Investigating Differential Gene Expression Patterns in Synthetic *Arabidopsis halleri x lyrata* Hybrids

STA426 Course Project
Elina Jansone
13.01.2023

1 Introduction

Polyploidization, also known as whole genome duplication, is a ubiquitous process in the evolutionary history of plants [1]. The reasons behind its success are still unclear, particularly the short-term survival of a newly formed polyploid species [2].

To address this question, researchers focused on genomic changes, with most studies looking at transcriptomic changes [3] [4] [5] [6] [7]. These changes have been explored from two different points of view: one looking at the bias in homoeolog expression between subgenomes and one looking at the expression changes between polyploid and their parent species [8]. When comparing polyploid and parent species, expression differences depended on the species but had relatively similar levels [9]. More specifically, most of the expression changes were attributed to hybridization rather than polyploidization per se [10]. This finding highlighted by the Spartina study system was based on microarray technology and focused on expression deviations from a mid-parent value (MPV). With the development of high- throughput sequencing technologies and polyploid-specific tools, it's now possible to quantify expression in a more accurate and comprehensive way [11].

1.1 Sample Background

A hybrid diploid line *A. halleri x lyrata* was generated by crossing *A. halleri* subsp. *gemmifera* (maternal) and *A. lyrata* subsp. *petraea* (paternal). Four biological replicates for the hybrid and parent species were incubated in a climate chamber set at 22 C, with 60% relative humidity and 16 hours light/8 hours dark cycles. These conditions should be mild for the species involved, meaning that no strong stress response is expected. Leaf tissue was collected from each individual when the individual started flowering, and the tissue was used for RNA extraction with RNeasy (QIAGEN). [14]

RNA libraries were synthesized using Illumina TruSeq stranded mRNA kit and all libraries were sequenced by Illumina NovaSeq 6000 in 150bp paired-end mode.

1.2 Goal of the analysis

To reassess the effect of hybridization in the polyploidization process and capitalize on recent methodological (and technological) advances, the goal of this project is to analyze gene expression changes in *Arabidopsis halleri x lyrata* hybrids with respect to their parent species *A. halleri* and *A. lyrata*. These *A. halleri x lyrata* hybrids represent precursors of the natural species *A. kamchatica* , a successfully established polyploid [12] [13]. Insights into the expression changes in hybrids could contribute towards understanding the expression patterns of the polyploid species, with some genes or gene functions potentially playing a central role in the survival and establishment of the polyploid species.

1.3 Sample preprocessing using command-line tools

Sample QC was done using fastqc and summarized with multiqc. 16 alignments were generated from FASTQ files aligned to 2 assemblies using STAR: 4 *A. halleri* samples and 4 *A. halleri x lyrata* samples aligned to the in-house chromosome-level *A. halleri* assembly [15], 4 *A. lyrata* samples and 4 *A. halleri x lyrata* samples aligned to the in-house chromosome-level *A. lyrata assembly* [15]. 16 feature count tables were obtained using 'featureCounts' from the 'subread' tool.

2 Statistical Analysis

```
library(edgeR)
library(tidyverse)
library(DESeq2)
library(topGO)
library(gghighlight)
library(ComplexHeatmap)
```

Statistical analysis of the samples is done using the libraries edgeR [16] and DESEq2 [17]. Heatmaps are generated using the ComplexHeatmap package [18] Supplementary to the

visualization with ggplot2 (part of tidyverse) [19], additional packages gghighlight [20] and ggrepel [21] were used.

2.1 Sample filtering and normalization

```
DGEList_hal_hyb <- read_rds('data_halleri_hybrid.rds')
DGEList_lyr_hyb <- read_rds('data_lyrata_hybrid.rds')

data_filt_halleri_hybrid <- calcNormFactors(
                            data_filt_halleri_hybrid)
names(data_filt_halleri_hybrid$samples$norm.factors) <-
c("Hal_1","Hal_2","Hal_3", "Hal_4", "Hyb_Hal_1",
"Hyb_Hal_2", "Hyb_Hal_3", "Hyb_Hal_4")

data_filt_lyrata_hybrid <- calcNormFactors(
                            data_filt_lyrata_hybrid)
names(data_filt_lyrata_hybrid$samples$norm.factors) <-
c("Lyr_1","Lyr_2","Lyr_3", "Lyr_4",
"Hyb_Lyr_1", "Hyb_Lyr_2", "Hyb_Lyr_3", "Hyb_Lyr_4")

# Normalization factors for halleri, hybrid samples
data_filt_halleri_hybrid$samples$norm.factors
# Normalization factors for lyrata, hybrid samples
data_filt_lyrata_hybrid$samples$norm.factors
```

NORMALIZATION FACTORS HERE AS TABLE

2.2 Analysis of Sample Biological Variance

```
data_filt_halleri_hybrid <- estimateDisp(
                        data_filt_halleri_hybrid)
data_filt_lyrata_hybrid <- estimateDisp(
                        data_filt_lyrata_hybrid)

plotBCV(data_filt_halleri_hybrid, main =
"BCV plot of samples, A. halleri v. hybrid",
                        yaxp = c(0.2,1.2,10))
plotBCV(data_filt_lyrata_hybrid, main =
"BCV plot of samples, A. lyrata v. hybrid",
                        yaxp = c(0.2,1.2,10))
```

The following code produces Figure 1. Biological coefficient of variation (BCV) describes how the (unknown) true abundance of the gene varies between replicate transcriptomes. It assumes indefinitely high sequencing depth and is a representation of natural variance. The common dispersion among samples in the A. halleri v. hybrid dataset is slightly under 0.2 with majority of genes being located in the 0.1 - 0.25 region. In the A. lyrata v. hybrid dataset, the common dispersion is 0.24 with a similar gene-densest region. Overall, the BCV plot indicates reasonable dispersion within the dataset to be able to identify the DEGs.

Next, it would be useful to investigate whether there are significant differences between individual samples in each dataset.

```
par(mfrow = c(1,2))
```

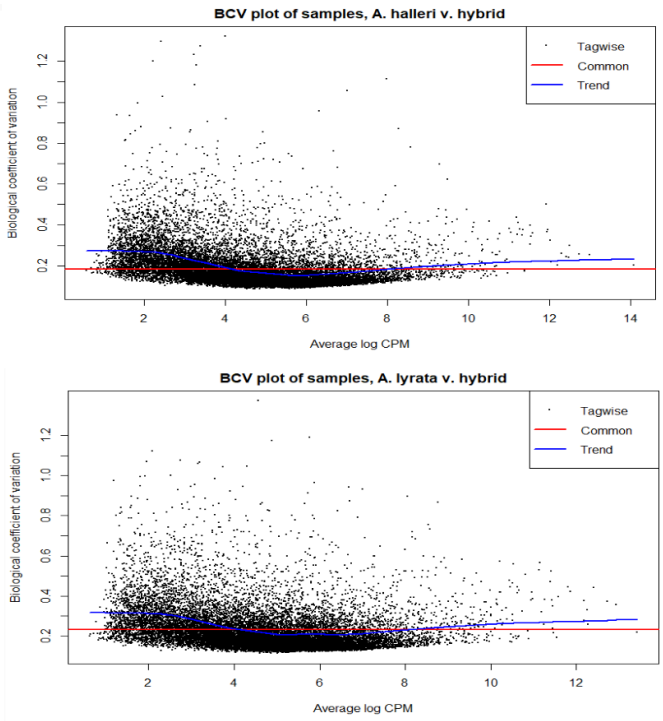

Figure 1: Biological coefficient of variation of both datasets. The trendlines as well as the point density suggests good quality of samples.

```
plotMDS(data_filt_halleri_hybrid, labels =
        data_filt_halleri_hybrid$samples$samples,
        method = "logFC", cex = 0.5,
        gene.selection = "common", main = "Sample
        expresssion MDS, A. halleri v. hybrid")
plotMDS(data_filt_lyrata_hybrid, labels =
        data_filt_lyrata_hybrid$samples$samples,
        method = "logFC", cex = 0.5,
        gene.selection = "common", main = "Sample
        expresssion MDS, A. lyrata v. hybrid")
```

The above output (2) shows the multivariate dissimilarity scaling of samples contained in both datasets. There is a notable separation of parent species and the hybrid organisms. Although 90% of expression variability lies in the difference between expression levels of the parent and the hybrid, there is a clear grouping of replicate pairs of both parent species, while hybrid samples are grouped much closer together. That suggests a prominent batch effect that is present among the parent species samples.

It would be therefore reasonable to investigate expression differences between A. halleri samples 1+2 v. 3+4 and A. lyrata samples 1+2 v. 3+4 in addition to the primary analysis goal.

```
logcpm_halleri_hybrid <- cpm(
        data_filt_halleri_hybrid, log=TRUE)
```

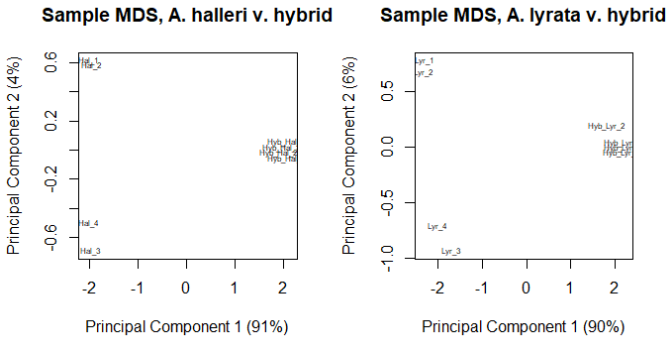

Figure 2: MDS of samples for both datasets. We see that, over both datasets, around 90% of the variability stems from the differences between parent and hybrid. However, there is a notable component separating pairs of parental samples from each other. The hybrid samples all group together on the right side, while the parental samples spread along the vertical axis on the left side.

```
logcpm_lyrata_hybrid <- cpm(
        data_filt_lyrata_hybrid, log=TRUE)

colnames(logcpm_halleri_hybrid) <-
        c("Hal_1","Hal_2","Hal_3", "Hal_4",
        "Hyb_Hal_1", "Hyb_Hal_2",
        "Hyb_Hal_3", "Hyb_Hal_4")
colnames(logcpm_lyrata_hybrid) <-
        c("Lyr_1","Lyr_2","Lyr_3", "Lyr_4",
        "Hyb_Lyr_1", "Hyb_Lyr_2",
        "Hyb_Lyr_3", "Hyb_Lyr_4")

clust_halleri_hybrid <- hclust(dist(t(
        logcpm_halleri_hybrid)), method = "average")
clust_lyrata_hybrid <- hclust(dist(t(
        logcpm_lyrata_hybrid)), method = "average")

par(mfrow = c(1,2))
plot(clust_halleri_hybrid, main =
    "Halleri v. Hybrid sample clustering")
plot(clust_lyrata_hybrid, main =
    "Lyrata v. Hybrid sample clustering)
```

Dendrograms show that A. halleri sample 1,2 and 3,4 clusters form early, but overall the samples converge much later than hybrid samples do (Figure 3). The fact that the first 2 convergences form the two distinct parental groups corresponds with the message of sample MDS plots for both datasets.

For A. lyrata samples, the distances between individual samples even in their closest pairing are higher than the distance within which all hybrid samples converge.

Interestingly, the distance at which all A. lyrata samples converge into a cluster is higher than the respective distance for A. halleri samples. The same difference in distance is also seen between the A. lyrata v. hybrid and A. halleri v. hybrid

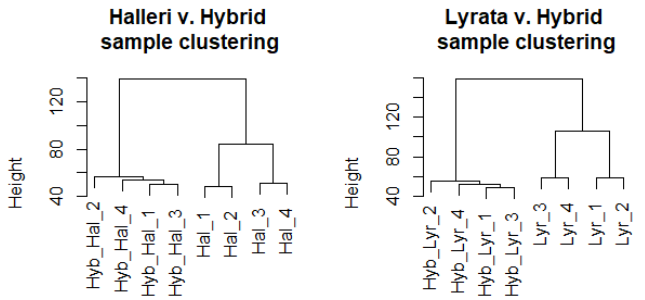

Figure 3: Dendrograms of both datasets paint a similar picture to the MDS plots - here we see that the distance between individual *A. lyrata* samples is greater than that between the hybrids or between the *A. halleri* samples.

datasets. Might it be that A. lyrata shares less transcriptomic similarity to the hybrid than the A. halleri samples do because of sharing less transcription factors with the hybrid? Or do certain genes of A. halleri dominate over those of A. lyrata?

```
Heatmap(logcpm_halleri_hybrid[,],
show_row_dend = FALSE, show_row_names = FALSE)
Heatmap(logcpm_lyrata_hybrid[,],
show_row_dend = FALSE, show_row_names = FALSE)
```

Heatmaps (Figure 4) of both datasets provide an overview of how transcript counts differ between parental and hybrid samples for each gene. In each of the heatmaps above we observe the 2 separated parental sample clusters for each species which show a distinct common gene expression pattern. For some genes, the difference in average logcpm between the different parental clusters is between 2 and 5. These differences show up more frequently among genes with lower logcpm.

Next, the identities and functions of top DEGs will be investigated.

3 Explorative Analysis

3.1 Differential Expression Analysis

3.1.1 Hybrid v. parental species

It would be difficult to compare differentially expressed genes between two different species which have different gene names. To make the following DGE plots more comparable, we can substitute the *A. halleri* and *A. lyrata* gene names with their reciprocal homologs in *A. thaliana.*

Performing an exact test on all genes of each datasets:

```
test_halleri_hybrid <-
        exactTest(data_filt_halleri_hybrid)
test_lyrata_hybrid <-
        exactTest(data_filt_lyrata_hybrid)
```

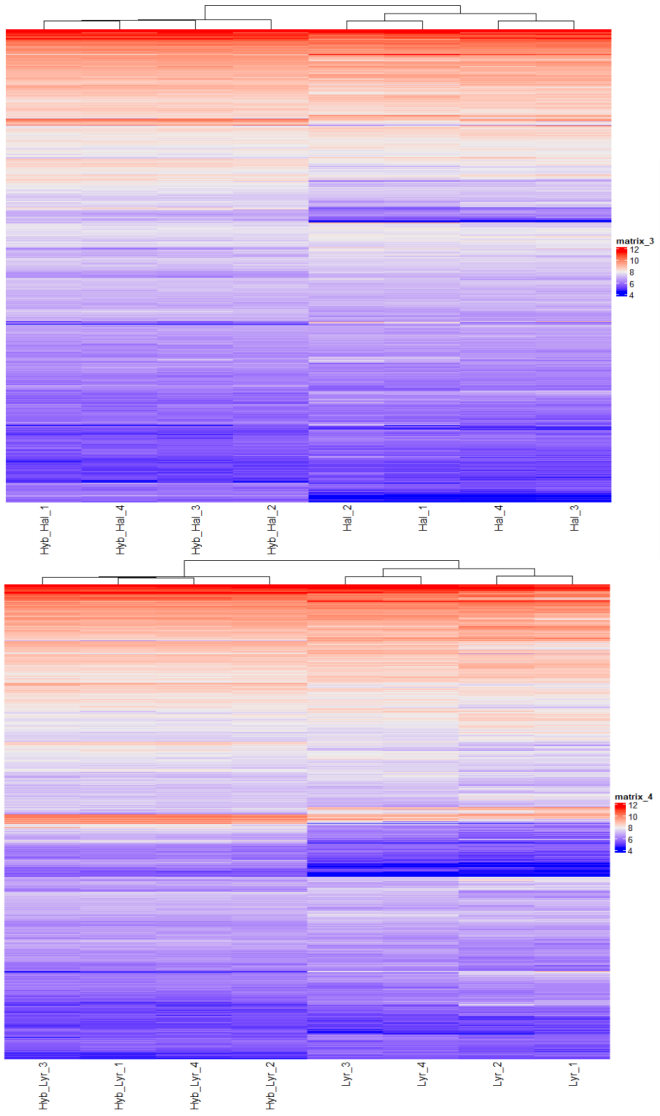

Figure 4: Heatmaps of variance in gene expression, in the form of logCPM. Heatmap above is for the *A. halleri* - *A. halleri x lyrata* comparison, while below is *A. lyrata v. A. halleri x lyrata*. Here, especially in the bottom heatmap, we see the pairings of the parental samples and the similarity of logCPM values that the grouping is based on.

```
all_tags_hh <- topTags(
        test_halleri_hybrid, n = 17167)
all_tags_lh <- topTags(
        test_lyrata_hybrid, n = 16965)
```

Next, the reciprocal BLAST hit files are read in for *A. halleri* and *A. thaliana* as well as *A. lyrata* and *A. thaliana*.

```
genenames_ahal_atha <- data.table::fread(
        "rbh_Ahal_Athal_221215", header = FALSE)
genenames_alyr_atha <- data.table::fread(
        "rbh_Alyr_Athal_221215", header = FALSE)

all_tags_hh_rename <- all_tags_hh[all_tags_hh$
        table$geneID %in% genenames_ahal_atha$V1,]
all_tags_lh_rename <- all_tags_lh[all_tags_lh$
        table$geneID %in% genenames_alyr_atha$V1,]

for (i in c(1:length(all_tags_lh_rename$
                     table$geneID))) {
  if (i <= length(all_tags_hh_rename$
                  table$geneID)) {

    all_tags_hh_rename$table$geneID[i] <-
      genenames_ahal_atha[which(
      genenames_ahal_atha$V1 ==
      all_tags_hh_rename$table$geneID[i])]$V2
  }

  all_tags_lh_rename$table$geneID[i] <-
    genenames_alyr_atha[which(
    genenames_alyr_atha$V1 ==
    all_tags_lh_rename$table$geneID[i])]$V2
}
```

Highlighting the left and right tail of the volcano plot and plotting all genes:

```
n_genes_sign_hh_rename <- all_tags_hh_rename$
table[-log10(all_tags_hh_rename$table$FDR) > 2 &
((abs(all_tags_hh_rename$table$logFC) >= 5 |
-log10(all_tags_hh_rename$table$FDR) > 50) |
all_tags_hh_rename$table$logFC < -3.7),]

n_genes_sign_lh_rename <- all_tags_lh_rename$
table[-log10(all_tags_lh_rename$table$FDR) > 50 |
((abs(all_tags_lh_rename$table$logFC) >= 5 |
-log10(all_tags_lh_rename$table$FDR) > 50) |
all_tags_lh_rename$table$logFC < -3.7),]

ggplot(data = all_tags_hh_rename$table,
        aes(y = -log10(FDR), x = logFC)) +
  geom_point(size = 0.5) +
  geom_text(aes(label = geneID),
            check_overlap = TRUE, size = 1.5,
            nudge_y = 3) +
  gghighlight(-log10(FDR) > 2 & ((abs(logFC) >= 5 |
  -log10(FDR) > 50) | logFC < -3.7 | logFC > 7),
            unhighlighted_params = aes(colour = 'grey'))
```

Figure 5: Volcano plots of differentially expressed genes in the comparison between the paternal *A. lyrata* and the hybrid *A. halleri x lyrata* (above) as well as the comparison of the maternal *A. halleri* and the hybrid (below). Both plots are slightly more crowded on the right side.

```
ggplot(data = all_tags_lh_rename$table,
       aes(y = -log10(FDR), x = logFC)) +
  geom_point(size = 0.5) +
  geom_text(aes(label = geneID),
            check_overlap = TRUE, size = 1.5,
            nudge_y = 3) +
  gghighlight(-log10(FDR) > 50 | ((abs(logFC) >= 5 |
  -log10(FDR) > 50) | logFC < -3.7 | logFC > 5),
            unhighlighted_params = aes(colour = 'grey'))
```

As seen in Figure 5, Genes appearing on the right side are upregulated in the hybrid, while those on the left side are expressed more in the compared parent species.

It is interesting to note that most DEGs show higher expression in the hybrid than the respective parent. We can now take only the right end of the plot and see where the respective genes are located in the other parental comparison dataset. However, it is to note that some genes found in one dataset might not be found in the other, as the reciprocal BLAST hit might have been found for one species' gene but not for the other's.

For that, we first filter for only positive logFC values, followed by highlighting points that are present among the positively regulated DEGs in the dataset of the other parent.

```
hh_rename_high <- n_genes_sign_hh_rename[
        n_genes_sign_hh_rename$logFC > 0,]

lh_rename_high <- n_genes_sign_lh_rename[
        n_genes_sign_lh_rename$logFC > 0,]

ggplot(data = all_tags_lh_rename$table,
        aes(y = -log10(FDR), x = logFC)) +
        geom_point(size = 0.5) +
        geom_text(aes(label = geneID),
                check_overlap = TRUE, size = 2,
                nudge_y = 3) +
        gghighlight(geneID %in% hh_rename_high$geneID) +
        labs(title = "Expression changes in A. halleri
        x lyrata hybrid compared to A. lyrata",
        subtitle = "Highlighted points from highly
        expressed hybrid genes compared to A. halleri")

ggplot(data = all_tags_hh_rename$table,
        aes(y = -log10(FDR), x = logFC)) +
        geom_point(size = 0.5) +
        geom_text(aes(label = geneID),
        check_overlap = TRUE, size = 2, nudge_y = 3) +
        gghighlight(geneID %in% lh_rename_high$geneID) +
        labs(title = "Expression changes in A. halleri
        x lyrata hybrid compared to A. halleri",
        subtitle = "Highlighted points from highly
        expressed hybrid genes compared to A. lyrata")
```

As we see in Figure 6, the datapoints of genes that were upregulated in the hybrid in one hybrid-parent comparison appear, on average, left of center of the plot of the other parent's comparison with the hybrid. That means that the genes of the hybrid are only highly expressed in comparison to one parent, while they seem to be "insufficient" when compared to the other parent. This, together with the fact that most genes overall seem to be upregulated, is a notable finding. However, the reasons for such a process can't be inferred based on this analysis.

3.1.2 Variance within parental samples

As we observed in Section 2.2, parental samples showed large variance. Now the actual significance of this variance will be investigated. The 2 subsets of 4 parental samples each are re-scaled and an exact test if performed to see which genes are significantly up- or downregulated.

```
parental_df_test <- function(df) {
   parental_df <- df[,1:4]
   parental_df$samples$group <- c(1,1,2,2)
   parental_df <- calcNormFactors(parental_df)
```

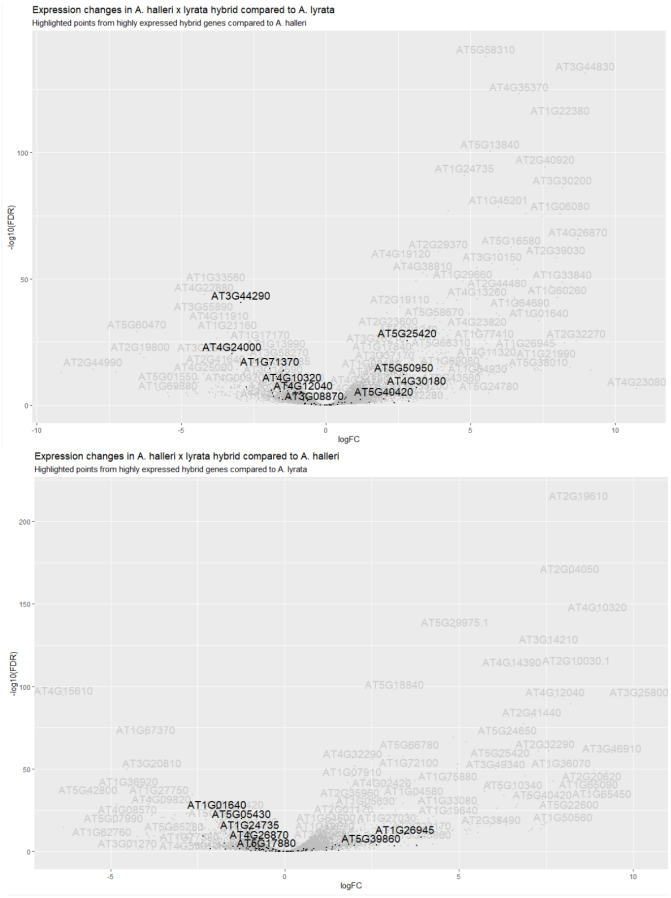

Figure 6: Highlighted locations of upregulated maternal *A. halleri* genes in the *A. lyrata- A. halleri x lyrata* volcano plot (above) as well as the locations of upregulated paternal *A. lyrata* genes in the *A. halleri-A. halleri x lyrata* volcano plot (below). Despite some loss due to absent gene translations, we can observe a general trend that highly expressed genes compared to one parent appear more on the negative side of the other parent's volcano plot.

geneID <chr>	logFC <dbl>	FDR <dbl>
AhW302_G006445	-10.290883	0.2146946
AhW302_G011433	-7.560099	0.2146946
AhW302_G029297	-7.500559	0.2146946

geneID <chr>	logFC <dbl>	FDR <dbl>
AhW302_G007492	4.604727	0.2146946
AhW302_G015821	4.857046	0.2146946
AhW302_G005608	7.656672	0.2146946

geneID <chr>	logFC <dbl>	FDR <dbl>
AlMed_G005994	-8.534536	0.1460684
AlMed_G009562	-8.122565	0.1460684
AlMed_G013737	-7.988319	0.1460684

geneID <chr>	logFC <dbl>	FDR <dbl>
AlMed_G014782	7.129237	0.1460684
AlMed_G006617	7.242574	0.1460684
AlMed_G005656	9.298266	0.1460684

Figure 7: Highest and lowest logFC values among the parental sample pairs for *A. halleri* (AhW...) and *A. lyrata* (AlMed...). The logFC values are very high, however, do not reach statistical significance with respect to multiple testing.

```
   parental_df <- exactTest(parental_df)
   parental_df <- topTags(parental_df, n =
                length(rownames(parental_df$table)),
                sort.by = c("logFC"))
   parental_df <- parental_df[order(as.data.frame(
                  parental_df)$logFC),]
   return(parental_df)
}

halleri_test <- parental_df_test(
                data_filt_halleri_hybrid)
lyrata_test <- parental_df_test(
                data_filt_lyrata_hybrid)

head(halleri_test[,c("geneID","logFC","FDR")], 3)
head(lyrata_test[,c("geneID","logFC","FDR")], 3)
```

The tables in Figure 7 show that, although with very high logFC values, the FDR of the intra-parental comparisons does not reach significant values. Therefore, no further analysis will be carried out, although it is worth expanding on this part of the investigation later on.

3.2 Gene Set Enrichment Analysis

Instead of looking at individual genes, it is more useful to look at patterns of pathway expression as it shows which biological activities are more prevalent in a sampled organism. For GSEA, the TopGO package was used [22]. First, a general function is made to carry out GSEA and for plotting the results:

```
# Import custom annotation for thaliana
# Create gene universe set
TAIRGO <- readMappings("GOslim.txt")
geneNames_all <- names(TAIRGO)

GSEA_analysis <- function(gene_sign_thaliana,
                          gene_names_all) {

   geneList <- factor(as.integer(gene_names_all
   %in% gene_sign_thaliana$geneID))
   names(geneList) <- gene_names_all
   str(geneList)

   #---------- GO enrichment analysis ------------
   #create go object
   GO_enrichment <- new("topGOdata", ontology = "BP",
                    allGenes = geneList,
                    annot = annFUN.gene2GO,
                    gene2GO = TAIRGO)

   # overrepresentation test
   BP_elim <- runTest(GO_enrichment, algorithm =
                        "elim", statistic = "fisher")

   # generate summary of the results
   table_genes <- GenTable(object = GO_enrichment,
                        elim = BP_elim,
                        topNodes = length(BP_elim@score))

   # select significantly enriched (elim < 0.05)
   # and annotated genes between 10-999

   table_reduced <- droplevels(subset(table_genes,
                            table_genes$Annotated>=10 &
                            table_genes$Annotated<1000 &
                            table_genes$elim<0.05))
   return(table_reduced)
}

# PLOT GSEA

GSEA_plot <- function(GSEA_table, set_color = "red") {
   ggplot(data = GSEA_table,
      aes(x = Significant/Expected, y =
            -log10(as.numeric(elim)))) +
      geom_point(aes(size = Significant),
            col = set_color, alpha = 0.5, shape = 16) +
      theme_bw() +
      theme(axis.title = element_text(size = 15),
            axis.text = element_text(size = 15)) +
      scale_size(range = c(1,20),
                 breaks = c(10, 100, 200, 300),
                 limits = c(0, 300)) +

      ggrepel::geom_text_repel(data = GSEA_table,
      aes(label = Term), box.padding = 1,
                       size = 3, max.iter = 1000000,
                       overlaps = 20, col = "blue")
}
```

For GSEA we select a FDR bottom cutoff of 10 and a logFC absolute value bottom cutoff at 1. This is to prevent the formation of many different gene set clusters that have similar functions, as well as to prevent the incorporation of gene set of interests into larger, less specific pathway clusters. Separate GSEA runs were done on the up- and downregulated significant DEGs of both parental v. hybrid comparison datasets.

```
genes_sign_hh <- all_tags_hh_rename$table[(
    -log10(all_tags_hh_rename$table$FDR) >= 10) &
    abs(all_tags_hh_rename$table$logFC) >= 1,]
genes_sign_lh <- all_tags_lh_rename$table[(
    -log10(all_tags_lh_rename$table$FDR) >= 10) &
    abs(all_tags_lh_rename$table$logFC) >= 1,]
```

Now we run GSEA and obtain the GSEA plots, which are shown in Figures 8 and 9.

```
genes_sign_hh_up <- genes_sign_hh[
                    genes_sign_hh$logFC > 0,]
genes_sign_hh_down <- genes_sign_hh[
                    genes_sign_hh$logFC < 0,]

genes_sign_lh_up <- genes_sign_lh[
                    genes_sign_lh$logFC > 0,]
genes_sign_lh_down <- genes_sign_lh[
                    genes_sign_lh$logFC < 0,]

GSEA_hh_up_table <- GSEA_analysis(
                    genes_sign_hh_up,geneNames_all)
GSEA_hh_down_table <- GSEA_analysis(
                    genes_sign_hh_down,geneNames_all)
GSEA_lh_up_table <- GSEA_analysis(
                    genes_sign_lh_up,geneNames_all)
GSEA_lh_down_table <- GSEA_analysis(
                    genes_sign_lh_down,geneNames_all)

GSEA_plot(GSEA_hh_up_table, tit = "Upregulated
           pathways in A. halleri x lyrata hybrid",
           sub = "compared to expression in A. halleri")
GSEA_plot(GSEA_hh_down_table, set_color = "blue",
           tit = "Downregulated pathways in A. halleri
           x lyrata hybrid", sub = "compared to
           expression in A. halleri")
GSEA_plot(GSEA_lh_up_table, tit = "Upregulated
           pathways in A. halleri x lyrata hybrid",
           sub = "compared to expression in A. lyrata")
```

```
GSEA_plot(GSEA_lh_down_table, set_color = "blue",
            tit = "Downregulated pathways in A. halleri
            x lyrata hybrid", sub = "compared to
            expression in A. lyrata")
```

4 Discussion

4.1 Biological Significance

The volcano plots of the hybrid compared with both parental species shows much more upregulation than downregulation. It is unclear what the reason for such behavior is, however, it is established that genes of the hybrid that are upregulated in comparison to one parent are not found among the significantly upregulated points in the comparison with the other parent.

There are (at least) two potential covariables here - one being the role of the gene (sex-specificity) and another one being the "loss in translation" for around 20-30% of the genes to enable a cross-species comparison in the first place. If it's the former reason, it makes sense that maternal genes would be quite enriched in the hybrid compared to the maternal genes of the male parent, similarly the enrichment of paternal genes in the maternal plant would be quite high. At the same time, the differences in expression of those genes would not be that notable when comparing them to the parent of the side whose genome they originate from (comparing maternal gene expression with the maternal plant and so on).

In GSEA of the *A. halleri x lyrata* v. *A. halleri* dataset, most pathways in downregulated plots seem to be related to defense mechanisms, which means that the hybrid plants seem to be experiencing less stress under the same growth conditions than parent species. However, there are three auxin-related processes among the downregulated pathways (auxin polar transport, cellular response to auxin stimulus, auxin-activated signaling pathway) while none of such are found among the upregulated pathways, suggesting lesser responsiveness to auxin, which is a growth hormone. Seed coat development-associated genes are also downregulated, which is notable as *A. halleri* is the maternal species. A conflicting result with that is the downregulation of abscisic acid processes, as it is the increase in this compound that could be responsible for developmental arrest patterns observed in the sample. Among the upregulated pathways, there is an apparent trend of increased photosynthesis and biomass accumulation (reductive pentose phosphate cycle, monocarboxylic acid metabolic process - an "umbrella process" for the generation of various primary and secondary metabolites, same for isopentenyl diphosphate).

Meanwhile, the GSEA results of *A. halleri x lyrata* v. *A. lyrata* are no less conflicting than those comparing to *A. halleri*: response to abscisic acid shows up in both downregulated and upregulated gene set plots (with higher abundance in the latter). Auxin responses is increased in the hybrid compared to the male parent, indicating that the growth patterns could have been inherited from the maternal plant but fail to be as efficient. Flowering-related genesets also seem to be downregulated in the hybrid.

No visual description of parental and hybrid plants themselves is available due to samples being obtained a long time before the beginning of the analysis. Therefore, it is not possible to verify the strength of certain gene set up- or downregulation. All samples were taken from leaf tissue, so any GSEA results containing processes occuring in other plant organs are dubious. For a better evaluation of processes happening in other plant organs, it would be useful to take samples from the flowering structures to determine whether the basis of sterility in the diploid hybrids truly is transcriptomic.

The interesting changes should be comprised of non-sex-specific genes, especially those responsible for plant development. However, these changes in form of whole geneset up- or downregulation could not be observed. Therefore, the results of this analysis are inconclusive.

4.2 Evaluation of the Analysis

It is hard to say how influential the strong pairing within the parental samples is on the overall analysis, as the statistical significance of a sample size of 2 is hard to evaluate. It would be recommended to repeat the experiment with more controlled conditions and, at best, take samples from one plant only, rather than having four individuals each sampled once - such sampling creates a covariate of inter-individual variability which is much harder to account for than simply processes occurring in one same parent plant but spread between different locations (leaf, flower etc.).

The major downside of the "gene translation" method used in this analysis is that only 68% of the DEGs remain after the "translation" from *A. halleri* to *A. thaliana*, while *A. lyrata* retains around 80% of its DEGs after translation to *A. thaliana*. There are also one-way hit gene lists available - these datasets have more proposed *A. thaliana* equivalents present and therefore allow to retain more genes. However, the one-way hit files are less reliable as the BLAST scores of the non-reciprocal hits among all entries may be quite low - after all, the reason for their exclusion from the reciprocal hit list is that the gene queries between investigated parental species and *A. thaliana* do not match. The "translation" step was unavoidable as it was a prerequisite to carry out GSEA. Alternative GSEA/ORA frameworks, such as camera which is native to the edgeR package, could also be used.

Further analysis could also be done on this data, investigating potential genomic origin of genes transcribed by the hybrid plant. For this, the tool *EAGLE* [23] could be used. This tool evaluates the probability with which a read belongs to a certain gene over another. By combining the probabilistic results of EAGLE with this analysis, it would be possible to get a better picture of the genomic differences that lie under the changes in expression patterns. In addition to that, epi-

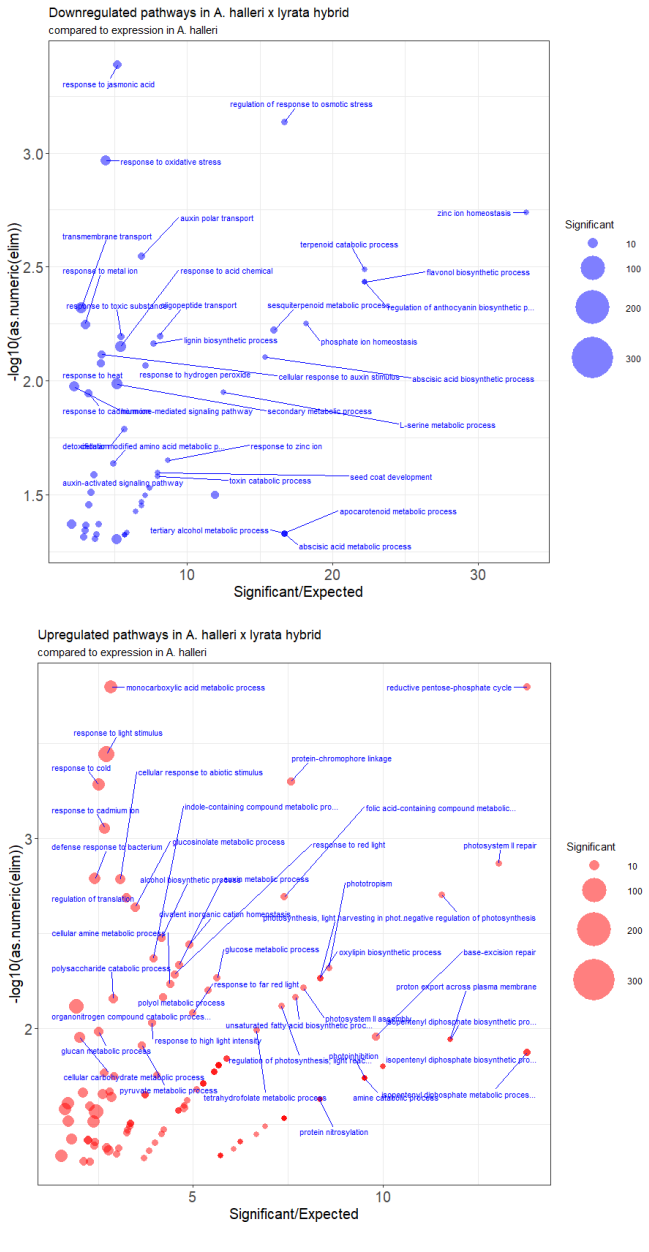

Figure 8: GSEA plots of downregulated (above) and upregulated (below) pathways in the *A. halleri x lyrata* hybrid compared to *A. halleri* parent.

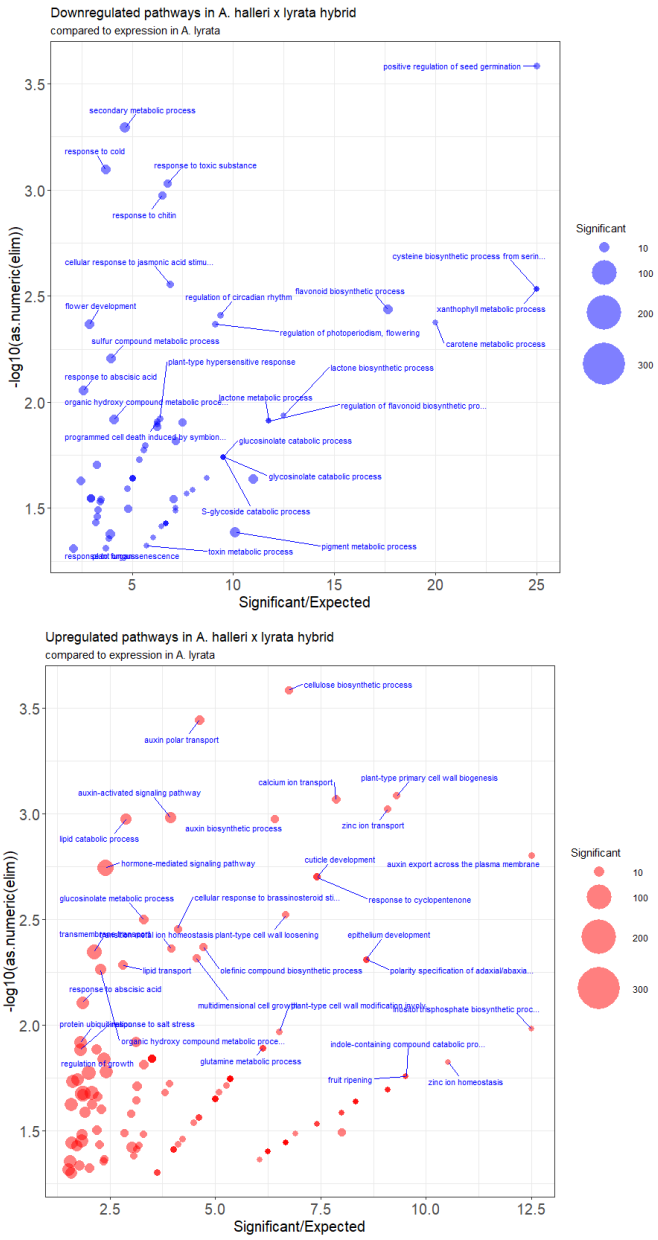

Figure 9: GSEA plots of downregulated (above) and upregulated (below) pathways in the *A. halleri x lyrata* hybrid compared to *A. lyrata* parent.

genetic changes might affect transcription patters and might even be responsible for the imbalance in normally antagonistic pathways.

References

[1] One thousand plant transcriptomes and the phylogenomics of green plants. Nature [Internet]. 2019 Oct 23;574(7780):679–85. Available from: http://www.nature.com/articles/s41586-019- 1693-2

[2] Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics,

and the temporal dynamics of genome fractionation. Curr Opin Genet Dev [Internet]. 2018 Apr;49:1–7. Available from: https://linkinghub.elsevier.com/ retrieve/pii/S0959437X17301557

[3] Jiang X, Song Q, Ye W, Chen ZJ. Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1382–93. Available from: https://www.nature.com/articles/s41559-021-01523-y

[4] Kashkush K, Feldman M, Levy AA. Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid. Genetics [Internet]. 2002 Apr 1;160(4):1651–9. Available from: https://academic.oup.com/genetics/ article/160/4/1651/6049775

[5] Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, et al. Evolutionary Genetics of Genome Merger and Doubling in Plants. Annu Rev Genet [Internet]. 2008 Dec 1;42(1):443–61. Available from: https://www.annualreviews.org/doi/ 10.1146/annurev.genet.42.110807.091524

[6] Yoo M-J, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) [Internet]. 2013 Feb 21;110(2):171–80. Available from: http://www.nature.com/articles/hdy201294

[7] Wang J, Tian L, Lee H-S, Wei NE, Jiang H, Watson B, et al. Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. Genetics [Internet]. 2006 Jan 1;172(1):507–17. Available from: https://academic.oup.com/genetics/ article/172/1/507/6065206

[8] Adams KL. Evolution of Duplicate Gene Expression in Polyploid and Hybrid Plants. J Hered [Internet]. 2007 Mar 1;98(2):136–41. Available from: http://academic.oup.com/jhered/article/98/2/136/ 2187871/Evolution-of-Duplicate-Gene-Expression-in

[9] Chen ZJ. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu Rev Plant Biol [Internet]. 2007 Jun;58(1):377–406. Available from:http://www.annualreviews.org/doi/ 10.1146/annurev.arplant.58.032806.103835

[10] Chelaifa H, Monnier A, Ainouche M. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina × townsendii and Spartina anglica (Poaceae). New Phytol [Internet]. 2010 Apr;186(1):161–74. Available from: https://onlinelibrary.wiley.com/doi/ 10.1111/j.1469-8137.2010.03179.x

[11] Kuo TCY, Hatakeyama M, Tameshige T, Shimizu KK, Sese J. Homeolog expression quantification methods for allopolyploids. Brief Bioinform [Internet]. 2018 Dec 27; Available from: https://academic.oup.com/bib/ advance-article/doi/10.1093/bib/bby121/5251019

[12] Paape T, Briskine R V., Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun [Internet]. 2018 Dec 25;9(1):3909. Available from: http://www.nature.com/articles/s41467-018-06108-1

[13] SHIMIZU-INATSUGI R, LIHOVÁ J, IWANAGA H, KUDOH H, MARHOLD K, SAVOLAINEN O, et al. The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Mol Ecol [Internet]. 2009 Oct;18(19):4024–48. Available from: http://doi.wiley.com/10.1111/ j.1365-294X.2009.04329.x

[14] Milosavljevic S, Kuo T, Decarli S, Mohn L, Sese J, Shimizu KK, Shimizu-Inatsugi R, Robinson MD. ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw. BMC Genomics. 2021 Jul 17;22(1):547. doi: 10.1186/s12864-021-07845-2. PMID: 34273949; PMCID: PMC8285871.

[15] Milosavljevic et al. Manuscript in progress.

[16] Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

[17] Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.

[18] Gu Z, Eils R, Schlesner M (2016). "Complex heatmaps reveal patterns and correlations in multidimensional genomic data." Bioinformatics. doi: 10.1093/bioinformatics/btw313.

[19] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[20] gghighlight by Hiroaki Yutani https://cran.r-project.org/web/packages/gghighlight/ vignettes/gghighlight.html

[21] ggrepel by Kamil Slowiowski https://cran.r-project.org/web/packages/ggrepel/ vignettes/ggrepel.html

[22] Alexa A, Rahnenfuhrer J (2022). topGO: Enrichment Analysis for Gene Ontology. R package version 2.50.0.

[23] Kuo, T., Frith, M.C., Sese, J. et al. EAGLE: Explicit Alternative Genome Likelihood Evaluator. BMC Med Genomics 11 (Suppl 2), 28 (2018). https://doi.org/10.1186/s12920-018-0342-1