Brachypodium distachyon infected with various isolates of *Zymoseptoria tritici*

Harriet R. Benbow

02/07/2020

Contents

1 Introduction **1**

2 Results **2**

2.1 RNAseq pre-processing 2

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 *Zymoseptora 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
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

Sample	Description	Rep	Timepoint	Brachy_genotype	Zymo_isolate
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 (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.

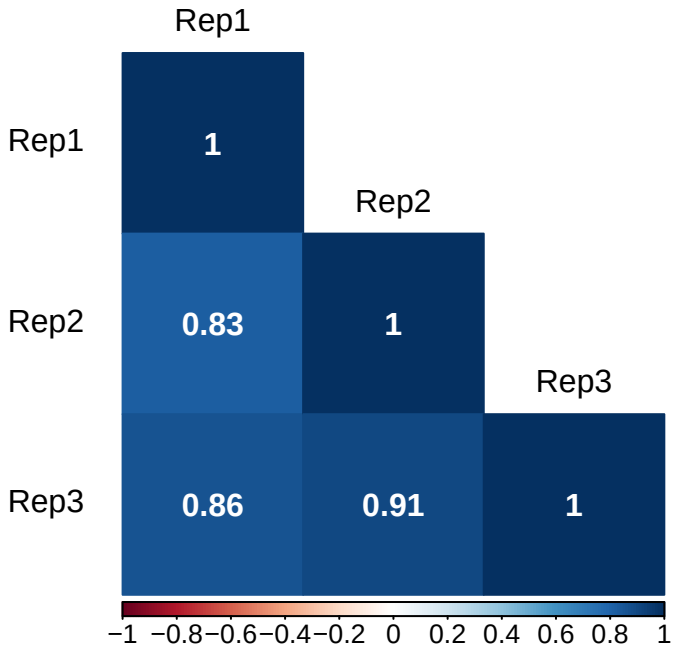

Figure 1: A correlation heatmap of the three reps. The number represents the correlation coefficient for each pair of reps.

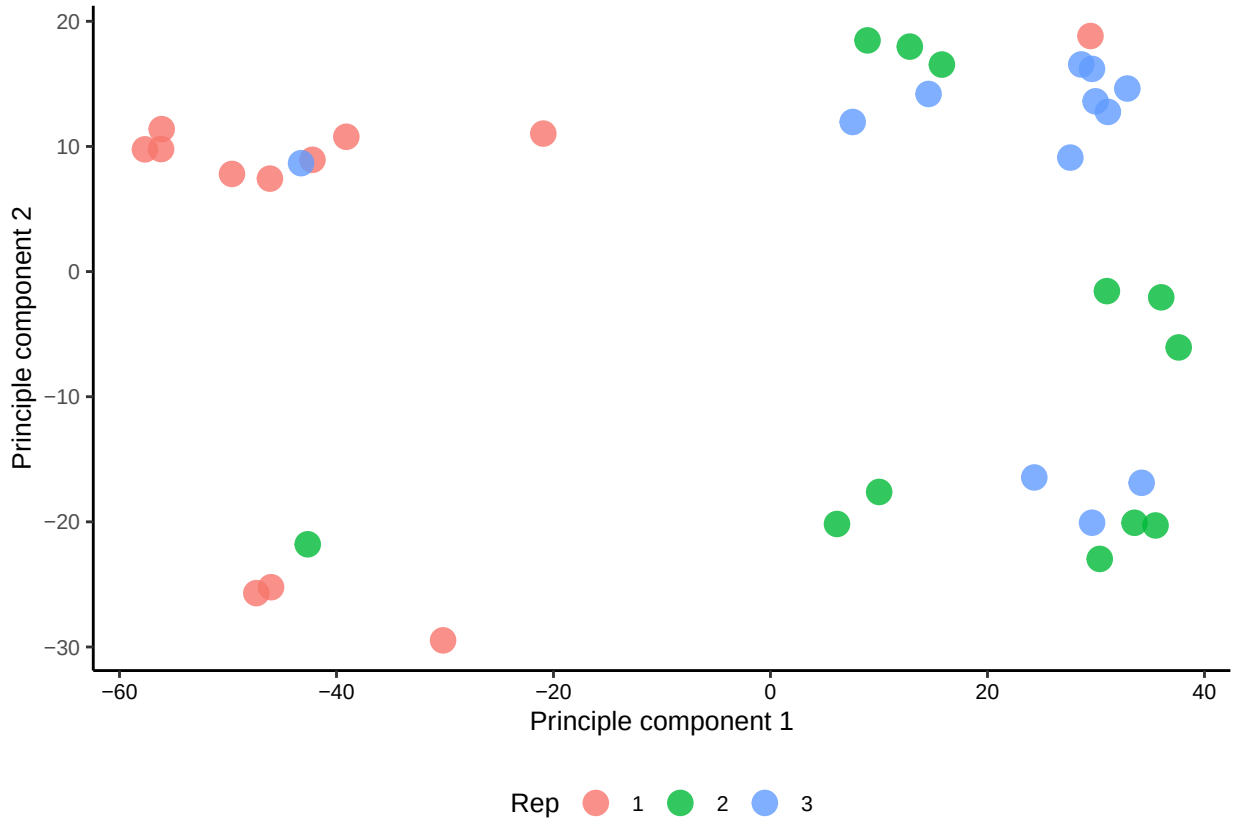

Figure 2: A principle component analysis of *B. distachyon* reads. Colours represent reps.

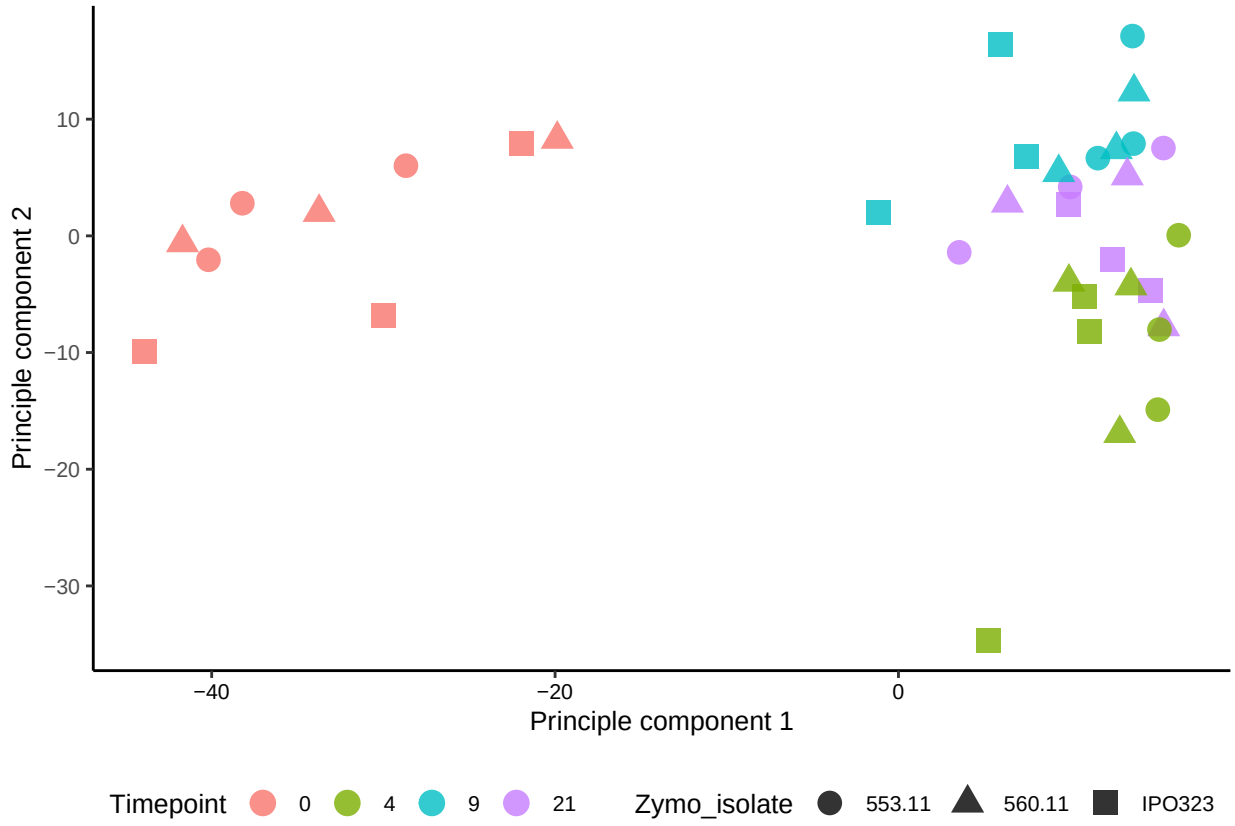

Figure 3: A principle component analysis of *Z. tritici* reads. Colours represent Timepoint and shape represents fungal isolate used.

2.1.2 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 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!

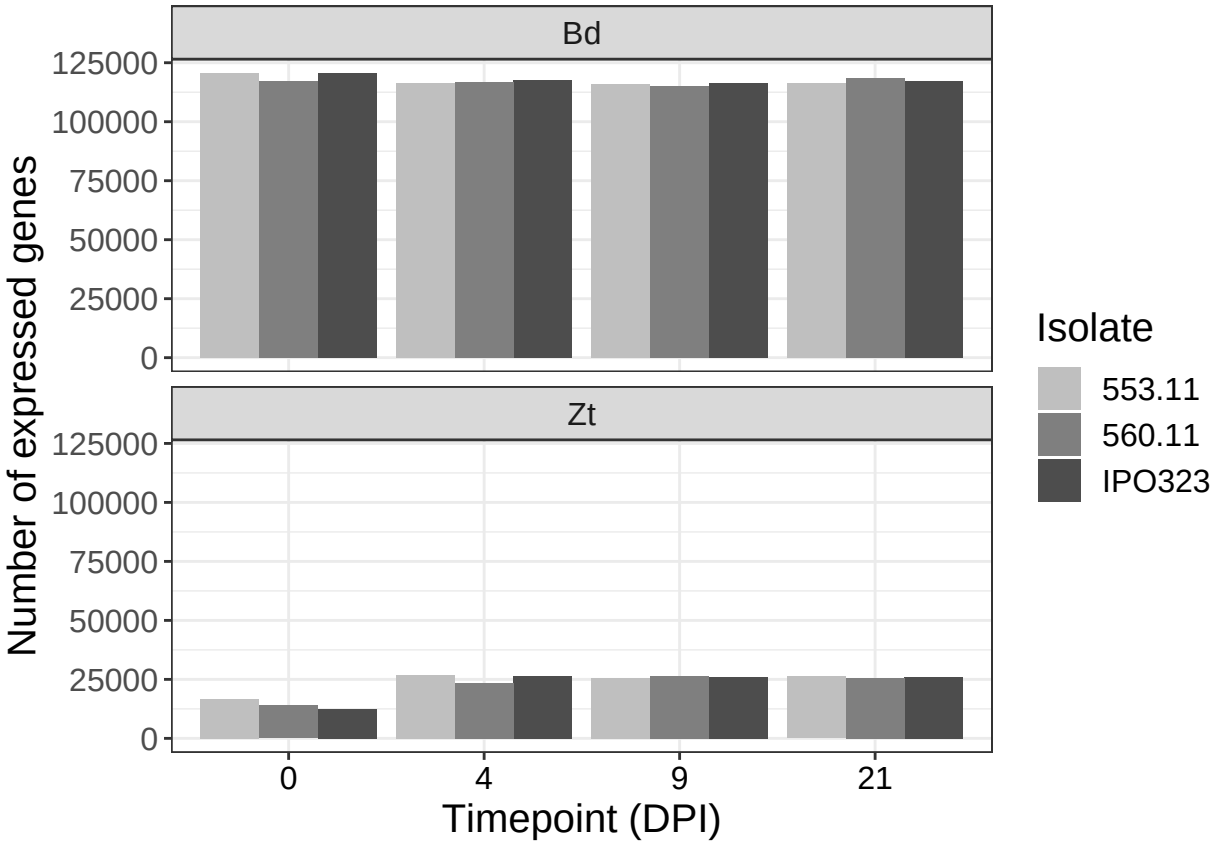

Figure 4: The number of expressed genes from both species across the 4 timepoints.