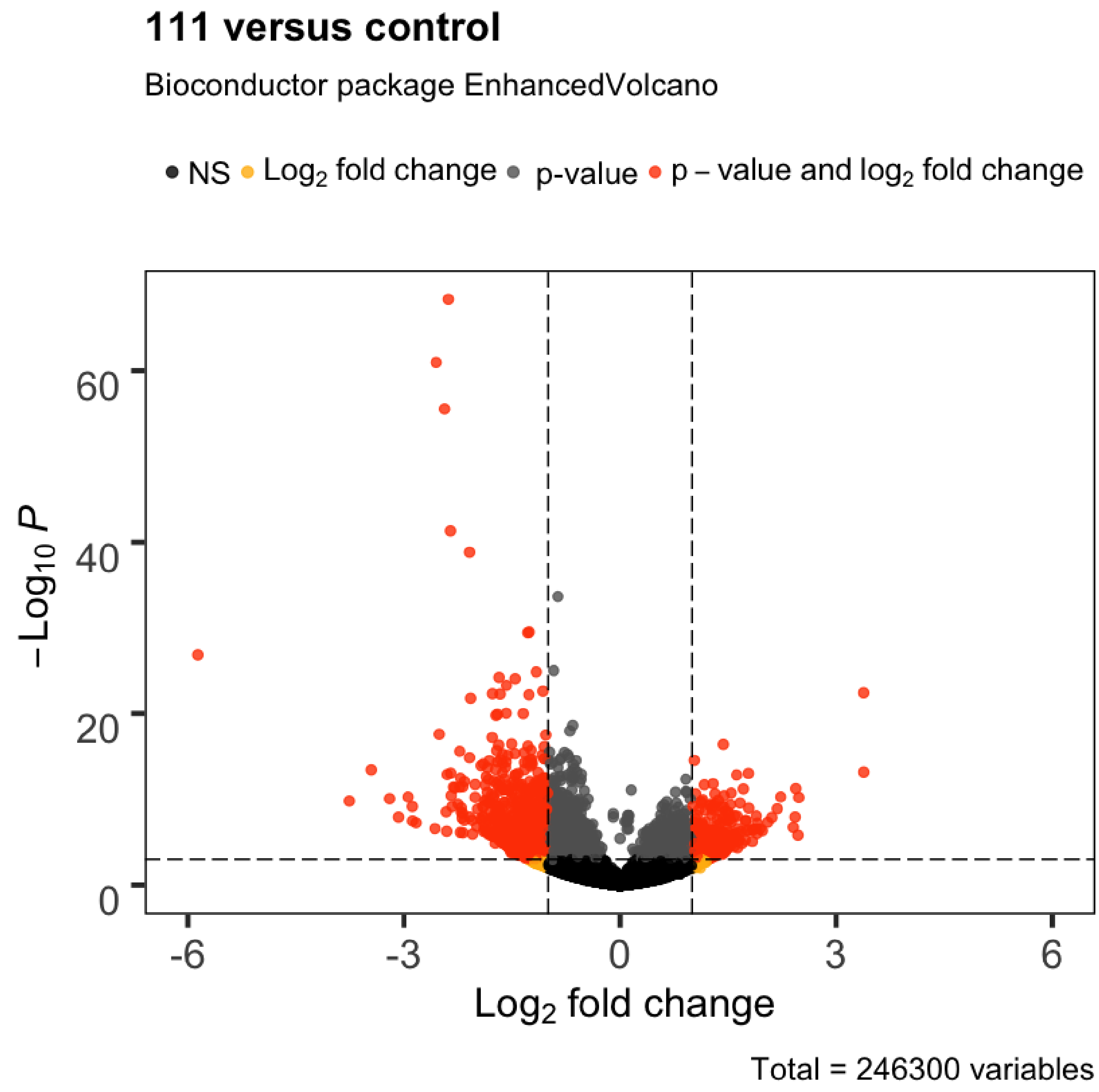
pdf: 2

- 111 vs control

In [125]:

```
Cv1=EnhancedVolcano(Cv111,
    lab = NA,
    x = 'log2FoldChange',
    y = 'pvalue',
    title = "111 versus control",
    legend=c("NS","Log2 fold-change","p-value",
    "p-value & Log2 fold-change"),
    legendPosition = "top",
    legendLabSize = 14,
    legendIconSize = 2.0,
    pCutoff = 0.001,
    FCcutoff = 1.0,
    transcriptPointSize = 1.75,
    transcriptLabSize = 3.0,
    colAlpha = 0.8,
    border = "full",
    gridlines.major = FALSE,
    gridlines.minor = FALSE,
    xlim = c(-6, 6),
    ylim = c(0, min(log10(Cv111$pvalue))),
    col=c("black", "darkgoldenrod1", "gray38", "orangered1"))
Cv1 + scale_color_manual(
  values=c(
    NS="black",
    FC="darkgoldenrod1",
    P="gray38",
    FC_P="orangered1"),
  labels=c(
    NS='NS',
    FC=expression(Log[2]~fold~change),
    P="p-value",
    FC_P=expression(p-value~and~log[2]~fold~change)))
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



```
In [126]: tiff("Mint_VP_Cv111.tiff", width=10, height=10, units='in', res=300)
          Cv1
          dev.off()
```

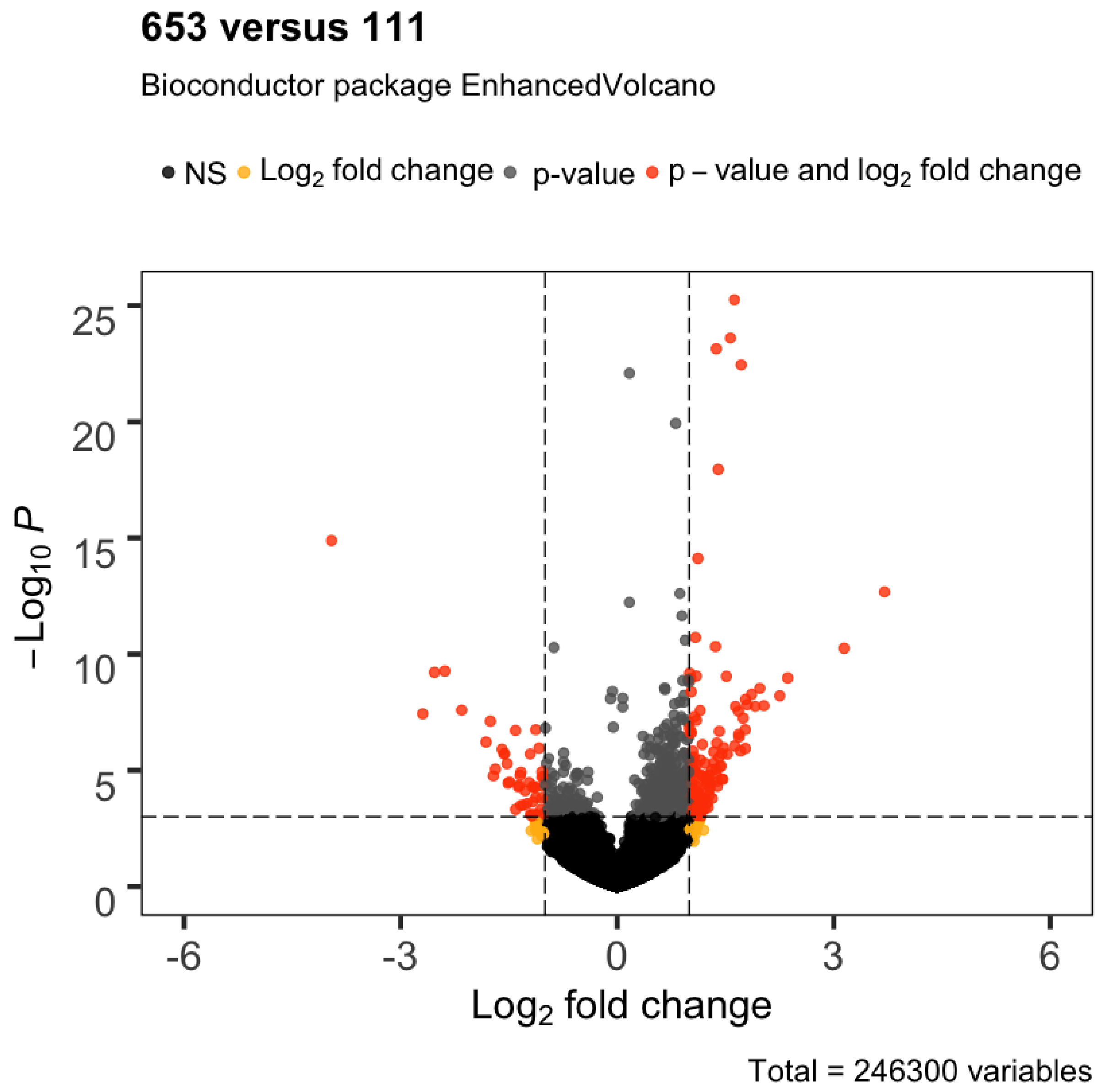
pdf: 2

- 653 vs. 111

In [127]:

```
i6vi1=EnhancedVolcano(i653v111,
    lab = NA,
    x = 'log2FoldChange',
    y = 'pvalue',
    title = "653 versus 111",
    legend=c("NS","Log2 fold-change","p-value",
    "p-value & Log2 fold-change"),
    legendPosition = "top",
    legendLabSize = 14,
    legendIconSize = 2.0,
    pCutoff = 0.001,
    FCcutoff = 1.0,
    transcriptPointSize = 1.75,
    transcriptLabSize = 3.0,
    colAlpha = 0.8,
    border = "full",
    gridlines.major = FALSE,
    gridlines.minor = FALSE,
    xlim = c(-6, 6),
    ylim = c(0, min(log10(i653v111$pvalue))),
    col=c("black", "darkgoldenrod1", "gray38", "orangered1"))
i6vi1 + scale_color_manual(
  values=c(
    NS="black",
    FC="darkgoldenrod1",
    P="gray38",
    FC_P="orangered1"),
  labels=c(
    NS='NS',
    FC=expression(Log[2]~fold~change),
    P="p-value",
    FC_P=expression(p-value~and~log[2]~fold~change)))
```

```
Scale for 'colour' is already present. Adding another scale for 'colour',
which will replace the existing scale.
```



```
In [128]: tiff("Mint_VP_653v111.tiff", width=10, height=10, units='in', res=300)
          i6vi1
          dev.off()
```

pdf: 2

Cluster Genes

Object for heatmap

```
In [70]: heat <- assay(vsd)[diffXGenes,]
```

Convert to dataframe

```
In [71]: hmDF = data.frame(heat)
```

Extract subset of genenames that are differentially expressed

- First, make new column of character gene IDs

In [72]:
```
gnDF["ID"] = as.character(gnDF$id)
```

- Differentially expressed genes

In [73]:
```
degs = subset(gnDF, gnDF$ID %in% rownames(hmDF))
```

- Use only unique/non duplicated rows

In [74]:
```
degs = degs[!duplicated(degs$ID),]
```

- Sanity checks

In [75]:
```
length(degs$ID) == length(rownames(hmDF))
```

```
TRUE
```

In [76]:
```
class(degs$ID) == class(rownames(hmDF))
```

```
TRUE
```

- Reorder rows of both dataframes to align with each other

In [77]:
```
hmDF <- hmDF[with(hmDF, order(rownames(hmDF))), ]
```

In [78]:
```
degs <- degs[with(degs, order(degs$ID)), ]
```

- Sanity check

In [79]:
```
rownames(hmDF) == degs$ID
```

```
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE
```

Convert dataframe to matrix

In [80]:
```
hmdf = as.matrix(hmDF)
```

Replace rownames in heatmap object with gene names

In [81]:
```
rownames(hmdf) <- degs$Hit1_acc
```

Order row labels to italicize

In [179]:
```
ord_row_labs = rownames(hmdf)[(ord_row_labs = rownames(hmdf)[(Mint_DEGs$tre
```

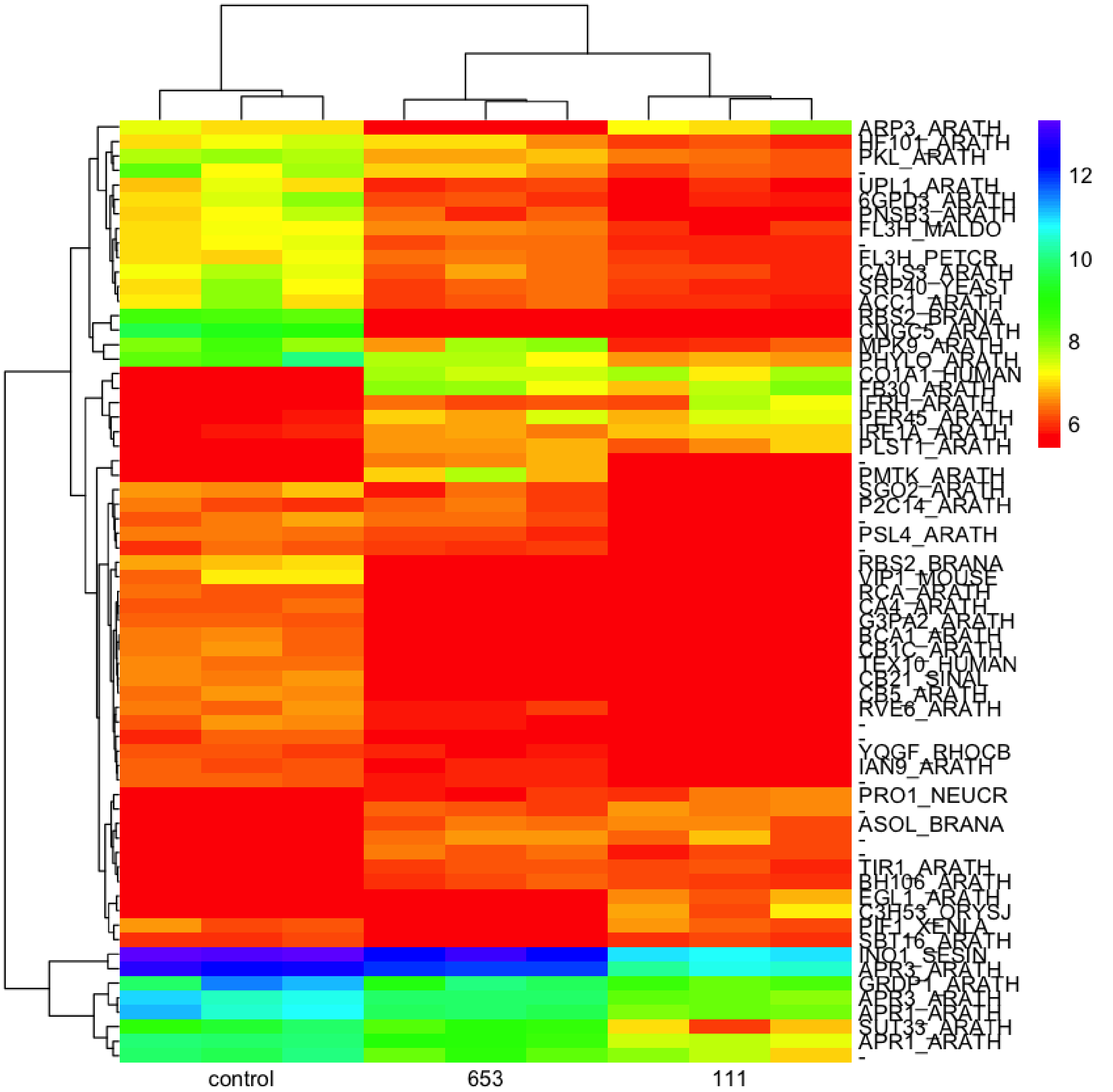
In [180]:
```
ord_row_labs = cat(paste0('"', paste(ord_row_labs, collapse="\", \""), '"')
```

```
"ARP3_ARATH", "HF101_ARATH", "PKL_ARATH", "-", "UPL1_ARATH", "6GPD3_ARAT
H", "PNSB3_ARATH", "FL3H_MALDO", "-", "FL3H_PETCR", "CALS3_ARATH", "SRP40
_YEAST", "ACC1_ARATH", "RBS2_BRANA", "CNGC5_ARATH", "MPK9_ARATH", "PHYLO_
ARATH", "CO1A1_HUMAN", "FB30_ARATH", "IFRH_ARATH", "PER45_ARATH", "IRE1A_
ARATH", "PLST1_ARATH", "-", "PMTK_ARATH", "SGO2_ARATH", "P2C14_ARATH", "-
", "PSL4_ARATH", "-", "RBS2_BRANA", "VIP1_MOUSE", "RCA_ARATH", "CA4_ARAT
H", "G3PA2_ARATH", "BCA1_ARATH", "CB1C_ARATH", "TEX10_HUMAN", "CB21_SINA
L", "CB5_ARATH", "RVE6_ARATH", "-", "-", "YQGF_RHOCB", "IAN9_ARATH", "-",
"PRO1_NEUCR", "-", "ASOL_BRANA", "-", "-", "TIR1_ARATH", "BH106_ARATH",
"EGL1_ARATH", "C3H53_ORYSJ", "PIF1_XENLA", "SBT16_ARATH", "INO1_SESIN",
"APR3_ARATH", "GRDP1_ARATH", "APR3_ARATH", "APR1_ARATH", "SUT33_ARATH",
"APR1_ARATH", "-"
```

Visualize heatmap

In [192]:

```
Mint_DEGs=pheatmap(hmdf,
        color= (rainbow(96,start=0.0,end=0.74,alpha=1)),#,s=1,v=0.6,start=0
        border_color = NA,
        show_colnames = TRUE,
        show_rownames = TRUE,
        labels_col=paste0(c("control"," "," ", "653"," "," ", "111"," "," "
        angle_col=0)
#         labels_row=expression(italic(c("ARP3_ARATH", "HF101_ARATH", "PKL_
#                                        "UPL1_ARATH", "6GPD3_ARATH", "PNSB
#                                        "FL3H_MALDO", "-", "FL3H_PETCR", "
#                                        "SRP40_YEAST", "ACC1_ARATH", "RBS2
#                                        "MPK9_ARATH", "PHYLO_ARATH", "CO1A
#                                        "IFRH_ARATH", "PER45_ARATH", "IRE1
#                                        "-", "PMTK_ARATH", "SGO2_ARATH", "
#                                        "PSL4_ARATH", "-", "RBS2_BRANA", "
#                                        "CA4_ARATH", "G3PA2_ARATH", "BCA1_
#                                        "TEX10_HUMAN", "CB21_SINAL", "CB5_
#                                        "-", "-", "YQGF_RHOCB", "IAN9_ARAT
#                                        "-", "ASOL_BRANA", "-", "-", "TIR1
#                                        "EGL1_ARATH", "C3H53_ORYSJ", "PIF1
#                                        "INO1_SESIN", "APR3_ARATH", "GRDP1
#                                        "APR1_ARATH", "SUT33_ARATH", "APR1
```



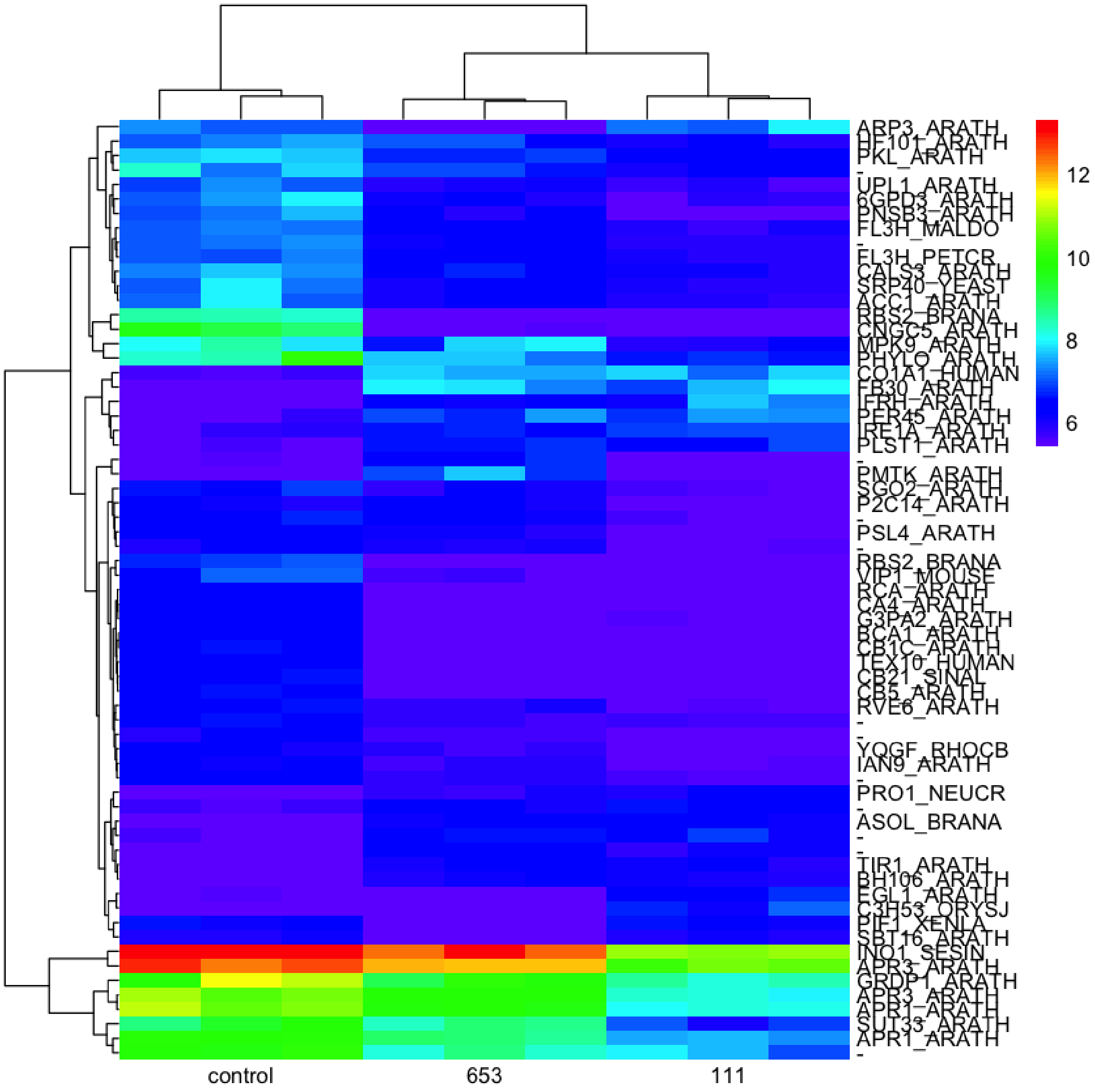
In [231]:

```
setwd('/Users/davidwheeler/Desktop/RESEARCH/Data/TRANSCRIPTOMICS/FIGURES//M
tiff("Mint_HeatMap_1.tiff", width = 5, height =10, units = 'in', res = 300)
Mint_DEGs
dev.off()
```

pdf: 2

In [233]:

```
Mint_DEGs=pheatmap(hmdf,
        color= rev(rainbow(96,start=0.0,end=0.74,alpha=1)),#,s=1,v=0.6,star
        border_color = NA,
        show_colnames = TRUE,
        show_rownames = TRUE,
        labels_col=paste0(c("control"," "," ", "653"," "," ", "111"," "," "
        angle_col=0)
        #labels_col=paste0("bar", 1:10))
```



In [234]:

```
tiff("Mint_HeatMap_2.tiff", width = 5, height =10, units = 'in', res = 300)
Mint_DEGs
dev.off()
```

pdf: 2

Venn Diagram

Control vs 653

```
In [65]: resSig_Cv653 = subset(Cv653, padj < 0.001)
         resSig_Cv653_fragments = row.names(resSig_Cv653)
```

Control vs 111

```
In [66]: resSig_Cv111 = subset(Cv111, padj < 0.001)
         resSig_Cv111_fragments = row.names(resSig_Cv111)
```

653 vs 111

```
In [67]: resSig_653v111 = subset(i653v111, padj < 0.001)
         resSig_653v111_fragments = row.names(resSig_653v111)
```

Build common dataframe

```
In [68]: vdDF = c(resSig_Cv653_fragments,
                  resSig_Cv111_fragments,
                 resSig_653v111_fragments)
```

Compare

```
In [69]: resSig_Cv653_fragments.2 <- vdDF %in% resSig_Cv653_fragments
         resSig_Cv111_fragments.2 <- vdDF %in% resSig_Cv111_fragments
         resSig_653v111_fragments.2 <- vdDF %in% resSig_653v111_fragments
```

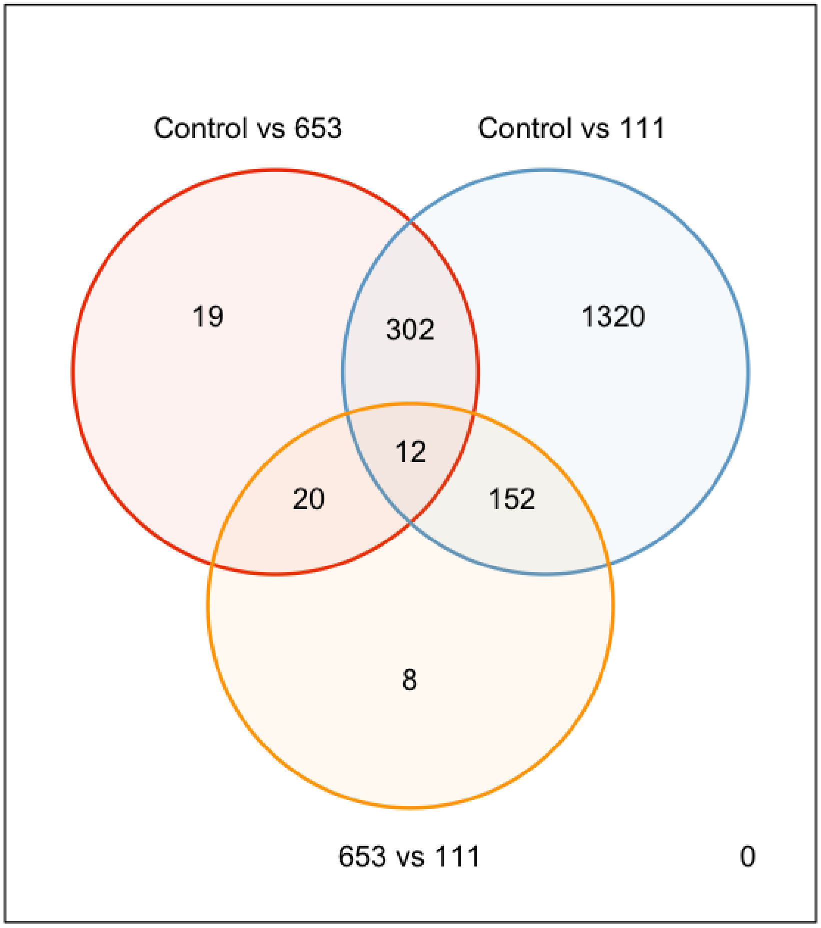
Compute venn diagram counts

```
In [70]: counts = cbind(resSig_Cv653_fragments.2, resSig_Cv111_fragments.2, resSig_6
         vdcounts = vennCounts(counts)
```

Plot

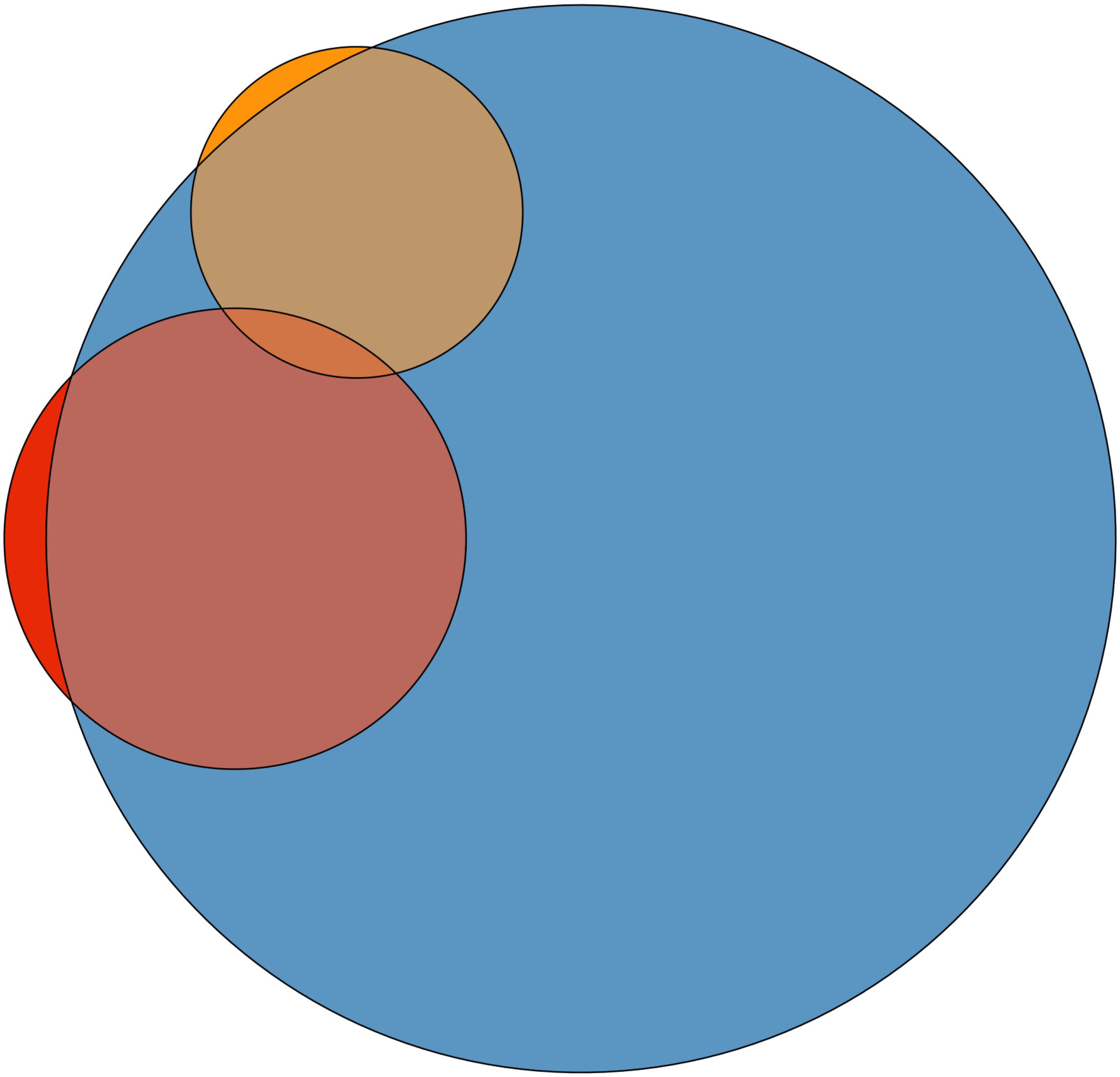
In [71]:
```
#tiff("Mint_VD.tiff", width=5, height=5, units = 'in', res = 300)
vennDiagram(vdcounts,
            cex=1,
            lwd=2,
            names=c("Control vs 653","Control vs 111","653 vs 111"),
            circle.col = c("orangered2","skyblue3","orange1"))

#dev.off()
```



In [65]:

```
#tiff("Mint_VD-2.tiff", width=5, height=5, units = 'in', res = 300)
VD = euler(c(A=19, B=1320, C=8,
             "A&B"=302, "A&C"=20, "B&C"=152,
             "A&B&C"=12))
plot(VD,
     fills = c("orangered2","skyblue3","orange1"),
     lwd=2,cex=1,
     labels = NULL
     )
options(repr.plot.width=15, repr.plot.height=15)
#dev.off()
```



Venn Diagram in Python

In [ ]:

```
!pip install matplotlib-venn
```

In [ ]:

```
venn3(subsets = (19, 1320, 302, 8, 20, 152 ,12),
     set_labels = ('A', 'B', 'D'))
plt.figure(figsize=(20,20))
plt.show()
```

Export data

Back to Table of Contents

- Control vs. 653

In [128]:

```
?read.csv()
```

In [50]:

```
write.csv(as.data.frame(diffXGenes),
          file="Mentha.csv")
```

- Control vs. 111

In [ ]:

```
write.csv(as.data.frame(resSig_Cv111),
          file="Mentha_DEGs_Cv111.csv")
```

- 653 vs 111

In [ ]:

```
write.csv(as.data.frame(resSig_653v111),
          file="Mentha_DEGs_653v111.csv")
```

Graveyard

Back to Table of Contents

In [ ]:

```
# For each sample/row in the sample column
for (row in DF$Sample){
    # Split the sample by the underscore
    sample = strsplit(row,"_")
    # If sample contains S2,
    if (grepl("S2",sample)){
    }
}
```

Resources

Back to Table of Contents

- R kernel installation: https://irkernel.github.io/installation/ (https://irkernel.github.io/installation/)

- DESeq2 installation: https://anaconda.org/bioconda/bioconductor-deseq2 (https://anaconda.org/bioconda/bioconductor-deseq2)
- DESeq2 data curation: https://www.bioconductor.org/packages/devel/bioc/vignettes/DEFormats/inst/doc/DEFormats.ht (https://www.bioconductor.org/packages/devel/bioc/vignettes/DEFormats/inst/doc/DEFormats.h
- DESeq2 vignette: http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html (http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html) http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnas (http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnas
- Volcano plots: https://rdrr.io/github/kevinblighe/EnhancedVolcano/f/README.md (https://rdrr.io/github/kevinblighe/EnhancedVolcano/f/README.md)
    - https://www.bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/l (https://www.bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/

Type *Markdown* and LaTeX: α^2

In [ ]: