WGCNA – Data Cleaning

```r
rm(list=ls())

library(jpeg)
library(dplyr)
library(tidyr)
library(tibble)
library(stringr)
library(ggplot2)

library(foreach)
library(iterators)
library(parallel)
library(doParallel)

library(WGCNA)
# library(KEGGREST)
# library(biomaRt)

set.seed(1)

# Enable WGCNA threads to speed up calculations
enableWGCNAThreads()


```

```r
##################################################
# Output folder
##################################################
output_path <- file.path("/home/ycth8/data/projects/05_30_2021_summer_WGCNA/Maize_proteomics_output/2021_06_10_data_cleaning")

if(!dir.exists(output_path)){
  dir.create(output_path, showWarnings=FALSE, recursive=TRUE)
  if(!dir.exists(output_path)){
    quit(status=1)
  }
}


```

```r
##################################################
# Read in input file
##################################################

folder_path = file.path("/home/ycth8/data/projects/05_30_2021_summer_WGCNA/Maize_proteomics_data")

dat = read.csv(
  file = file.path(folder_path, "3plus_2fold_genelist.csv"),
  header = TRUE,
  check.names = FALSE,
  stringsAsFactors = FALSE
)

datTraits = read.csv(
  file = file.path(folder_path, "PBAA_B73o2_Abs_Raw.csv"),
  header = TRUE,
  check.names = FALSE,
  stringsAsFactors = FALSE
)


```

Original Genotype Input File

Accession	Description	O2_10_1	O2_10_2	O2_10_3	O2_14_1	O2_14_2	O2_14_3	O2_18_1	O2_18_2	O2_18_3	O2_22_1	O2_22_2
Zm00001d004855	Zm00001d004855	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d004960	Zm00001d004960	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d005090	Clathrin heavy chain	210	207	234	194	209	242	201	218	217	222	198
Zm00001d006433	Ras-related protein Rab-6A	23	26	23	18	20	23	15	19	20	19	18
Zm00001d006596	Elongation factor 1-alpha	120	129	127	123	141	147	136	138	123	153	145
Zm00001d006651	TUBA5, Tubulin alpha chain	165	134	153	122	114	111	106	107	114	109	97
Zm00001d007048	Zm00001d007048	36	31	28	32	22	34	33	31	31	29	27
Zm00001d007758	Zm00001d007758	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d007900	Eukaryotic translation initiation factor 3 subunit I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d007936	Pyridoxin biosynthesis protein ER1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d008651	Pyruvate decarboxylase	22	22	25	30	31	36	38	33	38	41	39
Zm00001d008739	Zm00001d008739	7	9	11	10	4	6	5	0.5	0.5	8	4
Zm00001d008743	Ubiquitinyl hydrolase 1	20	16	22	24	33	22	22	32	33	33	36
Zm00001d009311	Zm00001d009311	8	9	10	10	6	9	7	7	8	9	9
Zm00001d009652	Zm00001d009652	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d009653	Zm00001d009653	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d009708	Putative calcium-dependent protein kise family protein isoform ...	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d011183	THI1-1, Thiamine thiazole synthase chloroplastic [S...	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d014073	60S ribosomal protein L23	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Zm00001d014844	0.5	19	21	17	40	53	54	59	63	61	54	60

Original Phenotype Input File

Genotype	DAP	Replicaiton	Ala	Arg	Asx	Glx	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser
B73	10	1	168.34	26.01	147.24	118.77	185.03	25.52	52.02	65.28	49.08	26.01	25.52	93.25	82.45
B73	10	2	182.93	25.85	144.39	120.00	180.98	24.88	53.17	64.39	48.78	24.39	26.34	92.20	80.98
B73	10	3	190.85	26.14	131.24	120.78	191.90	27.71	49.15	61.18	50.72	24.05	26.67	88.37	84.18
B73	14	1	120.34	20.06	82.96	89.34	155.44	25.53	30.54	41.48	40.57	14.13	20.97	59.26	53.33
B73	14	2	140.35	23.86	94.04	89.82	172.63	24.80	48.19	66.90	44.91	18.71	25.26	56.61	65.96
B73	14	3	130.87	22.70	87.00	74.14	137.30	19.29	49.93	73.38	38.20	17.78	21.18	48.42	65.44
B73	18	1	126.78	23.86	84.21	85.15	161.87	24.33	58.95	94.97	43.51	18.71	26.20	60.82	66.43
B73	18	2	116.69	24.97	81.53	93.25	171.21	27.01	61.66	101.91	44.84	20.89	28.03	66.24	68.28
B73	18	3	114.58	23.74	74.32	88.26	160.52	27.87	54.19	90.32	43.35	17.03	27.87	59.87	60.90
B73	22	1	153.83	24.68	89.04	95.21	163.53	21.60	71.85	132.67	40.99	21.60	25.56	78.02	71.85
B73	22	2	135.80	20.23	80.90	82.35	132.91	16.98	69.71	123.88	33.95	22.39	21.31	68.98	65.73
B73	22	3	128.52	21.99	78.92	87.55	150.51	21.99	62.10	116.87	39.25	18.98	24.58	69.43	62.96
B73	26	1	135.53	23.53	85.18	93.18	162.82	24.47	60.71	106.82	38.59	17.88	26.35	66.35	62.59
B73	26	2	144.78	26.28	94.06	102.36	165.07	22.59	75.16	136.48	41.50	24.44	26.74	82.54	79.77
B73	26	3	199.47	29.47	107.89	117.37	193.68	26.32	95.79	168.42	45.79	25.26	31.05	96.32	87.37
B73	30	1	167.27	28.05	89.35	107.01	183.38	26.49	78.96	148.57	42.60	20.26	29.61	84.16	75.84
B73	30	2	143.66	25.13	80.42	88.38	148.27	20.52	71.62	121.88	38.53	20.94	23.87	68.27	69.53

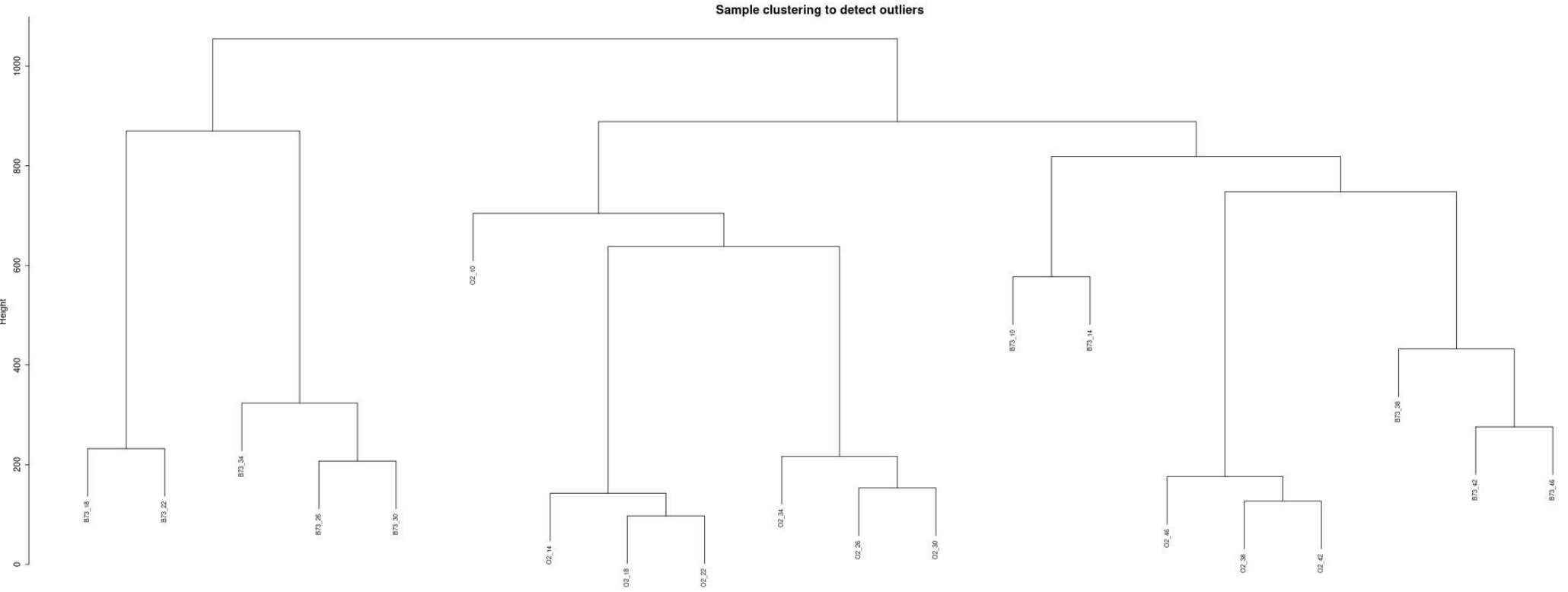
```
69  ##################################################
70  # Process the input file
71  ##################################################
72
73  datExpr0 = dat %>%
74    pivot_longer(!c(Accession, Description), names_to = "Genotype", values_to = "Measurement") %>%
75    separate(Genotype, c("Genotype", "Repetition"), sep = "\\_(?=[^\\_]+$)", extra = "drop", fill = "right") %>%
76    group_by(Accession, Description, Genotype) %>%
77    summarise_at(vars(Measurement), list(Measurement = mean)) %>%
78    ungroup() %>%
79    select(-Description) %>%
80    pivot_wider(names_from = Accession, values_from = Measurement, values_fill = NA) %>%
81    as.data.frame(stringsAsFactors = FALSE)
82
83  # Copy values in column 1 into row names
84  rownames(datExpr0) <- datExpr0[,1]
85
86  # Remove column 1
87  datExpr0 <- datExpr0[,-1]
88
89  print(head(datExpr0))
90  print(tail(datExpr0))
91  print(dim(datExpr0))
92
```

Processed Input File

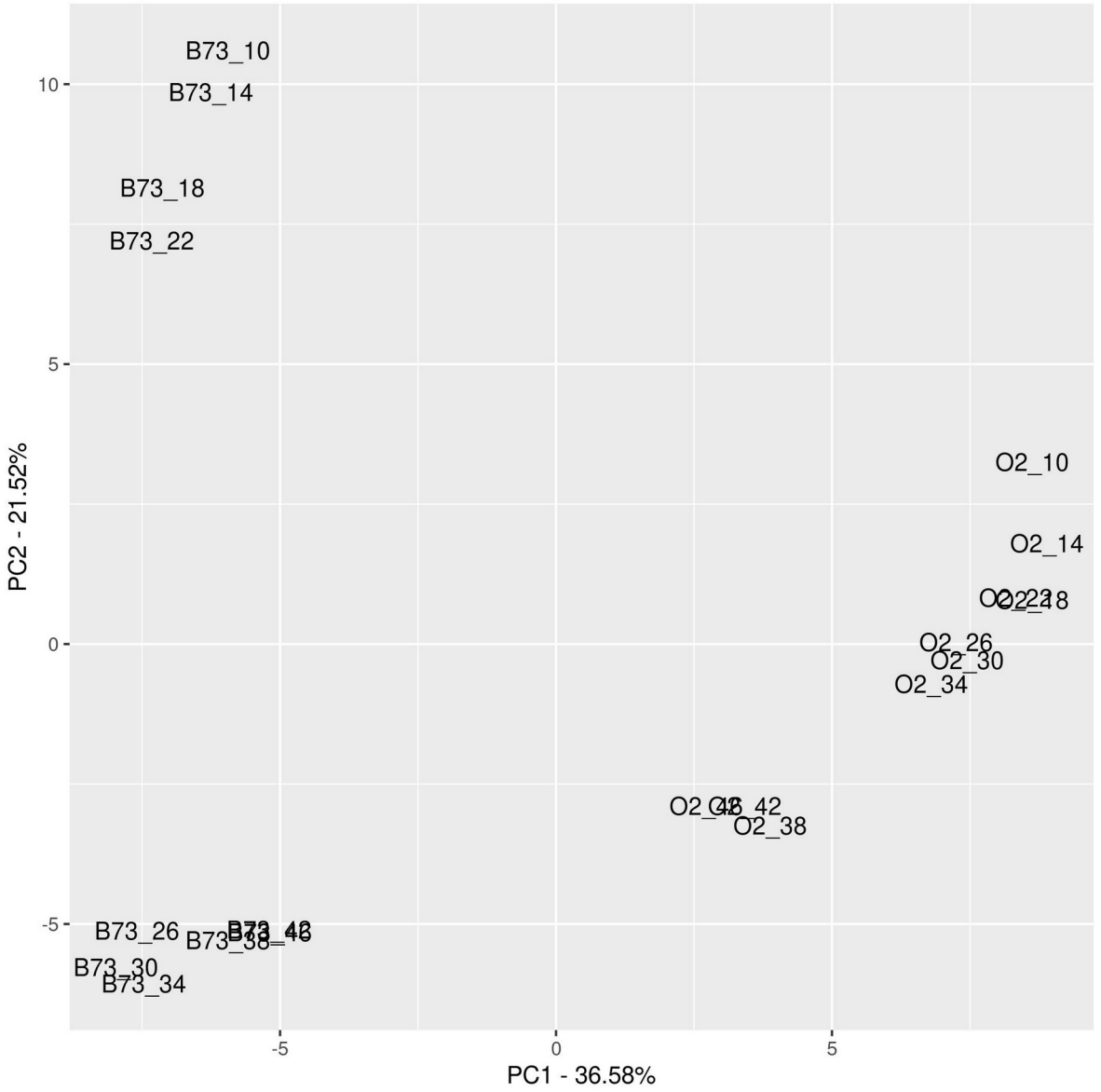
	Zm00001d001895	Zm00001d002065	Zm00001d004837	Zm00001d004855	Zm00001d004960	Zm00001d005062	Zm00001d005090
B73_10	3.5	18	5.6666666666667	216.33333333333	70.333333333333	0.5	0.5
B73_14	5.1666666666667	21.666666666667	10.333333333333	229	97.666666666667	0.5	0.5
B73_18	9	22	14.333333333333	206	137.66666666667	0.5	0.5
B73_22	11.666666666667	15	16	209.66666666667	140.66666666667	0.5	0.5
B73_26	11.666666666667	17.666666666667	11	0.5	140.66666666667	0.5	216.33333333333
B73_30	9.3333333333333	11.666666666667	10.666666666667	0.5	149.66666666667	0.5	213.33333333333
B73_34	8.3333333333333	13.666666666667	8.3333333333333	0.5	133.66666666667	0.5	179.33333333333
B73_38	8.6666666666667	10	10.333333333333	0.5	136.33333333333	0.5	0.5
B73_42	6	11	11.333333333333	0.5	111.33333333333	0.5	0.5
B73_46	6.3333333333333	13	0.5	0.5	114.66666666667	0.5	125
O2_10	7.3333333333333	8.6666666666667	0.5	0.5	0.5	0.5	217
O2_14	11.333333333333	6.3333333333333	0.5	0.5	0.5	0.5	215
O2_18	15.333333333333	5.6666666666667	0.5	0.5	0.5	0.5	212
O2_22	14.666666666667	9	0.5	0.5	0.5	0.5	201.66666666667
O2_26	17	7	0.5	0.5	118.66666666667	16.333333333333	205.66666666667
O2_30	21.333333333333	3.8333333333333	0.5	0.5	107	18.333333333333	213.66666666667
O2_34	18	4.6666666666667	0.5	0.5	101	19.666666666667	181
O2_38	17.666666666667	6	0.5	0.5	87.666666666667	26.333333333333	166.33333333333
O2_42	8.3333333333333	5.3333333333333	0.5	0.5	81.666666666667	20.333333333333	129
O2_46	6	6.6666666666667	0.5	0.5	79.333333333333	15.666666666667	89.333333333333

```
94  ###################################################
95  # Quality check
96  ###################################################
97
98  gsg = goodSamplesGenes(datExpr0, verbose = 3)
99
100 cat(rep("\n", 2))
101 print(gsg$allOK)
102
103 # If gsg$allOK is FALSE, remove genes and samples
104 if (!gsg$allOK)
105 {
106   # Optionally, print the gene and sample names that were removed:
107   if (sum(!gsg$goodGenes)>0)
108     printFlush(paste("Removing genes:", paste(names(datExpr0)[!gsg$goodGenes], collapse = ", ")))
109   if (sum(!gsg$goodSamples)>0)
110     printFlush(paste("Removing samples:", paste(rownames(datExpr0)[!gsg$goodSamples], collapse = ", ")))
111
112   # Remove the offending genes and samples from the data:
113   datExpr0 = datExpr0[gsg$goodSamples, gsg$goodGenes]
114 }
115
```

```r
##################################################
# Cluster the data
##################################################

sampleTree = hclust(dist(datExpr0), method = "average")

# Plot tree
cat(rep("\n", 2))
jpeg(filename = file.path(output_path, "sampleTree.jpeg"),  width = 1920, height = 720)
par(cex = 0.6)
par(mar = c(0,4,2,0))
plot(sampleTree, main = "Sample clustering to detect outliers", sub="", xlab="", cex.lab = 1.5, cex.axis = 1.5, cex.main = 2)
# abline(h = 15, col = "red")
dev.off()

# # Determine cluster under the line
# clust = cutreeStatic(sampleTree, cutHeight = 15.2, minSize = 10)
#
# print(table(clust))
#
# # Remove outlier sample
# keepSamples = (clust==1)
# datExpr = datExpr0[keepSamples, ]

datExpr = datExpr0
nGenes = ncol(datExpr)
nSamples = nrow(datExpr)

# Collect garbage
collectGarbage()

```


Sample clustering to detect outliers
Height
0
200
400
600
800
1000
B73_18
B73_22
B73_34
B73_26
B73_30
O2_10
O2_14
O2_18
O2_22
O2_34
O2_26
O2_30
B73_10
B73_14
O2_46
O2_38
O2_42
B73_38
B73_42
B73_46

```r
################################################################################
## Performs principal components analysis (PCA)
################################################################################
pca_df <- datExpr
pca <- prcomp(pca_df, center = TRUE, scale. = TRUE)

pca_variance <- pca$sdev^2
pca_variance_percentage <- round(pca_variance/sum(pca_variance, na.rm = TRUE)*100, digits=2)

x <- pca$x %>% as.data.frame(stringsAsFactors=TRUE)

## Plot PC1 vs PC2 with GGplot2
p <- ggplot(x, aes(x=PC1, y=PC2)) +
  geom_text(label=rownames(x)) +
  labs(
    x=paste0("PC1 - ", pca_variance_percentage[1], "%"),
    y=paste0("PC2 - ", pca_variance_percentage[2], "%")
  )

ggsave(
  filename="PCA_Plot.jpg",
  plot=p,
  path=output_path
)
```


B73_10
B73_14
B73_18
B73_22
O2_10
O2_14
O2_26
O2_30
O2_34
O2_38
B73_26
B73_30
B73_34
10
5
0
-5
-5
0
5
PC2 - 21.52%
PC1 - 36.58%

```
175  ##################################################
176  # Save processed data
177  ##################################################
178  write.csv(
179    x = datExpr,
180    file = file.path(output_path, "..", "datExpr.csv"),
181    na = "",
182    quote = FALSE
183  )
184
```

Original Phenotype Input File

Genotype	DAP	Replicaiton	Ala	Arg	Asx	Glx	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser
B73	10	1	168.34	26.01	147.24	118.77	185.03	25.52	52.02	65.28	49.08	26.01	25.52	93.25	82.45
B73	10	2	182.93	25.85	144.39	120.00	180.98	24.88	53.17	64.39	48.78	24.39	26.34	92.20	80.98
B73	10	3	190.85	26.14	131.24	120.78	191.90	27.71	49.15	61.18	50.72	24.05	26.67	88.37	84.18
B73	14	1	120.34	20.06	82.96	89.34	155.44	25.53	30.54	41.48	40.57	14.13	20.97	59.26	53.33
B73	14	2	140.35	23.86	94.04	89.82	172.63	24.80	48.19	66.90	44.91	18.71	25.26	56.61	65.96
B73	14	3	130.87	22.70	87.00	74.14	137.30	19.29	49.93	73.38	38.20	17.78	21.18	48.42	65.44
B73	18	1	126.78	23.86	84.21	85.15	161.87	24.33	58.95	94.97	43.51	18.71	26.20	60.82	66.43
B73	18	2	116.69	24.97	81.53	93.25	171.21	27.01	61.66	101.91	44.84	20.89	28.03	66.24	68.28
B73	18	3	114.58	23.74	74.32	88.26	160.52	27.87	54.19	90.32	43.35	17.03	27.87	59.87	60.90
B73	22	1	153.83	24.68	89.04	95.21	163.53	21.60	71.85	132.67	40.99	21.60	25.56	78.02	71.85
B73	22	2	135.80	20.23	80.90	82.35	132.91	16.98	69.71	123.88	33.95	22.39	21.31	68.98	65.73
B73	22	3	128.52	21.99	78.92	87.55	150.51	21.99	62.10	116.87	39.25	18.98	24.58	69.43	62.96
B73	26	1	135.53	23.53	85.18	93.18	162.82	24.47	60.71	106.82	38.59	17.88	26.35	66.35	62.59
B73	26	2	144.78	26.28	94.06	102.36	165.07	22.59	75.16	136.48	41.50	24.44	26.74	82.54	79.77
B73	26	3	199.47	29.47	107.89	117.37	193.68	26.32	95.79	168.42	45.79	25.26	31.05	96.32	87.37
B73	30	1	167.27	28.05	89.35	107.01	183.38	26.49	78.96	148.57	42.60	20.26	29.61	84.16	75.84
B73	30	2	143.66	25.13	80.42	88.38	148.27	20.52	71.62	121.88	38.53	20.94	23.87	68.27	69.53

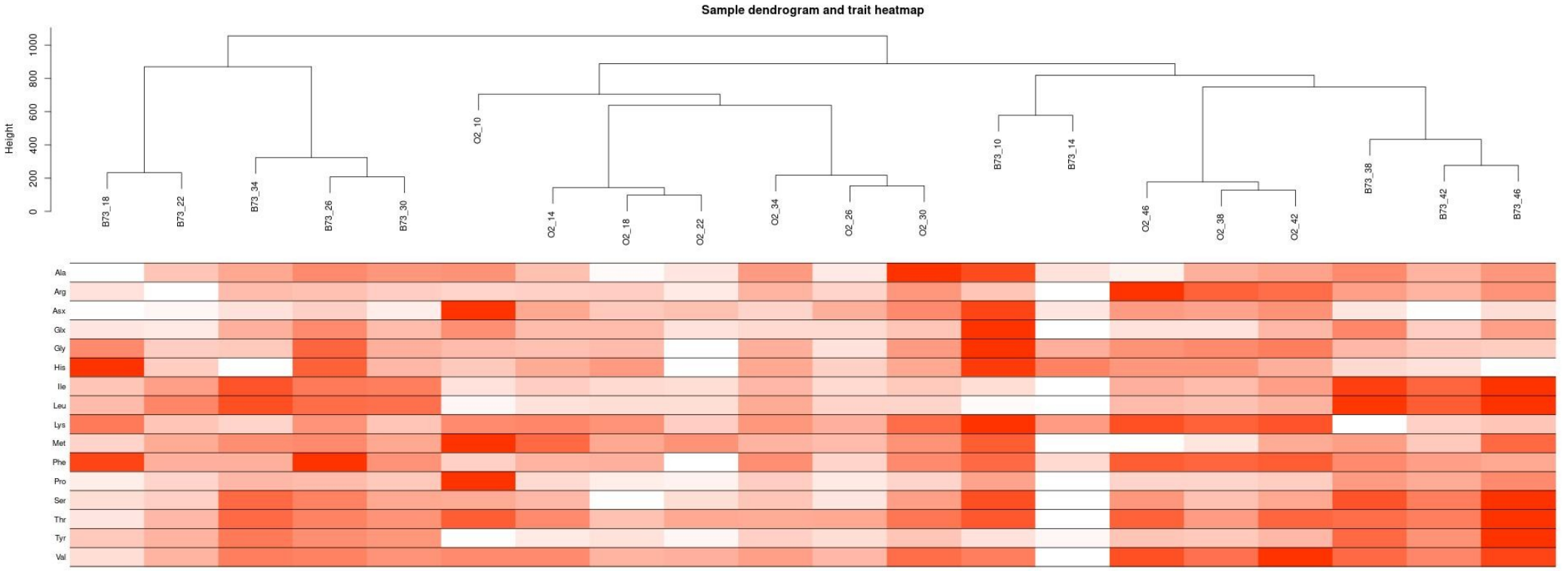
```
186 ##################################################
187 # Process the trait data
188 ##################################################
189
190 datTraits[datTraits[,1] == "o2", 1] <- "O2"
191
192 datTraits = datTraits %>%
193   pivot_longer(!c(Genotype, DAP, Replicaiton), names_to = "Trait", values_to = "Measurement") %>%
194   group_by(Genotype, DAP, Trait) %>%
195   summarise_at(vars(Measurement), list(Measurement = mean)) %>%
196   ungroup() %>%
197   unite("Genotype", Genotype:DAP, sep = "_") %>%
198   pivot_wider(names_from = Trait, values_from = Measurement, values_fill = NA) %>%
199   as.data.frame(stringsAsFactors = FALSE)
200
201 # Copy values in column 1 into row names
202 rownames(datTraits) <- datTraits[,1]
203
204 # Remove column 1
205 datTraits <- datTraits[,-1]
206
207 print(head(datTraits))
208 print(tail(datTraits))
209 print(dim(datTraits))
210
```

Processed Phenotype Input File

	Ala	Arg	Asx	Glx	Gly	His	I
B73_10	180.70666666667	26	140.95666666667	119.85	185.97	26.036666666667	
B73_14	130.52	22.206666666667	88	84.433333333333	155.12333333333	23.206666666667	
B73_18	119.35	24.19	80.02	88.886666666667	164.53333333333	26.403333333333	
B73_22	139.38333333333	22.3	82.953333333333	88.37	148.98333333333	20.19	
B73_26	159.92666666667	26.426666666667	95.71	104.30333333333	173.85666666667	24.46	
B73_30	155.11333333333	25.53	85.236666666667	95.873333333333	155.26666666667	21.026666666667	
B73_34	149.81666666667	26.46	89.566666666667	98.043333333333	149.23333333333	18.22	
B73_38	159.05666666667	28.43	87.906666666667	105.13	153.66	19.693333333333	
B73_42	145.66	27.23	79.52	93.22	149.85	19.393333333333	
B73_46	155.34	29.193333333333	90.473333333333	100.91666666667	148.21666666667	18.066666666667	
O2_10	156.9	25.113333333333	147.76666666667	104.1	153.05333333333	20.323333333333	
O2_14	140.63666666667	25.153333333333	108.94	95.906666666667	152.01333333333	21.476666666667	
O2_18	121.94666666667	25.52	98.116666666667	96.206666666667	152.60333333333	22.133333333333	
O2_22	129.03333333333	23.76	100.72	89.433333333333	136.11	18.186666666667	
O2_26	126.86	25.126666666667	106.40666666667	90.983333333333	143.65333333333	19.946666666667	
O2_30	189.93	29.07	118.04666666667	94.48	160.17333333333	21.696666666667	
O2_34	153.55333333333	27.106666666667	94.406666666667	90.926666666667	155.99333333333	21.653333333333	
O2_38	146.48	32.313333333333	110.6	89.653333333333	164.26	22.39	
O2_42	151.67666666667	31.7	115.13333333333	96.94	167.67	21.38	
O2_46	124.71333333333	35.483333333333	112.5	89.816666666667	162.35	22.246666666667	

```r
##################################################
# Save processed trait data
##################################################
write.csv(
  x = datTraits,
  file = file.path(output_path, "..", "datTraits.csv"),
  na = "",
  quote = FALSE
)
```

```r
#################################################
# Construct heatmap to visualize trait data
#################################################
# Re-cluster samples
sampleTree2 = hclust(dist(datExpr), method = "average")

# Convert traits to a color representation: white means low, red means high, grey means missing entry
traitColors = numbers2colors(datTraits, signed = FALSE)

cat(rep("\n", 2))
jpeg(filename = file.path(output_path, "tree_trait_heatmap.jpeg"), width = 1920, height = 720)
# Plot the sample dendrogram and the colors underneath.
plotDendroAndColors(sampleTree2, traitColors, groupLabels = names(datTraits), main = "Sample dendrogram and trait heatmap")
dev.off()
```


Sample dendrogram and trait heatmap
Height
0
200
400
600
800
1000
B73_18
B73_22
B73_34
B73_26
B73_30
O2_10
O2_14
O2_18
O2_22
O2_34
O2_26
O2_30
B73_10
B73_14
O2_46
O2_38
O2_42
B73_38
B73_42
B73_46
Ala
Arg
Asx
Glx
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Tyr
Val