Brachypodium distachyon infected with various isolates of *Zymoseptoria tritici*

Harriet R. Benbow

2020-07-16

Contents

1 Introduction

The purpose of this analysis is to yield more information from the RNAseq samples of *B. distachyon* ecotype Bd21 inoculated with different isoluate os *Zymoseptoria tritici*.

The RNAseq samples are paired end RNAseq samples. The samples were aligned to a reference sequence, and transcript abundance was estimated using Kallisto. As both host and pathogen transcripts are of interest, the reference was a combined reference of the gene annotations of *B. distachyon* and *Z. tritici*. In total, 36 transcript abundance files were created, one for every set of paired end reads (Table 1).

Table 1: Samples

Sample	Description	Rep	Timepoint	Brachy_genotype	Zymo_isolate
R1-553-0dpi	Rep 1 553.11	1	0	Bd21	553.11
R1-553-4dpi	Rep 1 553.11	1	4	Bd21	553.11
R1-553-9dpi	Rep 1 553.11	1	9	Bd21	553.11
R1-553-21dpi	Rep 1 553.11	1	21	Bd21	553.11
R1-560-0dpi	Rep 1 560.11	1	0	Bd21	560.11
R1-560-4dpi	Rep 1 560.11	1	4	Bd21	560.11
R1-560-9dpi	Rep 1 560.11	1	9	Bd21	560.11
R1-560-21dpi	Rep 1 560.11	1	21	Bd21	560.11
R1-323-0dpi	Rep 1 IPO323	1	0	Bd21	IPO323
R1-323-4dpi	Rep 1 IPO323	1	4	Bd21	IPO323
R1-323-9dpi	Rep 1 IPO323	1	9	Bd21	IPO323
R1-323-21dpi	Rep 1 IPO323	1	21	Bd21	IPO323

Sample	Description	Rep	Timepoint	Brachy_genotype	Zymo_isolate
R2-553-0dpi	Rep 2 553.11	2	0	Bd21	553.11
R2-553-4dpi	Rep 2 553.11	2	4	Bd21	553.11
R2-553-9dpi	Rep 2 553.11	2	9	Bd21	553.11
R2-553-21dpi	Rep 2 553.11	2	21	Bd21	553.11
R2-560-0dpi	Rep 2 560.11	2	0	Bd21	560.11
R2-560-4dpi	Rep 2 560.11	2	4	Bd21	560.11
R2-560-9dpi	Rep 2 560.11	2	9	Bd21	560.11
R2-560-21dpi	Rep 2 560.11	2	21	Bd21	560.11
R2-323-0dpi	Rep 2 IPO323	2	0	Bd21	IPO323
R2-323-4dpi	Rep 2 IPO323	2	4	Bd21	IPO323
R2-323-9dpi	Rep 2 IPO323	2	9	Bd21	IPO323
R2-323-21dpi	Rep 2 IPO323	2	21	Bd21	IPO323
R3-553-0dpi	Rep 3 553.11	3	0	Bd21	553.11
R3-553-4dpi	Rep 3 553.11	3	4	Bd21	553.11
R3-553-9dpi	Rep 3 553.11	3	9	Bd21	553.11
R3-553-21dpi	Rep 3 553.11	3	21	Bd21	553.11
R3-560-0dpi	Rep 3 560.11	3	0	Bd21	560.11
R3-560-4dpi	Rep 3 560.11	3	4	Bd21	560.11
R3-560-9dpi	Rep 3 560.11	3	9	Bd21	560.11
R3-560-21dpi	Rep 3 560.11	3	21	Bd21	560.11
R3-323-0dpi	Rep 3 IPO323	3	0	Bd21	IPO323
R3-323-4dpi	Rep 3 IPO323	3	4	Bd21	IPO323
R3-323-9dpi	Rep 3 IPO323	3	9	Bd21	IPO323
R3-323-21dpi	Rep 3 IPO323	3	21	Bd21	IPO323

2 Results

2.1 RNAseq pre-processing

The first step of the RNAseq analysis is to assess the quality of the data. Firstly, we look at correlation between the reps. All three reps were strongly correlation with eachother, indicating good agreement of gene expression between the three reps (Figure 1).

2.1.1 Principle component analysis to identify sources of variance

Here, we use principle component analysis as a method if dimension reduction, to identify key drivers of variation in your data. We do this on the plant and fungal data separately, as otherwise the signal of the fungal data is lost due to the dominance of the plant reads.

2.1.1.1 *B. distachyon* reads When we insepct the *B. distachyon* reads, we can see that the clustering shows clusters with an obvious reason. We see a slight rep effect (to be expected), but no other patterns in the data that explain the clustering (Figure 2). This suggests natural biological variation is the main driver. N.B. don't be suprised that there are not any obvious clusters based on fungal isolate used. The effect of the isolate would not be sufficient enough to cause enough variation in the brachy reads that could be picked up by principle componends 1 and 2.

2.1.1.2 *Z. tritici* reads When we explore the *Z. tritici* reads only, we see a clear clustering based on timepoint, with T0 reads clustered together, separated from the other timepoints. This is to be expected

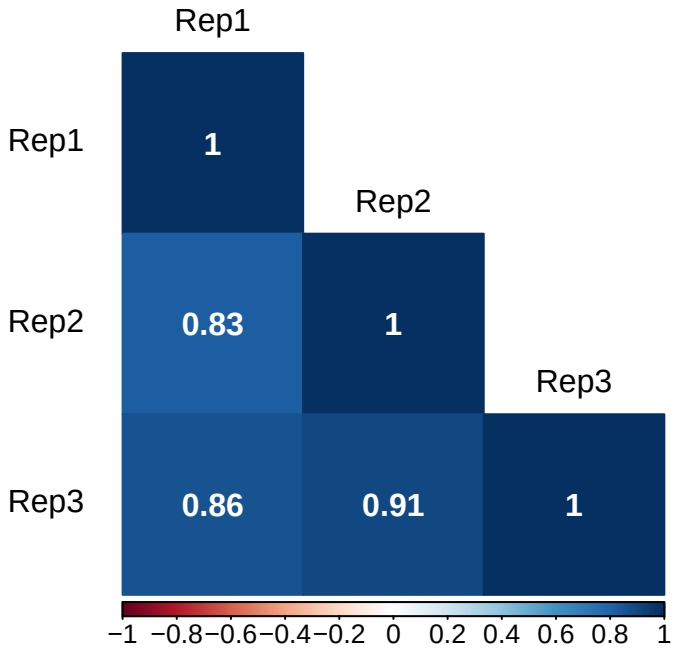

Figure 1: A correlation heatmap of the three reps. The number represents the correlation coefficient for each pair of reps.

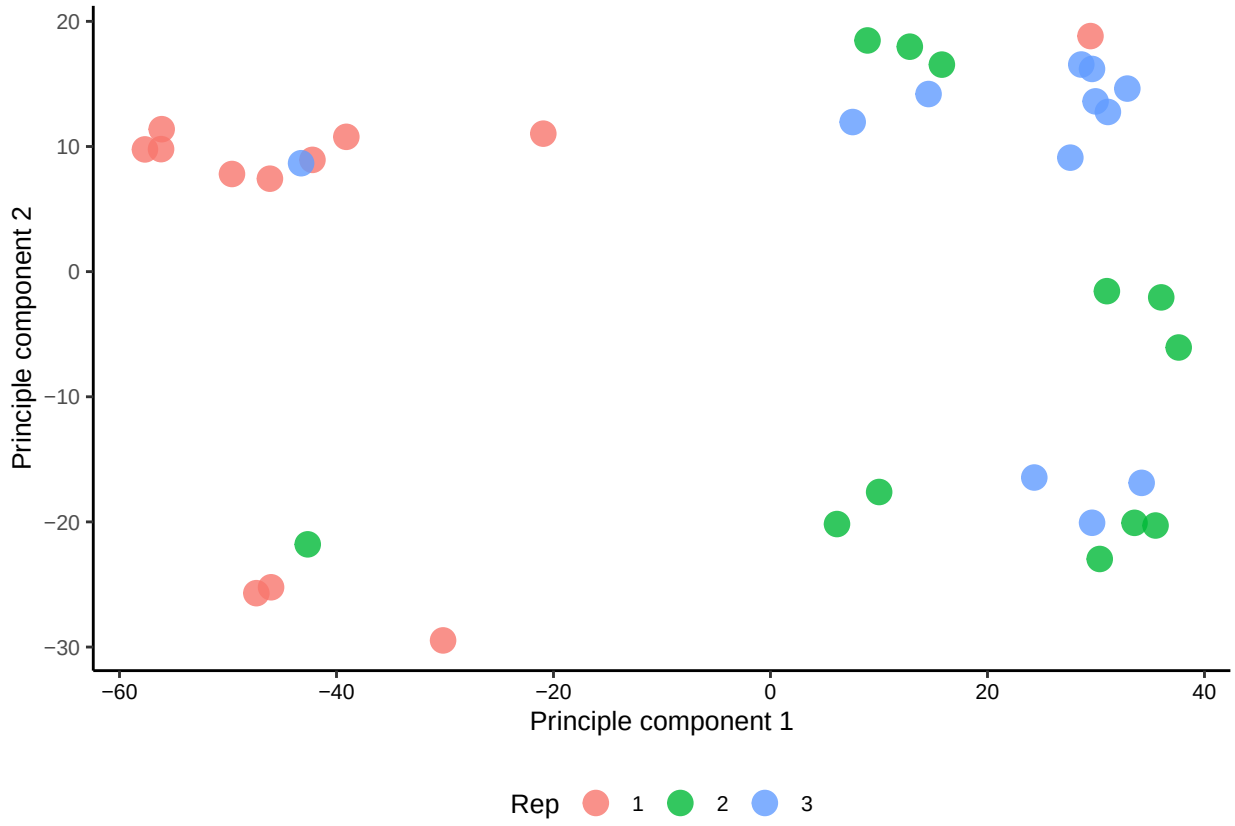

Figure 2: A principle component analysis of B. distachyon reads. Colours represent reps.

(and is something I have seen before in my data - the early timepoint on its own). You can also see some level of clustering between the other timepoints, but these groups are not very clearly defined (Figure 3). The shapes of the points represent the isolates, and as you can see, there is not much variance that we can attribute to isolate.

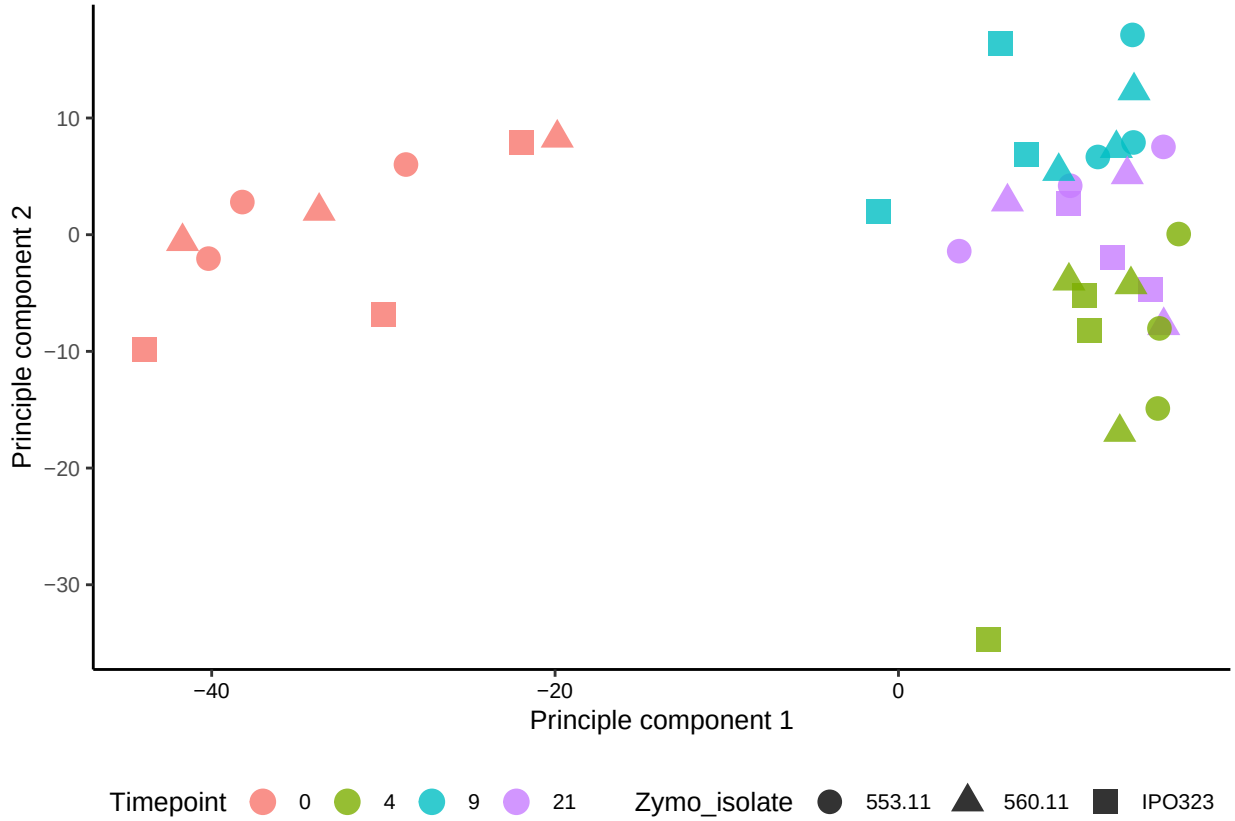

Figure 3: A principle component analysis of Z. tritici reads. Colours represent Timepoint and shape represents fungal isolate used.

2.2 Read count statistics

As part of the pre-processing of RNAseq data, it is good to asses and quantify the number of genes expressed in each sample. This is of particular importance when looking at dual RNAseq of a host + pathogen system. To do this, we filter out genes based on their expression level, and the consistency of their expression across the reps. Here, I have filtered genes that are expressed at 0.5 transcripts per million (TPM) or more, in 2 out of the three reps. For example, for 'Gene A' to be considered 'expressed' in Sample 1, timepoint 1, it would have to be expressed (TPM >=0.5) in 2 out of the three reps for that sample. Based on these criteria, we found a total of 55,420 genes to be expressed in this data, with an average of 46,781 genes expressed per sample (Table 2).
Please note that these are fairly stringent filtering parameters, and may not be necessary, especially if you wish to detect pathogen genes that may be present at very low levels.

Table 2: Number of expressed genes per sample (averaged across reps)

Sample	Number of expressed genes
553.11 0	45736
553.11 21	47595
553.11 4	47779
553.11 9	47133
560.11 0	43716
560.11 21	48013
560.11 4	46770
560.11 9	47115
IPO323 0	44305
IPO323 21	47733
IPO323 4	48077
IPO323 9	47410

We can break these down into host and pathogen genes and see how many genes from each species are expressed in each sample. As we can see from figure 4, the number of *B. distachyon* genes is fairly stable cross the timepoints, and we see a slight increase in *Z. tritici* reads from 0 DPI to 4 DPI, and the number of genes expressed is then stable across timepoints. This is really interesting and informative, it could mean that the fungus is growing between days 0 and 4, but because it is the *number* of genes expressed, rather than the *abundance* of genes expressed, it more likely indicates that the fungus is active and doing stuff at 4 DPI that its not doing at 0 DPI. This is logical really as at 0 DPI its just been chilling on a petri dish!

2.3 Differential expression analysis

In this sections, we explore differential gene expression between samples. As there are no *Z. tritici* controls (i.e. samples without *Z. tritici*), we can explore everything with respect to timepoint 0. We can also look at differences between isolates. Basically at this point I went rogue and did all comparisons; for each isolate, I compared every timepoint to 0 DPI (table 3), and for every timepoint, I compared every pair of isolates (table 4). This yielded to sets of differentially expressed genes.

Table 3: Number of differentally expressed genes

Species	Timepoint	Isolate	Number of DEGs
Bd	4	553.11	828
Zt	4	553.11	1093
Bd	9	553.11	667
Zt	9	553.11	4151
Bd	21	553.11	270
Zt	21	553.11	1289
Bd	4	560.11	1309
Zt	4	560.11	584
Bd	9	560.11	365
Zt	9	560.11	504
Bd	21	560.11	301
Zt	21	560.11	465
Bd	4	IPO323	282
Zt	4	IPO323	412
Bd	9	IPO323	164

Species	Timepoint	Isolate	Number of DEGs
Zt	9	IPO323	113
Bd	21	IPO323	710
Zt	21	IPO323	371

Table 4: Number of differentally expressed genes

Species	Timepoint	Comparison	Number of DEGs
Bd	0	553.11 v IPO323	46
Zt	0	553.11 v IPO323	61
Bd	4	553.11 v IPO323	40
Zt	4	553.11 v IPO323	11
Bd	9	553.11 v IPO323	53
Zt	9	553.11 v IPO323	61
Bd	21	553.11 v IPO323	231
Zt	21	553.11 v IPO323	26
Bd	0	560.11 v 553.11	56
Zt	0	560.11 v 553.11	57
Bd	4	560.11 v 553.11	71
Zt	4	560.11 v 553.11	11
Bd	9	560.11 v 553.11	53
Zt	9	560.11 v 553.11	14
Bd	21	560.11 v 553.11	65
Zt	21	560.11 v 553.11	20
Bd	0	560.11 v IPO323	62
Zt	0	560.11 v IPO323	42
Bd	4	560.11 v IPO323	51
Zt	4	560.11 v IPO323	11
Bd	9	560.11 v IPO323	61
Zt	9	560.11 v IPO323	32
Bd	21	560.11 v IPO323	259
Zt	21	560.11 v IPO323	28

2.3.1 Isolate specific DEGs (per timepoint)

We can start to subset the differentially expressed genes based on interesting biological questions. Firsly, we can look at genes that are differentially expressed at a partcular timepoint (compared to 0 DPI) in 1, 2 or 3 of the isolates. Given the observed 'aggresiveness' of 553.11, it might be interesting to look at genes that are differentially expressed in this isolate, and not the others. Similarly, it might be interesting to see genes that are differentially expressed between the isolates, at each timepoint.
First, we look at genes that are differentially expressed by timepoint (Figure 5).

2.3.2 DEGs between the isolates

Secondly, we can see genes that are differentially expressed by isolate, across the timepoints (Figure 6).

As we can see from Figure 6, tghere are no genes that are differentially expressed between all three comparisons. Given the interesting phenotype observed on leaves inoculated with 553.11, it might be interesting to look at genes that are differentially expressed between 553.11 and the other two, but not differentially expressed between 560.11 and IPO323. For example in Figure 6 part A, the 12, 61 and 12 genes might be of interest when searching for a 553.11-specific transcriptional response.

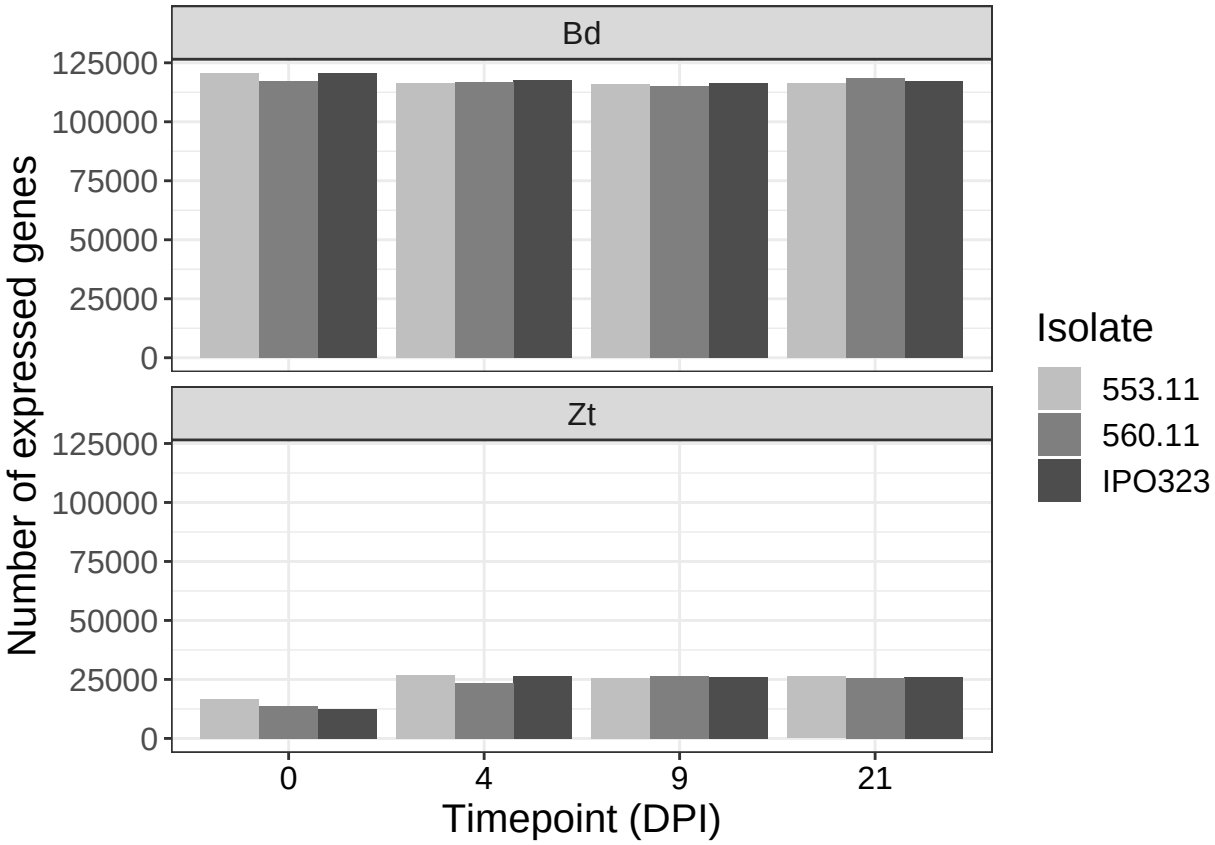

Figure 4: The number of expressed genes from both species across the 4 timepoints.

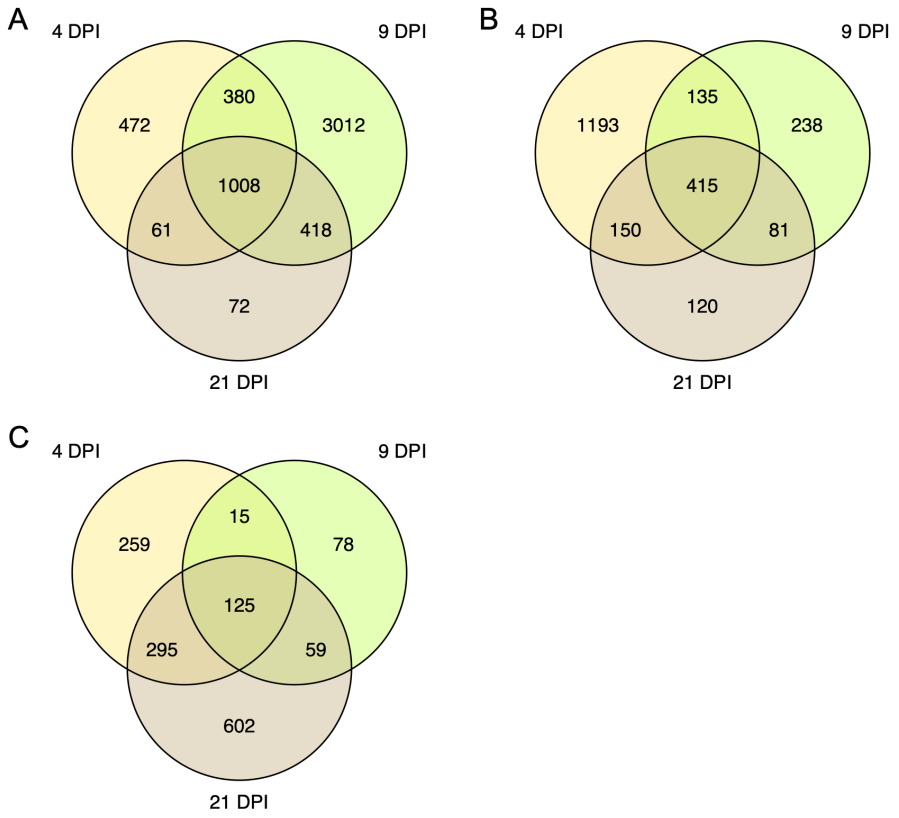

Figure 5: Venn diagrams of differentially expressed genes between timepoints, using timpoint '0 DPI' as the control A) Isolate 553.11, B) Isolate 560.11, C) Isolate IPO323.

For all of the DEGs, and their subsets thereof, I have put csv tables in the directory for you. Each of these tables has the gene name as rows, and the condition (so either timepoint, or comparison) as rows, and each gene has a 1 if it is differentially expressed and a 0 if it is not. I find the easiest way to find genes in subsets I am interested in is to just sort by those columns, and extract genes that have a 1 in those columns, and a 0 in the others. You can then extract their differential expression data from the files 'all_genetype_sig.csv' and 'all_timepoint_sig.csv'.

2.3.3 Species breakdown of differentially expressed genes

As we are looking at dual RNAseq, it is important to dileneate the DEGs by species. As we can see in Figure 7, there is a high proprtion of the DEGs in the 553.11-treated saples that are *Z. tritici* genes. This seems specific to the 553.11-treated samples. Another interesting/nice observatioun is that the number of DEGs increases over time across the comparisons between isolates. Of course these are just preliminary observation- there is alot of fishing that can be done within these data to find the patterns. We can also see that there are alot more DEGs in the samples where the timepoints were compared than the isolates. This will be because there are no controls so many of these DEGs will be because of differences in expression that is independant of treatment. However these genes may be interesting to look at with respect to DEGs that are also differentially expressed between the isolates (at that should be timepoint-independant).

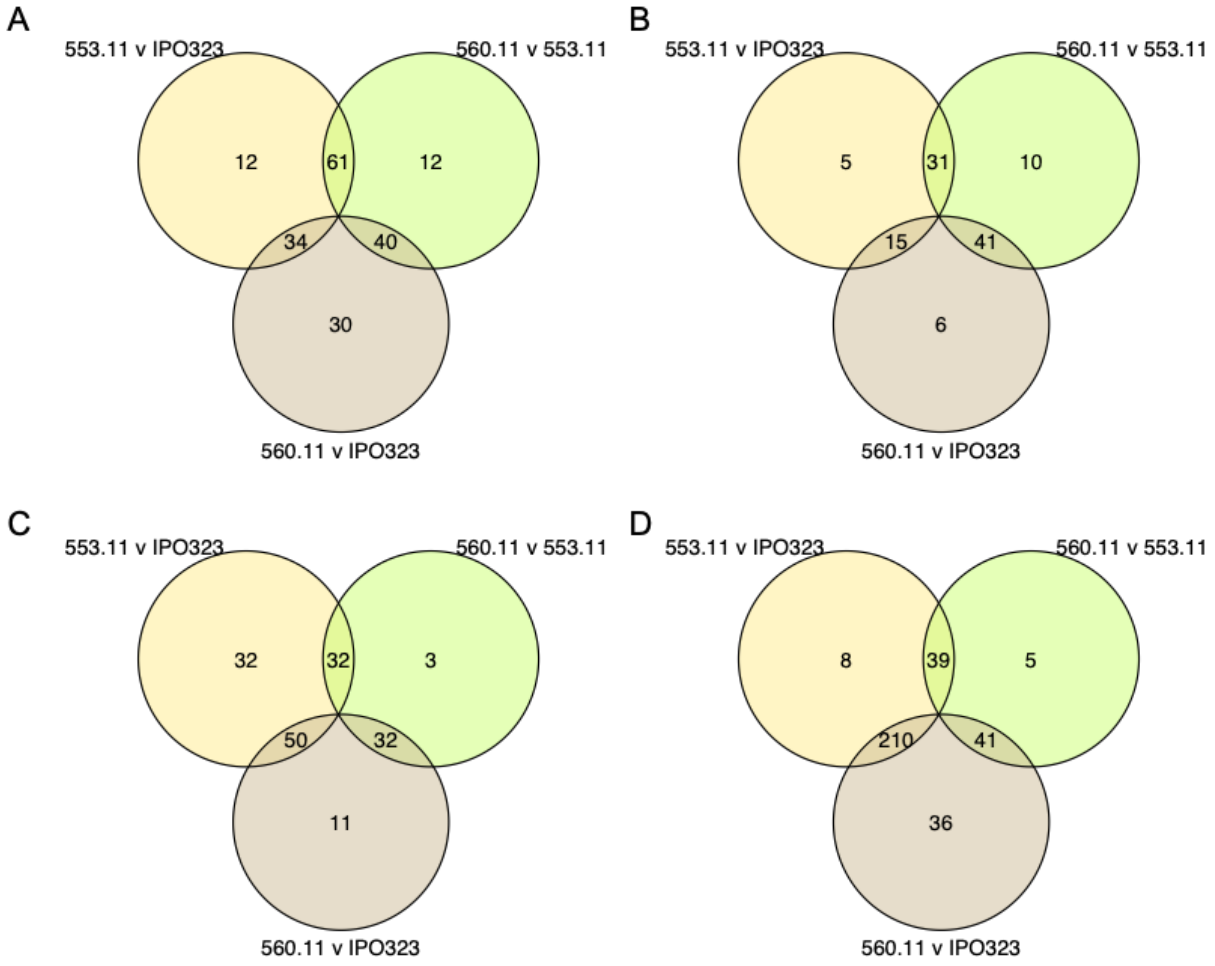

Figure 6: Venn diagrams of differentially expressed genes between isolates. A) 0 DPI, B) 4 DPI, C) 9 DPI, D) 21 DPI

2.4 Mining data for small secreted proteins (SSPs)

Now to the actual reason I started this! The goal here was to identify putative SSPs from *B. distachyon* and *Z. tritici* that are expressed/differentially expressed in the RNAseq data. Firstly, I use the SSP-hunting pipeline from Zhou et al., (2020) to identify SSPs in across the whole genomes of both species. Briefly, we chack for length (<=250 AA), presence of a signal peptide, presence of a transmembrane helix domain, and presence of a GPI anchor. Protiens that are <=250 AA in length with a signal peptide, no TM domain and no GPI anchor are considered putative SSPs.
Based on these criteria, I found *B. distachyon* to have 610 putative SSPs, and *Z. tritici* to have 234. The *B. distachyon* gene annotation contains 52,972 genes and the *Z. tritici* gene annotation contains 10,931. Therefore the percentage of each annotation that consists of SSPs is 1.15% and 2.14%, respectively. Just for comparison, we found the wheat genome to comprise 2.3% SSPs, so not too different.

Next, we find which, if any of these SSPs are differentially expressed in your RNAseq data. We found 27 SSPs differentially expressed between isolates (table 5), and 133 SSPs differentially expressed by timepoint (table 6).

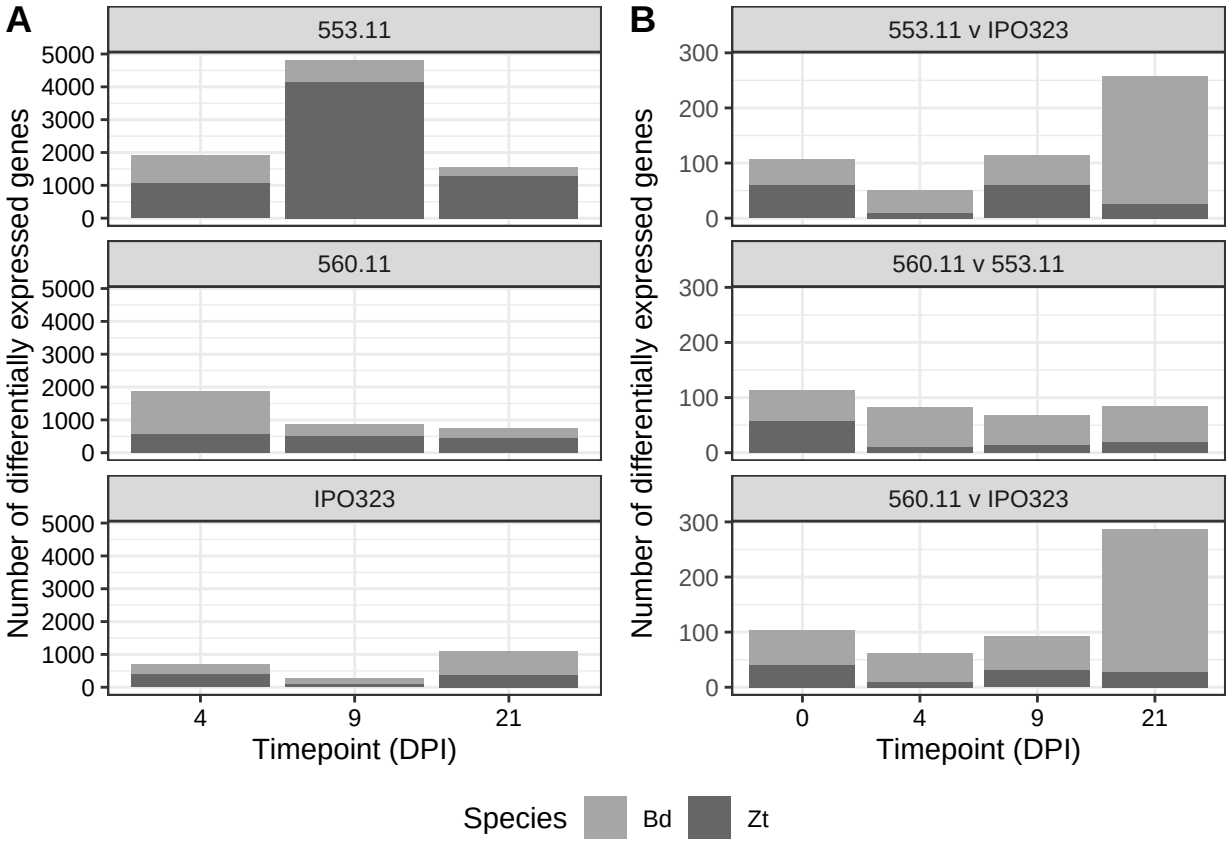

Figure 7: The number of differentially expressed genes by species A) Differentially expressed genes between each timepoint and 0 DPI, separated by isolate, B) Differentially expressed genes betweeen isolates.

Table 5: Number of differentally expressed SSPs (by isolate)

Timepoint	Comparison	Bd DE SSPs	Zt DE SSPs
0	553.11 v IPO323	0	2
0	560.11 v 553.11	0	9
0	560.11 v IPO323	0	10
4	553.11 v IPO323	0	0
4	560.11 v 553.11	0	0
4	560.11 v IPO323	0	1
9	553.11 v IPO323	0	10
9	560.11 v 553.11	0	3
9	560.11 v IPO323	0	5
21	553.11 v IPO323	7	2
21	560.11 v 553.11	0	4
21	560.11 v IPO323	7	4

Table 6: Number of differentally expressed SSPs (by timepoint)

Timepoint	Isolate	Bd DE SSPs	Zt DE SSPs
4	553.11	17	48
4	560.11	21	38
4	IPO323	6	29
9	553.11	17	73
9	560.11	5	24
9	IPO323	5	9
21	553.11	5	35
21	560.11	3	18
21	IPO323	18	26