Identifying QTLs for improving acclimation of photosynthetic efficiency after periods of fluctuating light

Author: Bram Duurland
Degree programme: Master Plant Biotechnology, specialisation Functional Genomics
Student number: 1012491
Course code: GEN80436
Supervisor: Dr. Phuong Nguyen
Start date: 16-01-2023
End date: 11-07-2023
Chair group: Laboratory of Genetics

Abstract

Photosynthesis is one of the most important mechanisms found in nature, on which many organisms rely as energy source. During the light reaction, light energy is converted into chemical energy and fixed during the dark reaction in carbohydrates. In high light conditions, excess energy is dissipated as heat through the photoprotection mechanism called non photochemical quenching (NPQ). However, under fluctuating light conditions, NPQ inhibits photochemistry decreasing the quantum yield of photosystem II (ϕPSII). This in turn decreases the amount of biomass accumulation. So, in order to improve acclimation of photosynthetic efficiency after fluctuating light (FL) periods, two FL experiments were performed with a Dutch *Arabidopsis thaliana* population. These two experiments had the exact same conditions, only the FL treatments were reversed to account for plant age, since plant age is an important factor in variation of photosynthetic efficiency. For one of the two experiments, photosynthetic parameters for each individual leaf were collected, to further decrease influences of plant age. The resulting phenotypic data was used for genome wide association studies (GWAS), to identify QTLs involved in the acclimation of photosynthetic parameters after periods of FL. The resulting QTLs showed that most of the variation in photosynthetic efficiency can be attributed to plant development. The most promising QTLs to be involved in the acclimation of photosynthetic efficiency after periods of FL, were identified on chromosome 1, 3 and 4 of *A. thaliana.*

Index

Introduction

Background

Life on earth is highly dependent on the process of photosynthesis by plants, algae and some bacteria. Photosynthesis is a complex mechanism in which light energy is converted into chemical energy, forming the base for every ecosystem. The energy is fixed in organic compounds using inorganic, atmospheric and soil compounds. Because of this photosynthesis not only provides other organisms with energy, but it is also important for maintaining earth's atmosphere and in turn climate.

The light reaction of photosynthesis is mediated by a chain of several proteins in the membrane of thylakoids inside chloroplasts (Figure 1). The main protein complexes are photosystems I and II (PSI & PSII respectively), PSII uses the energy from incoming light to split water molecules into hydrogen and an electron. The resulted hydrogen gradient is then used by ATP-synthase to create Adenosine Triphosphate (ATP) from Adenosine Diphosphate (ADP). The electron is transported through a cytochrome complex and plastocyanin to reach PSI, where the electron is used to drive the reduction of Nicotinamide adenine dinucleotide phosphate ($NADP^+$) into NADPH (Hou, 2012).
The other important reaction of photosynthesis is called the dark reaction. The dark reaction or Calvin cycle happens simultaneously with the light reaction but is called dark reaction as it does not require light to function. In the Calvin cycle the previously formed ATP and NADPH are used to add carbon molecules from carbon dioxide (CO_2) to unstable sugar molecules and form glucose and other carbohydrates useful for the plant (*A Dictionary of Biology*, 2019).

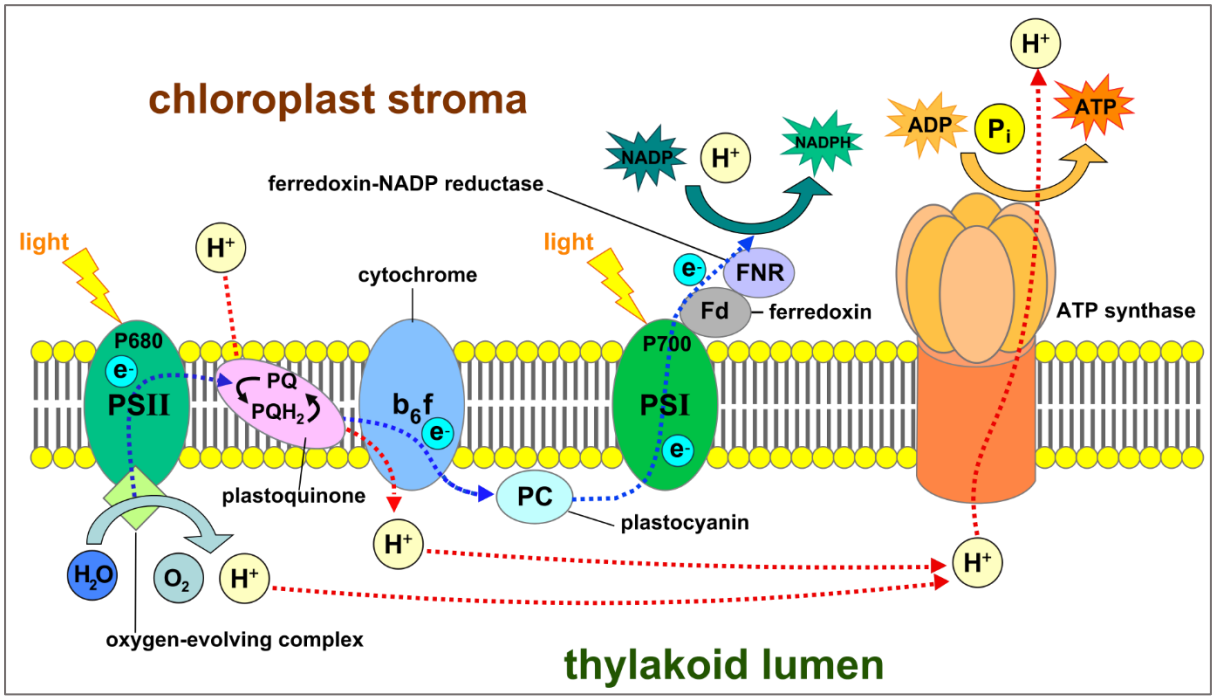


Figure 1: Schematic representation of the thylakoid membrane with the photosynthesis machinery of the light reaction (Wikimedia Commons, 2015). Light energy is absorbed by photosystem II (PSII), which mediates electron transport through the membrane. This electron is obtained from separating hydrogen from water, resulting in a hydrogen gradient between the inside and outside membrane. This gradient is used to create Adenosine Triphosphate (ATP) from Adenosine Diphosphate (ADP). The electron travels through the membrane, through the cytochrome complex, to Photosystem I (PSI) where it is used to create NADPH from Nicotinamide adenine dinucleotide phosphate (NADP) (A Dictionary of Biology, 2019).

The theoretical energy conversion of photosynthesis is only 4-6% in plants, which means that 4-6% of all incoming light is used in fixating chemical energy in carbohydrates. This inefficiency is due to the fact that only 45% of solar energy is photosynthetically active, and further energy loss is due to

reflection and respiration (Zhu et al., 2008). Compared to the most recent advancements in photovoltaics, with solar cells with an efficiency of around 25% (Green et al., 2021), this is a very low efficiency. On top of that the 4-6% energy conversion is only theoretical, the maximum recorded energy conversion is only 3-4%. Generally the observed maximum energy conversion is a third of the theoretical energy conversion (Zhu et al., 2010).

This difference between maximum theoretical energy conversion and maximum energy conversion is due to a variety of factors, like photorespiration and the photoprotection mechanism present in PSII. When in high light conditions, the Light Harvesting Complex (LHC) in PSII becomes oversaturated and needs to dissipate the excess energy. Otherwise the excess energy will form reactive oxygen species that will damage the cell (Tietz et al., 2017). I will focus further on the photoprotection mechanism, because most of the previous research on improving energy conversion has been on improving rubisco, but there is less known about photosynthetic efficiency and photoprotection mechanisms.

All incoming light that is absorbed by the LHC, is either used for photosynthesis (photochemistry), re-emitted as heat, or re-emitted as fluorescence. The part of the light that is used for photosynthesis is called the quantum yield of PSII (ϕPSII)(Murchie & Lawson, 2013). The dissipation of excess energy through heat is called Non-Photochemical Quenching (NPQ) and is dependant on the ratio between ϕNO/ϕNPQ, of which ϕNPQ is the regulated dissipation of energy, and ϕNO is the non-regulated dissipation. Because all of these processes constitute the total energy absorbed by the LHC, they exist in competition with each other, so the sum of ϕPSII + ϕNPQ + ϕNO = 1 (Kramer et al., 2004). In nature plants are exposed to fluctuating light intensities, from being in the shade of other plants or its own leaves. When a plant is exposed to what it perceives as high light intensity, NPQ is activated. But when the light intensity again decreases, because of shade for example, it takes some time for NPQ to relax causing it to compete with ϕPSII (Müller et al., 2001). The rate of NPQ induction and relaxation are important for efficient photosynthesis in fluctuating light conditions. To increase the photosynthetic efficiency, it is worthy to improve the regulation of ϕPSII as well as of NPQ, as they directly correlate with each other.

To study ϕPSII and NPQ, these parameters need to be measured. This can be done by measuring the chlorophyll fluorescence of PSII (Murchie & Lawson, 2013). In a set-up with dark adapted plants and actinic light (light that drives photochemistry), several parameters of fluorescence can be measured. The first is the maximum quantum efficiency of PSII photochemistry, Fv/Fm. Fm is the maximum fluorescence value in a dark-adapted leaf, where all LHCs are open, so no NPQ is taking place. Fv is the difference between the minimum fluorescence value (Fo) and the maximum fluorescence (Fm). This maximum value is obtained by exciting the leaf with a saturating pulse, which ensures all reaction centres in PSII are closed, but no electron transport takes place. The then measured fluorescence is Fm, which is visible as a vertical line figure 2A.

Then turning on the actinic light allows for measuring the fluorescence in light adapted leaves. At first there is an initial spike in fluorescence just as high as the dark-adapted state, but quenches (seen as a curve in figure 2B) as NPQ activates until it reaches a steady state fluorescence in a light-adapted state, called Fp. Giving the leaf again a saturating pulse, gives us the maximal fluorescence in a light adapted leaf (Fmp). With the parameters Fm, Fp and Fmp ϕPSII, ϕNO, ϕNPQ and NPQ can be calculated using the following equations.

$$\frac{Fmp-Fp}{Fmp} = \varphi\mathrm{PSII} \qquad \textit{(Eq. 1)}$$

$$\frac{Fp}{Fm} = \varphi\mathrm{NO} \qquad \textit{(Eq. 2)}$$

$$\left(\frac{Fp}{Fmp}\right) - \left(\frac{Fp}{Fm}\right) = \varphi\text{NPQ} \qquad \textit{(Eq. 3)}$$

$$\left(\frac{Fm}{Fmp}\right) - 1 = NPQ \qquad \textit{(Eq. 4)}$$

As seen in figure 2C, it takes 20-30 minutes before NPQ is relaxed and the plant is in optimal photosynthesis conditions. It is known that the decreased rate of photosynthesis due to NPQ relates to crop productivity through stomatal conductance and carbon assimilation rates (Kaiser et al., 2017; Kromdijk & Walter, 2023; Murchie & Ruban, 2020). And it has been shown that improving NPQ relaxation can have a positive effect on plant performance in fluctuating light conditions (Kromdijk & Walter, 2023).

Earlier research in improving the relaxation time of NPQ showed that up-regulating violaxanthin de-epoxidase (VDE), PSII subunit S (PsbS), and Zeaxanthin Epoxidase (ZEP) (together also known as the VPZ construct), significantly increased NPQ relaxation (De Souza et al. (2022); Kromdijk et al. (2016)). This improvement relies on improving the energy dependent quenching (qE), but there are more aspects to improving NPQ relaxation. Like improving quenching caused by state transitions (qT) and photoinhibitory quenching (qI).

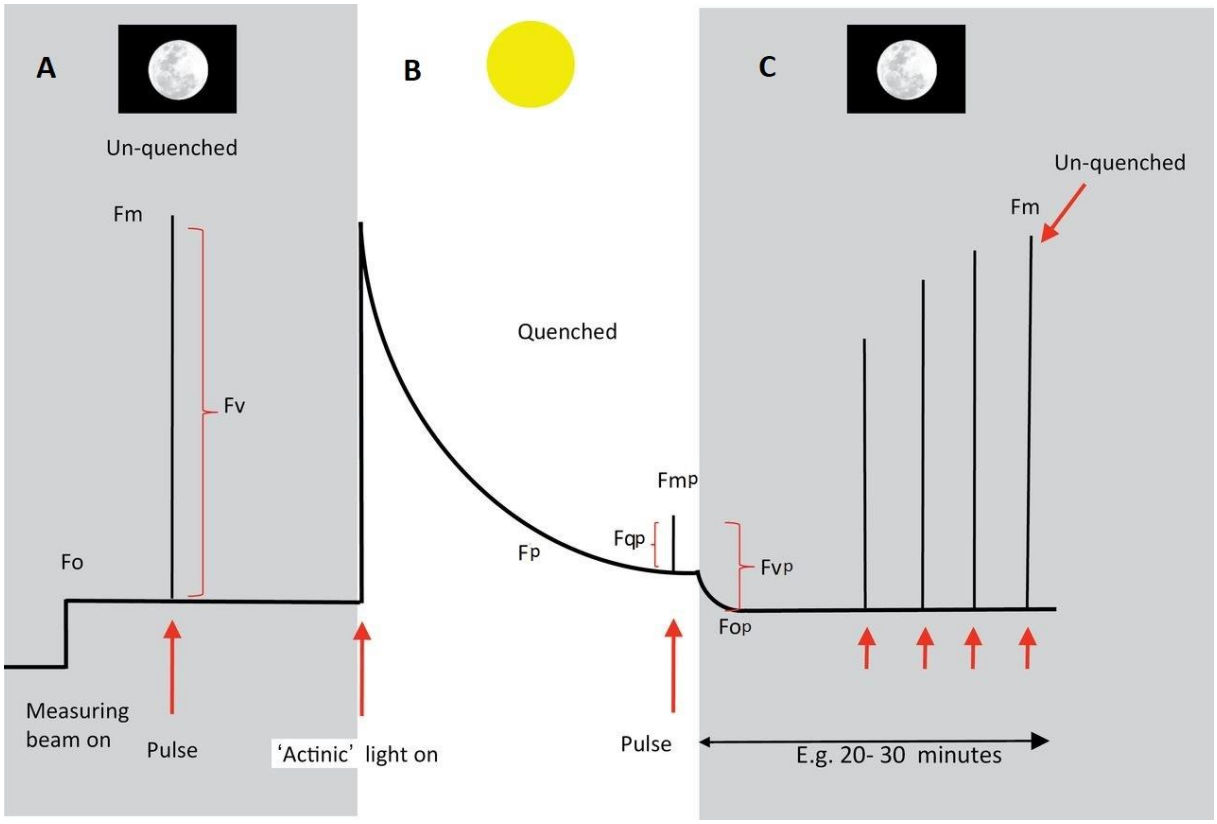


Figure 2: Graph explaining the fluorescence parameters of photosynthesis. In the dark-adapted state of a plant, a measuring beam is turned which is strong enough to elicit fluorescence, but low enough to not induce electron transport. The measured fluorescence in this state is Fo. Then a saturating pulse of light is given, closing all reaction centres. This is gives the maximum fluorescence of an un-quenched plant, Fm. The difference between Fm and Fo is Fv. Then the actinic light is turned on, resulting in a spike of fluorescence, which then quenches to a steady state fluorescence in a light adapted plant, called Fp. Again giving a saturated pulse in the light adapted state gives Fmp, the maximum fluorescence in a light adapted state. This is noticeably lower than the dark-adapted state, which is un-quenched by NPQ. After turning of the light, it takes 20-30 minutes before the plant is back in an un-quenched state. (Adapted from Murchie and Lawson (2013))

This research focuses on obtaining a better understanding of the photosynthetic efficiency and relaxation of NPQ in *Arabidopsis*, and builds upon an earlier experiment by Bos Calderó (2022). Bos Calderó (2022) identified candidate genes involved in the variation of photosynthetic efficiency after fluctuating light in the Dutch population of *Arabidopsis thaliana*, DartMap (Dutch *Arabidopsis thaliana* Map)(Wijfjes et al., 2021).

A. thaliana is a model organism for genotypic research mostly due to its relatively small chromosome size and having only 5 chromosomes. Combined with a short generation time and self-pollination, it proved to be well-suited for genetic and molecular research (Koornneef & Meinke, 2010). The genotypic data for DartMap was obtained by Wijfjes et al. (2021). Wijfjes et al. (2021) performed Genome Wide Association Studies (GWAS) using this Dutch population of *A. thaliana* to study traits such as flowering time and response of photosynthesis efficiency to iron deficiency.

GWAS connects phenotype to genotype by comparing phenotypic variance to genotypic variance within a population, with which specific genes or Quantitative Trait Loci (QTLs) can be identified. Having high genetic variance can increase effect size (measure of phenotypic difference of two alleles at a locus) of the associated genotype, however high variance also increases the chance for heterogeneity. Heterogeneity occurs when two different genetic mechanisms result in a similar phenotype. This causes the effect size to decrease, because the correlation between the phenotype and any of the variants is weakened (Korte & Farlow, 2013). Small effect size poses a problem for GWAS, because GWAS depends on the phenotypic variance explained by the genotypic variance. For this reason, rare variants are also a problem when performing genome wide association analyses. If a variant is found in only one individual in the population, many other SNPs exclusive to that individual are also associated with the trait.

Increasing sample size by having higher variance counteracts the problem of rare variants, but again increases heterogeneity. One way to counter heterogeneity is to densely sample a local population, which is exactly what Wijfjes et al. (2021) did with DartMap. Even though there is low diversity in environmental conditions in the Netherlands, they still found a high genetic variance. The mild environmental cline allows for identification of smaller differences in phenotypes.

When doing GWAS it is also important to take into account relatedness of individuals among the population (Korte et al., 2012), and population stratification (Price et al., 2010). Population stratification means that there is a systematic difference in allele frequencies between subpopulations within a population (Huang, 2022, December 02). With a LMM these factors can be taken into account.

Previous research

Previously two similar fluctuating light (FL) experiments were performed by Bos Calderó (2022) and Dr. Nguyen, where they had the phenovator II (a high-throughput phenotyping machine) measure Fv, Fm, Fv/Fm, projected leaf area (PLA), Fp and Fmp, of the Dutch *A. thaliana* population.

The experiment by Bos Calderó (2022) will be referred to as NWO22_01, and the experiment by Dr. Nguyen will be referred to as NWO22_03.

The protocol for NWO22_01 starts with the plants being treated with a period of fast fluctuating light followed by a period of slow fluctuating light (Figure 3). NWO22_03 is the exact same experiment but the two FL treatments reversed, doing the slow FL treatment first, followed by the fast FL treatment (Figure 4). The goal of these experiments is to identify QTLs associated with the acclimation of photosynthetic efficiency to fluctuations in light intensity. The reason for performing the second, reversed experiment is to confirm or deny any variation in photosynthetic efficiency due to

developmental stage of the plants, rather than the FL treatments. All conditions except the light treatments are the same for both experiments.

Aim of this research

The focus of this research is to process the data obtained from NWO22_03 and identify QTLs that share the same or similar functions as those identified in NWO22_01. The aim is to locate significant QTLs in both experiments and examine the function of the associated QTLs. QTLs found significant at the same time point in both experiments but not corresponding with the same treatment may be involved in plant development rather than fluctuating light response, whereas QTLs that correspond with the same treatments and have similar functions are likely involved in the fluctuating light response.

The leading question throughout the experiment is: Which QTLs in the *Arabidopsis thaliana* genome are involved in the acclimation of photosynthetic efficiency to fluctuating light?

To answer this question, I will try to answer the following questions:

- Which QTLs are found significant in both experiments, and do these correspond with the same treatment?
- What function do the identified QTLs have in *A. thaliana*?

My expectation is that there will be QTLs found significant for the same treatment in both experiments. However, I think there will also be many QTLs which correspond to developmental stage of the plant.

To further explore the effect of plant development and light treatment on photosynthetic efficiency, I will make use of individual leaf data. Jurado-Ruiz et al. (2022) developed an algorithm for tracking individual leaves in a phenotyping experiment. This algorithm was used during NWO22_01 to track Fv/Fm and ϕPSII for each individual leaf of a plant. This data will give higher resolution of the results because leaf age and order might influence the photosynthetic efficiency. By focusing on only select leaves a more complete image of the photosynthetic efficiency could be obtained. My expectation is that QTLs significantly associated with both the leaf data and the whole plant data of NWO22_01, are QTLs actually involved in acclimation of photosynthetic efficiency. The QTLs that are only significant in the whole plant data are likely developmental QTLs.

Materials & Method

All analyses mentioned as being done using R, were performed using R statistical Software (v4.2.2; R Core Team (2022)).

Dutch Arabidopsis population

For this experiment, the aforementioned DartMap population by Wijfjes et al. (2021) was used. From this population a total of 169 accessions was used in both experiments (appendix 1). However, both experiments did not use the same set of accessions, because some seeds got lost in between both experiments. Only two accessions are different, NWO22_01 has accession number 736 which NWO22_03 does not include, and NWO22_03 has accession number 1 which NWO22_01 does not include.

Experiment data

The light protocol for NWO22_01 is depicted in figure 3, where plants received the first 11 days after sowing (DAS) constant light (CL) at 300 µmol.m^{-2}.s^{-1} (in short µmol). Then from 12 to 15 DAS, they received fluctuating light every 15 minutes (FL15), the fast FL treatment. At 16 DAS the plants received fluctuating light every 60 minutes (FL60), the slow FL treatment, to assess if the previous fast fluctuating treatment influences the response to a slower fluctuating treatment.

Then the plants received CL for 4 days, to eliminate possible influence of the previous FL15 treatment. From 21 to 24 DAS the plants received FL60 followed by 1 day of FL15, also to assess the influence of the previous FL treatment. Finally, the plants received one more day of CL (figure 3). Both FL treatments fluctuated light intensity was between 100 and 900 µmol.

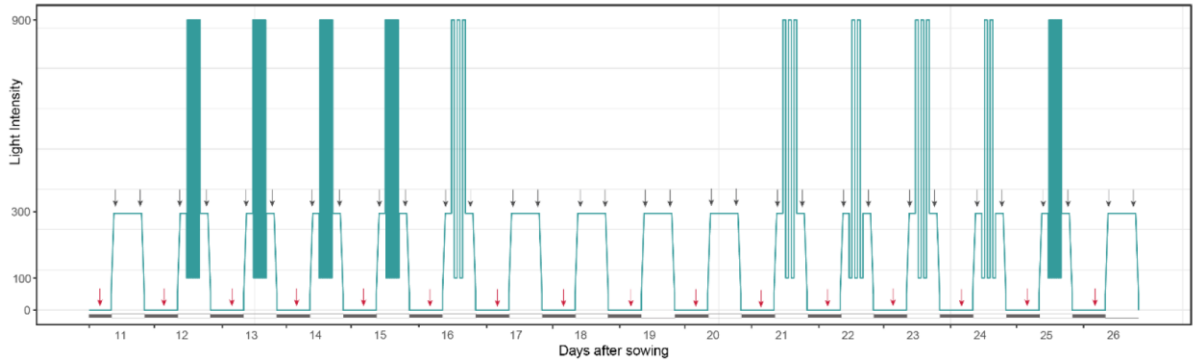


Figure 3: Light treatment protocol of experiment NWO22_01. Constant light (CL) at 300 µmol.s-1.m-2 (in short µmol), fluctuating light every 15 minutes (FL15) changing between 100 and 900 µmol and fluctuating light every 60 minutes (FL60) also between 100 and 900 µmol. Red arrows indicate measuring time point of Fv and Fm. Black arrows measuring time point of Fp and Fmp.

NWO22_03 is the exact same experiment as NWO22_01 but with the two FL treatments reversed, doing the FL60 treatment from 12 to 15 DAS, and the FL15 treatment from 21 to 24 DAS (figure 4).

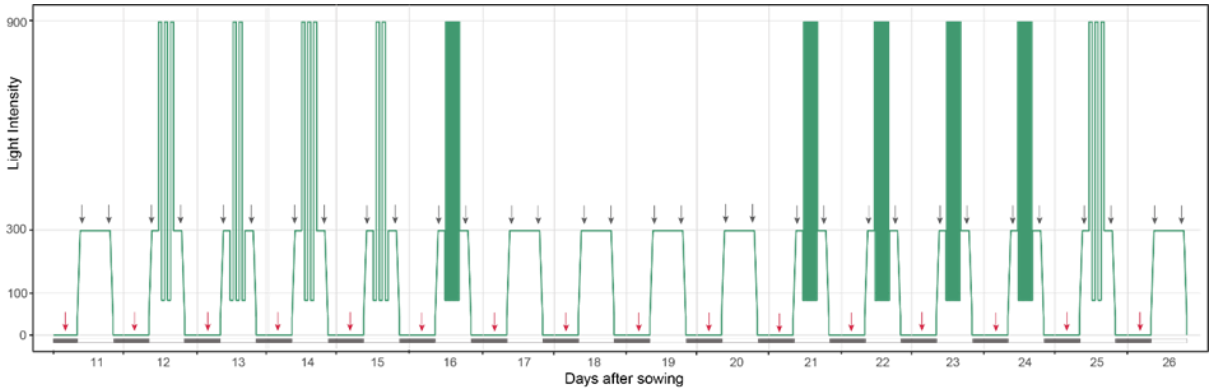


Figure 4: Light treatment protocol of experiment NWO22_03. Constant light (CL) at 300 µmol.s-1.m-2 (in short µmol), fluctuating light every 60 minutes (FL60) changing between 100 and 900 µmol and fluctuating light every 15 minutes (FL15) also between 100 and 900 µmol. Red arrows indicate measuring time point of Fv and Fm. Black arrows measuring time point of Fp and Fmp.

For both experiments, after the raw phenotypic data was obtained, some quality treatment was performed and a BLUEs (Best Linear Unbiased Estimator) analysis. For each genotype there are 8 replicates in a complete randomized block design, but instead of just taking the average of all those individuals, ignoring any variance due to other factors, like position in the climate room, the BLUE estimates the value for that time point based on the estimator with the least variance (Moser, 1996).

Individual leaf data

The algorithm by Jurado-Ruiz et al. (2022) was used on the images of NWO22_01, it tracked each leaf on a plant and the parameters Fo, Fm, Fp, and Fmp along with the PLA were measured for each individual leaf. This also means that for the leaf data only Fv/Fm and ϕPSII are included.

To account for leaf age during the experiment I needed to select true leaves (leaf 6 and 7) and retrieve the photosynthesis data (Fv/Fm and ϕPSII) of these leaves from 16 to 25 DAS. The average between these 2 leaves were used in downstream analyses.

Quality control

For the whole plant data of NWO22_01 and NWO22_03, a quality control was performed visually using excel, and by plotting the data in a boxplot. The data was conditionally formatted with a colour gradient, to detect outlying data. Accessions which data show abnormal values (very low or very high, compared to the rest), were selected as being outliers. This quality control was performed for each phenotypic trait individually, so for example in some cases the data of an accession was only removed for Fv/Fm. Bos Calderó (2022) also removed the data of two accessions, 1675 and 736, because these did not grow properly. In appendix 2 is an overview of exactly which phenotypes of which accessions were selected as outliers.

For the leaf data, quality control had to be performed for filtering out incorrect data due to the algorithm making a mistake. The mask overlay, that detects the position of an individual leaf in the image, was in many cases incorrectly placed. A mask was considered incorrect if the size of the mask was exactly the same for 4 or more timepoints in a row. Using a python script (“remove_duplicates.py”) these data points were replaced by NA. A further visual quality control in the same way as with the whole plant data was done, a colour gradient to detect abnormal data. For the leaf data the same outliers as for NWO22_01 were removed. The complete set of data used in the leaf data analysis is in appendix 3.

Correlation analysis

Before comparing the results of NWO22_01 to those of NWO22_03, a correlation analysis was performed. In this correlation analysis, the Pearson correlation between the same genotypes for each parameter at each timepoint was calculated. However, as mentioned before the experiments did not use the exact same set of genotypes, there were two genotypes unique in both experiments. These genotypes were not included in the correlation analysis, so only 168 genotypes were used. Also, the data of the first day of NWO22_03 was lost, resulting in a timepoint less for NWO22_03. For this reason, the first day of NWO22_01 also wasn't included in the correlation analysis, resulting in only using the data from 9 DAS to 26 DAS. The correlation analysis was done using R (appendix 4).

Heritability

It is useful to know whether or not the parameters Fv/Fm, ϕPSII, NPQ, ϕNPQ and ϕNO, are actually heritable traits. For this reason, the heritability is calculated for each of these traits. The broad sense heritability is the ratio between the genetic variance and the total variance in the population (Griffiths et al., 2015). The variance components, genotype, tray, blocks, and the residuals are calculated using the lme4 package (Bates et al., 2015) in R (appendix 5). Then with equation 5 the broad sense heritability for each trait at each timepoint is calculated and visualised in R (appendix 5).

$$H^2 = \frac{V_g}{V_X} \qquad \textit{(Eq. 5)}$$

Pipeline

After obtaining the data and performing quality control, each dataset will follow the same pipeline for the GWA analysis (figure 5). First genotype will be linked to phenotype using the software GEMMA. The resulting association files will be used to create QQ plots and Manhattan plots of the p-values using R. The p-values from the association files will be converted to a LOD-score by taking the -log10 of the p-value. Then to visualize the most significant windows, the LOD-scores are plotted in a heatmap. Finally, the results of NWO22_01 will be compared to the results of NWO22_03 and the individual leaf data of NWO22_01.

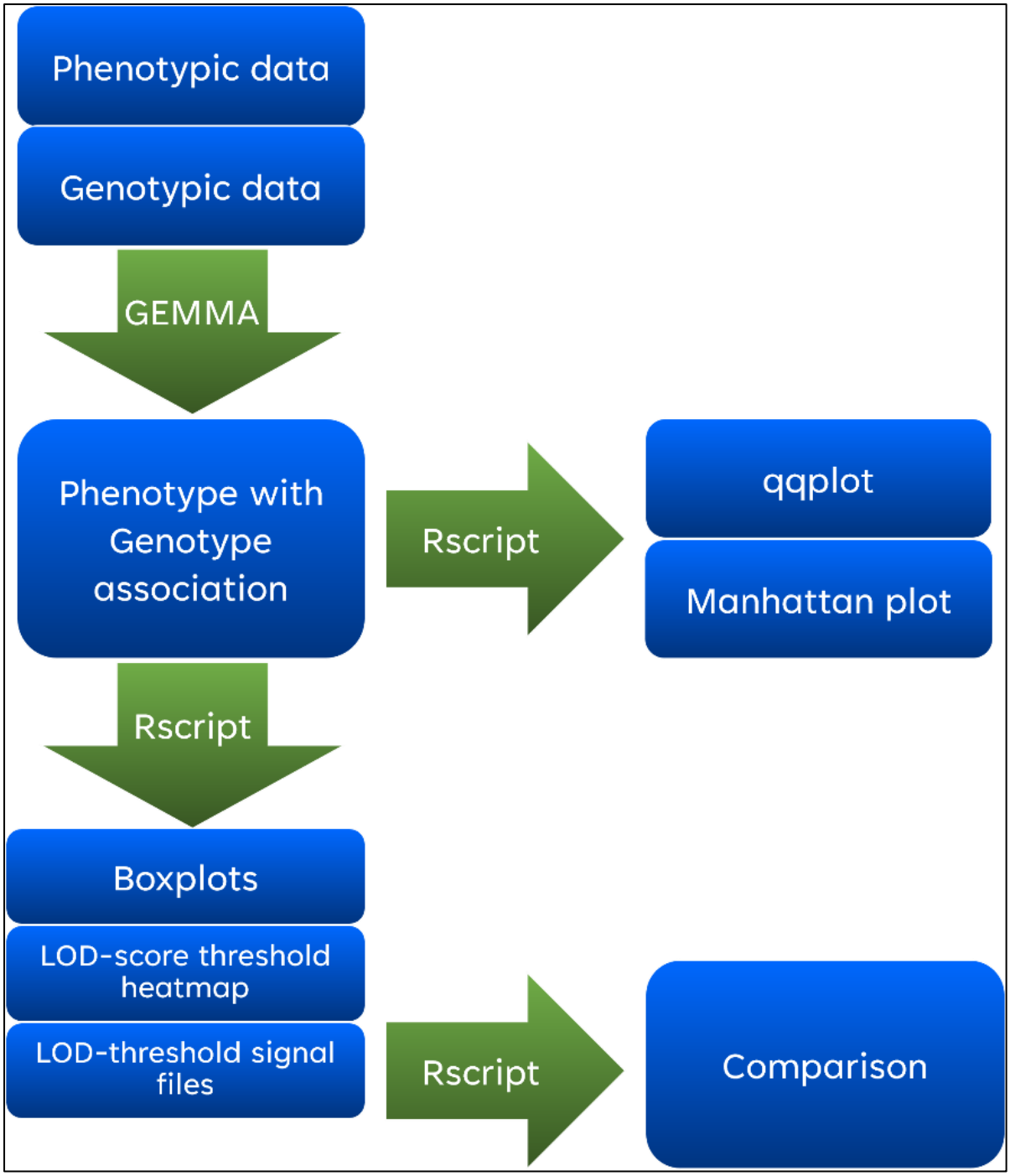


Figure 5: Pipeline for performing GWA analysis. The blue boxes are input and output for and from the programs that are used in the green arrows. The phenotypic and genotypic data are the PLINK format files required to run GEMMA, in which the .fam file contains the phenotypic information, and the .bim and .bed files the genotypic information. GEMMA produces association files, which are then used for two separate Rscript analyses. The first is analysing the results of the GWAS, by making QQ plots and Manhattan plots. In this script the association files are also filtered from very low p-values, to reduce the file size and computation time in the next analysis. The reduced association files are used to combine the individual GWAS results in a heatmap, and the phenotypic data is used to create boxplots. The heatmap shows which windows are significantly associated with one of the traits per time point. The LOD-scores per window, per time point are also saved in the signal file. The signal file of each trait is then used in a combined comparison, to show which windows are significantly associated with the whole experiment.

GEMMA

The genome-wide efficient mixed model (GEMMA) software by Zhou and Stephens (2014) was used for fitting a LMM on the genotype and phenotype data. The input for a LMM using GEMMA is a set of 4 files: two genotype files in PLINK format, having the .bim and .bed suffixes, a phenotype file also in PLINK format, having the .fam suffix, and a relatedness matrix in .cXX file format. The .bim and .bed files were provided to me by Dr. Nguyen, the .fam file was made with the phenotypic data, and the kinship matrix could be made also using GEMMA (make kinship matrix command).

To account for rare variants within the population, the minor allele frequency (maf) threshold is set at 0.05. So, alleles with a frequency lower than 5% will not come up in the results of the analysis.

To run GEMMA the following command is used as input:

```
./gemma-0.98.5-linux-static-AMD64 -bfile [prefix] -maf 0.05 -k [filename] -lmm 1 -n 1 -o [prefix]
```

Where after "-bfile" the prefix for the genotype and phenotype files is given, after "-k" the filename of the relatedness matrix, and after "-o" the prefix of the output file. However, this line would have to manually run for each phenotype column in the .fam file, which is indicated with the number after "-n". So, to speed up the process I wrote a python script that automatically ran this line for each phenotype in the .fam file (appendix 6).

The output of GEMMA is a file in which for each SNP the p-value of a Wald test is given. The Wald test is used to determine if a SNP is significantly associated with a given trait, where the p-value indicates the chance that a SNP is associated by chance (Glen, 2020).

Evaluation of GWAS results

To evaluate the results of the GWAS performed using GEMMA, QQ plots and Manhattan plots were made for each phenotype, at each timepoint. Both these plots were made using the R package qqman (Turner, 2018) (appendix 7).

QQ plots

QQ plots are a way of evaluating whether or not the resulting p-values from a GWAS are actually indicating of causal SNPs. In the QQ plot all p-values are ordered from highest to lowest value and plotted against expected p-values from a chi-squared test. In case of no causal SNPs, the observed p-values are the same as the expected p-values, which results in a straight diagonal line in the QQ plot. If the observed p-values are higher than the expected p-values, this might indicate there are causal SNPs for that phenotype. However, if the lower observed p-values already differ too much from the expected values, it is an indication that the GWAS went wrong. In the ideal situation, the QQ plot is a straight line with a tail going up in the end (Ehret, 2010; Lee & Lee, 2021).

Manhattan plots

The Manhattan plot is a visual overview of the p-values of each SNP divided over the genome. It is literally a scatterplot of the p-values, with on the x-axis the position in the genome. QTLs in the genome that are significantly associated with the trait will come up as large peaks in this plot, resembling a Manhattan skyline, because the SNPs are in linkage disequilibrium and rise above the rest together (Ehret, 2010; Slatkin, 2008).

Comparing GWAS results

To see which SNPs are significantly associated with a trait at a certain timepoint, the genome is split into windows of 50,000 base pairs (bp). Then with an R script (appendix 8) the SNP with the highest

LOD-score is selected and saved as the LOD-score for that window. With that same R script, a heatmap is made, this heatmap shows per timepoint which window in the genome has a score above the LOD-threshold of 6.0, and higher LOD-scores are presented in red. Combined with this heatmap are boxplots of the corresponding trait, showing the distribution of the phenotype at each timepoint.

The data for these heatmaps is saved in a signal file, a .csv file containing the LOD-score of each window that is above the LOD-threshold. These signal files are used to compile the data of each individual trait in another heatmap combined with a Manhattan plot, to show the significant windows for the overall experiment per trait. To compare two datasets, a similar heatmap and Manhattan plot combination is made with the compiled data of each dataset (appendix 9 & 10).

Results

Heritability of the phenotypes of experiment NWO22_03

The traits ϕPSII, NPQ, ϕNPQ and ϕNO were measured twice a day, once in the morning before the light treatment and once in the afternoon after the light treatment, and they showed clear differences in expression between those two measurements (figure 6 AB). For ϕPSII this variation between morning and afternoon is consistent throughout the experiment, but for NPQ, ϕNPQ and ϕNO the variation is greater in the beginning of the experiment, compared to the end of the experiment. The overall values for all traits, including Fv/Fm, do increase over time. The phenotypic data of NWO22_01 show a similar pattern, also increasing over time. (All boxplots are included in appendix 11, with the heatmaps)

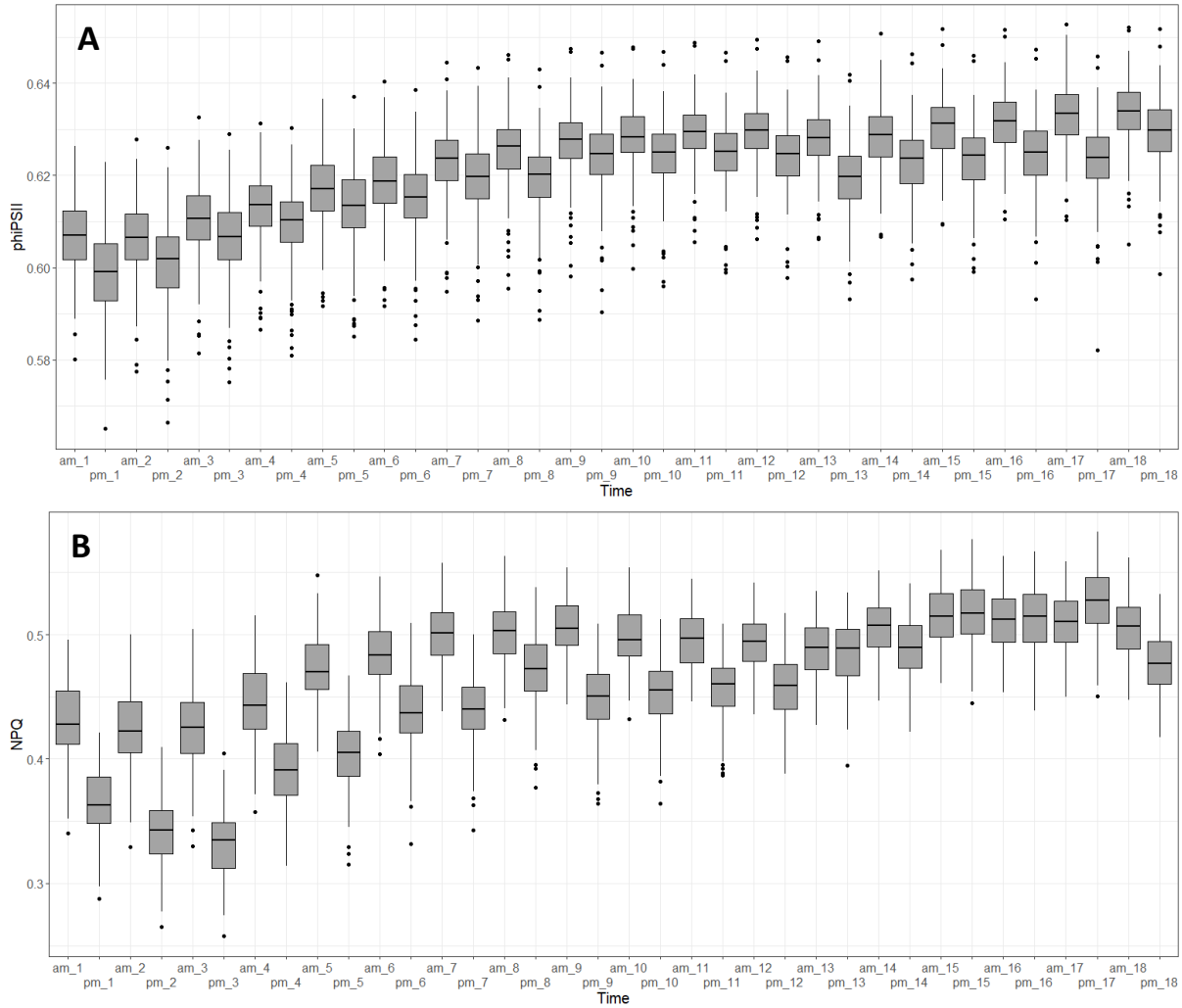


Figure 6: Boxplots of φPSII (A) and NPQ (B) measurements. Each boxplot represents measured φPSII or NPQ values of all genotypes at that timepoint. NWO22_03 started at 9 days after sowing (DAS), so timepoint am_1 represents the first measurement at 9 DAS. ΦPSII and NPQ were measured twice a day, once in the morning before the light treatment and once in the afternoon after the light treatment, represented by am and pm.

The heritability of the traits starts of quite low, but increases as the plants develop, similar to how the measured values increase (figure 7). How much each trait is heritable seems to be dependent on the height of the measured value. This is confirmed by the fluctuations in heritability between morning and afternoon measurements. For example, in figure 6B the measured values of NPQ drop in the afternoon after the FL treatment. The same drop is reflected in the heritability of NPQ, the blue line in figure 7. And similar to how this variation between morning and afternoon decreases as the experiment progresses, the difference in heritability between morning and afternoon decreases as

well. From 12 DAS, when the FL60 treatment starts, the heritability of all traits is above 30%. Meaning that at least 30% of the phenotypic variation can be explained by genetic variation.

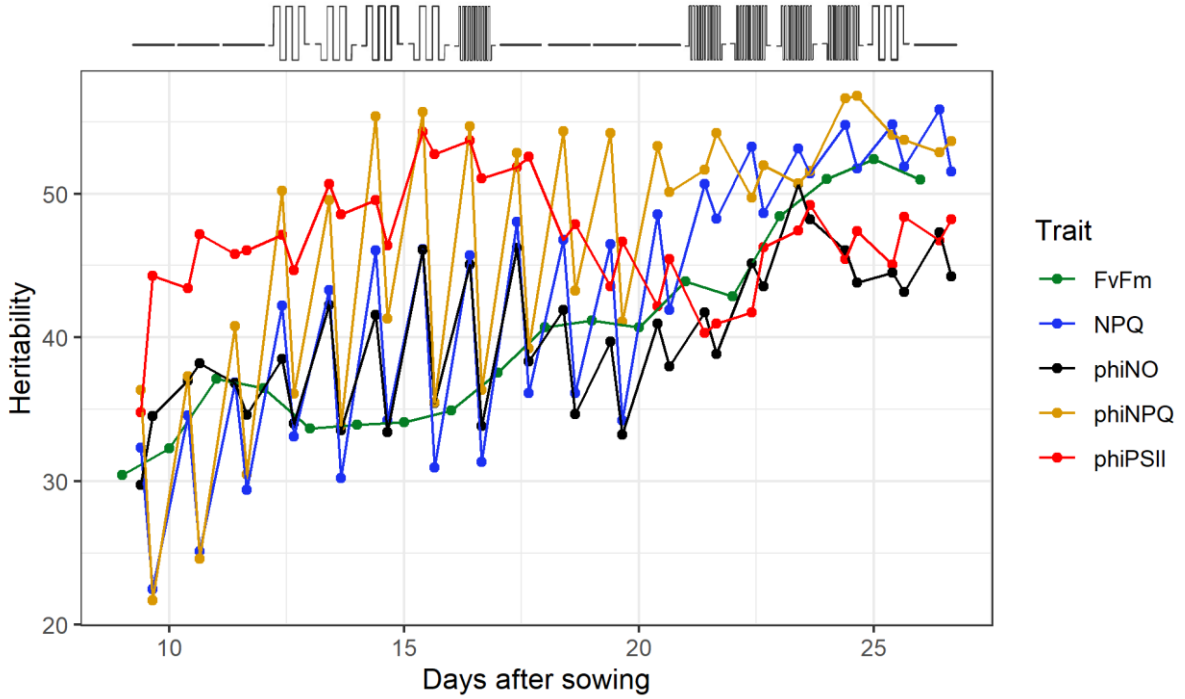


Figure 7: Heritability was calculated for each trait at each timepoint over the course of the experiment from 9 to 26 days after sowing. The traits are Fv/Fm (maximum quantum yield of photosystem II), NPQ (non-photochemical quenching), φNO (non-regulated dissipation of energy), φNPQ (regulated dissipation of energy), and φPSII (quantum yield of photosystem II). Each dot represents a measurement. For NPQ, φNO, φNPQ and φPSII there are 2 measurements per day, and for Fv/Fm there is only one measurement per day. The y-axis is heritability in percentages. The bar above the graph indicates the light regime for each day throughout the experiment. A flat line is constant light, a waving line is slow fluctuating light, and a condensed waving line is fast fluctuating light.

GWAS for phenotypes of experiment NWO22_03

To identify genetic factors underlying phenotypic variation for photosynthetic traits, genome wide association analyses were performed using GEMMA for Fv/Fm, ϕPSII, NPQ, ϕNPQ and ϕNO, on each timepoint the parameters were measured. This resulted in over 160 separate GWAS for both NWO22_01 and NWO22_03. For most of the QQ plots, that indicate whether or not the GWAS was a success, the graph did not show a tail end, meaning there were no causal SNPs. However, several QQ plots showed a perfect tail end (figure 8C), indicating there are causal SNPs found at those timepoints.

On the other hand, the Manhattan plots, which plotted the -10log of the p-values for each SNP in its position in the genome, showed a less promising result. The Manhattan plots that correspond to these nice QQ plots do not show clear peaks, but rather just a few dots shooting out (figure 8A). So only one SNP from that QTL is associated with the trait at that timepoint.

In some other cases the QQ plot deviated from the expected values quite early (figure 8D), supposedly indicating that there are a lot of significant SNPs, which is unlikely. However, looking at the Manhattan plot you can see a clear peak at one point (figure 8B). Together with a threshold, the unlikely significant SNPs can be filtered out.

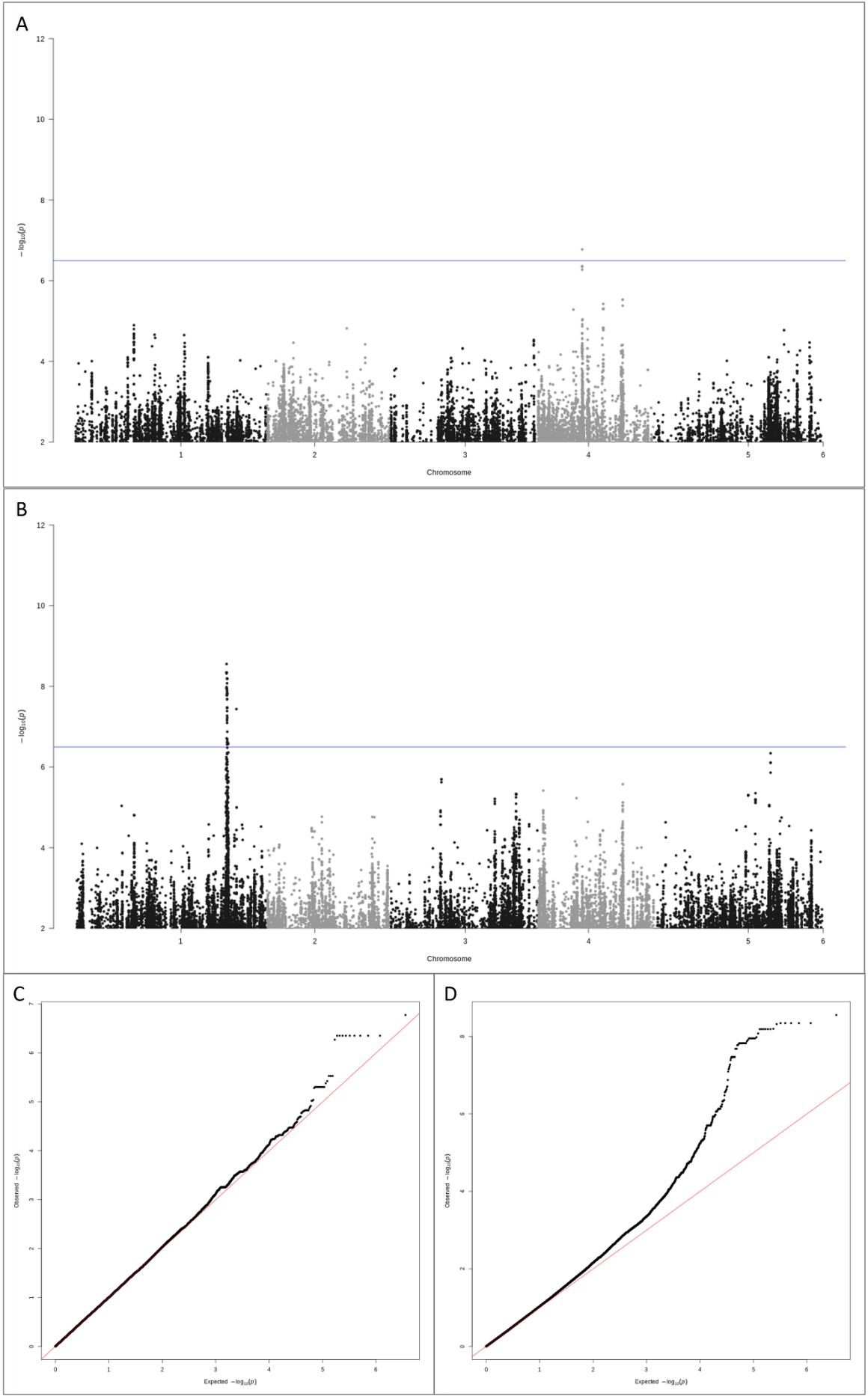


Figure 8: A,B) Manhattan plots of the GWAS result of a single timepoint in NWO22_03 and NWO22_01. In the Manhattan plot each dot represents the -10log of the p-value of each SNP included in the GWAS. They are ordered on position in the genome per chromosome. The horizontal blue line is a LOD-threshold set at 6.5. C,D) QQ plots of the same timepoints as the Manhattan plots in A and B. In the QQ plot the -10log of the p-values are ordered from lowest to highest. Because it is the -10log of the p-values, lower p-values become higher and vice versa. The observed p-values are plotted against the expected p-values. The red line indicates what the plot would look like if the observed values were the same as the expected values. If the p-values differ from the expected values it could indicate that there are causal SNPs for the phenotype.

Compiled results of QTLs for photosynthetic traits from experiment NWO22_03

The results of each separate GWAS for all traits, were combined in genome windows of 50kbp of which the significance is represented by the most significant SNP within that window. These windows were used to analyse which QTLs are significantly associated with which treatment. For ϕPSII there are some clear significant windows on chromosomes 1, 4 and 5. The windows on chromosomes 1 and 5 are clearly associated with the FL60 treatment, as they are only significantly associated on the days where the FL60 treatment was given (figure 9A). The windows on chromosome 4 are more likely associated with the FL15 treatment, because the lines showing significant windows, are only there on the later days in the experiment. Remember that during NWO22_03 from 12 to 16 DAS the FL15 treatment was given, and from 21 to 25 DAS, FL60. Interesting to note is that for NPQ, ϕNPQ and ϕNO, most windows were only found significant on the timepoints right after the FL period. You can see this in figure 9B for NPQ on 15, 16 and 17 DAS on chromosome 2.

In figure 10 the LOD-scores of each window for each trait at every timepoint is compiled and plotted in a Manhattan plot. In this plot you can see which windows are significantly associated with multiple traits. A few to point out are several windows in the back end of chromosome 1, here are several windows significantly associated with Fv/Fm, ϕNO and ϕPSII, as indicated by the larger, light blue dots in figure 10. Another one is in the back end of chromosome 4, which has a very high compiled LOD-score and is significantly associated with ϕNO and ϕPSII.

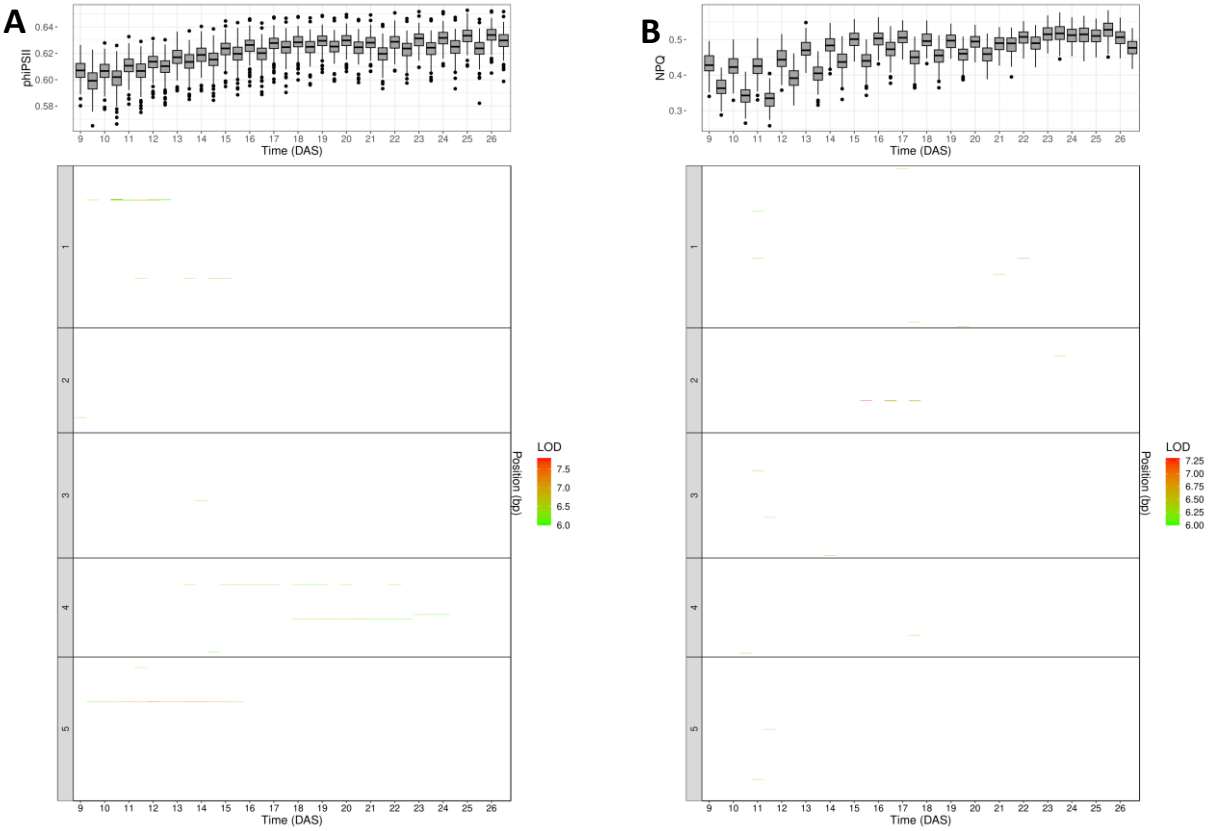


Figure 9: Boxplots and heatmaps of the GWAS results for φPSII (A) and NPQ (B). Each boxplot represents measured φPSII or NPQ values of all genotypes at that timepoint. The heatmaps plot the LOD-score of each window with a LOD-score above the threshold of 6.0 The windows are ordered in physical position on the chromosome on the x-axis.

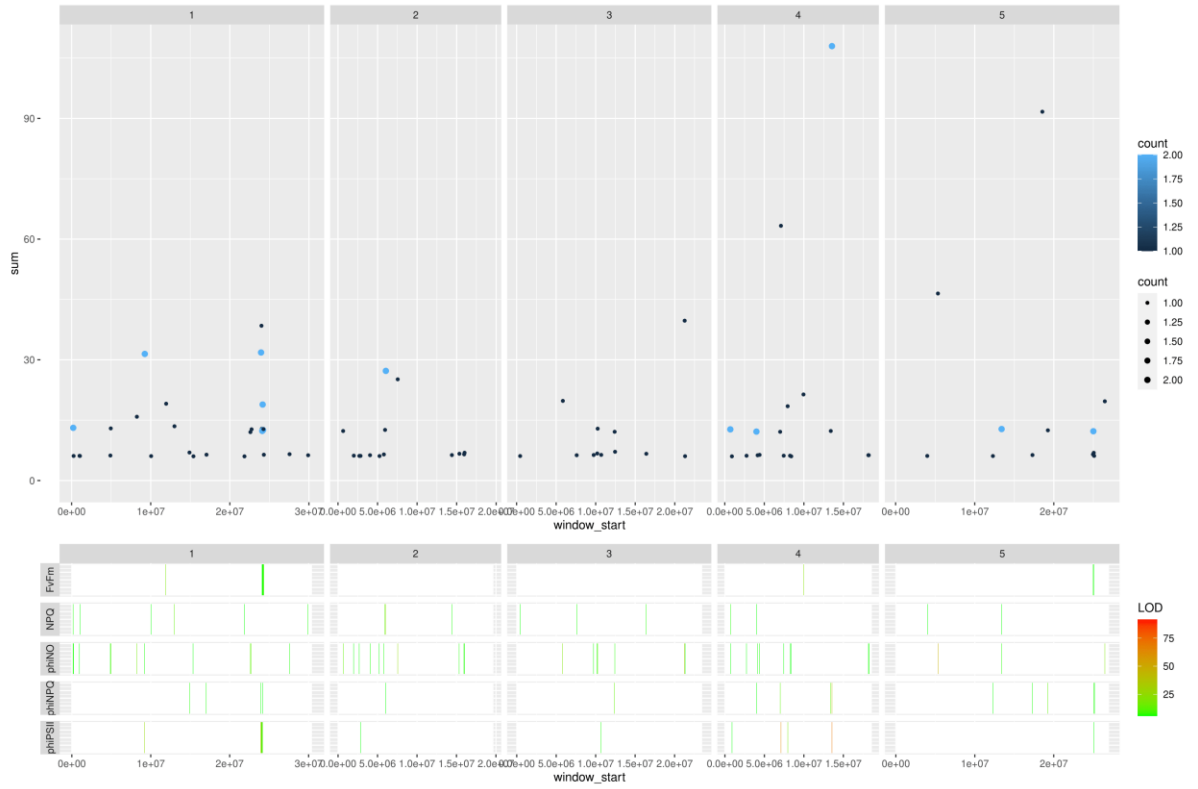


Figure 10: Manhattan plot and heatmap of the traits Fv/Fm, NPQ, φNPQ, φNO and φPSII. The Manhattan plot plots the sum of all LOD-scores above the threshold of 6.0, of that window. So, if a window has a LOD-score above 6.0 for NPQ and φNO, the blue dot is the sum of those scores. If a window is significantly associated with two traits, the dot in the Manhattan plot is coloured light blue and larger than the others. The heatmap on the bottom, plots the sum of all LOD-scores above the threshold of 6.0 for each window. The windows are ordered in physical position on the chromosome on the x-axis.

Phenotypic correlation analysis between the two experiments

Before comparing the results of NWO22_03 to the results of NWO22_01, a correlation analysis between NWO22_01 and NWO22_03 was performed to check whether or not the phenotypes are influenced by the FL treatments in the same manner. If the data did not correlate, it would not make sense to do the comparison. The correlation analysis showed a high correlation between NWO22_01 and NWO22_03, 70 percent or higher (figure 11). Interesting to note is that the correlation between the two experiments increases over time. The correlation of NPQ and ϕNPQ varies between morning and afternoon measurements, where the data is much more correlated in the morning compared to the afternoon. This variation pattern is actually the same as in the heritability analysis (figure 7), where the heritability also dropped in the afternoon. And again, in this case the correlation increases as the values of the traits increase over time.

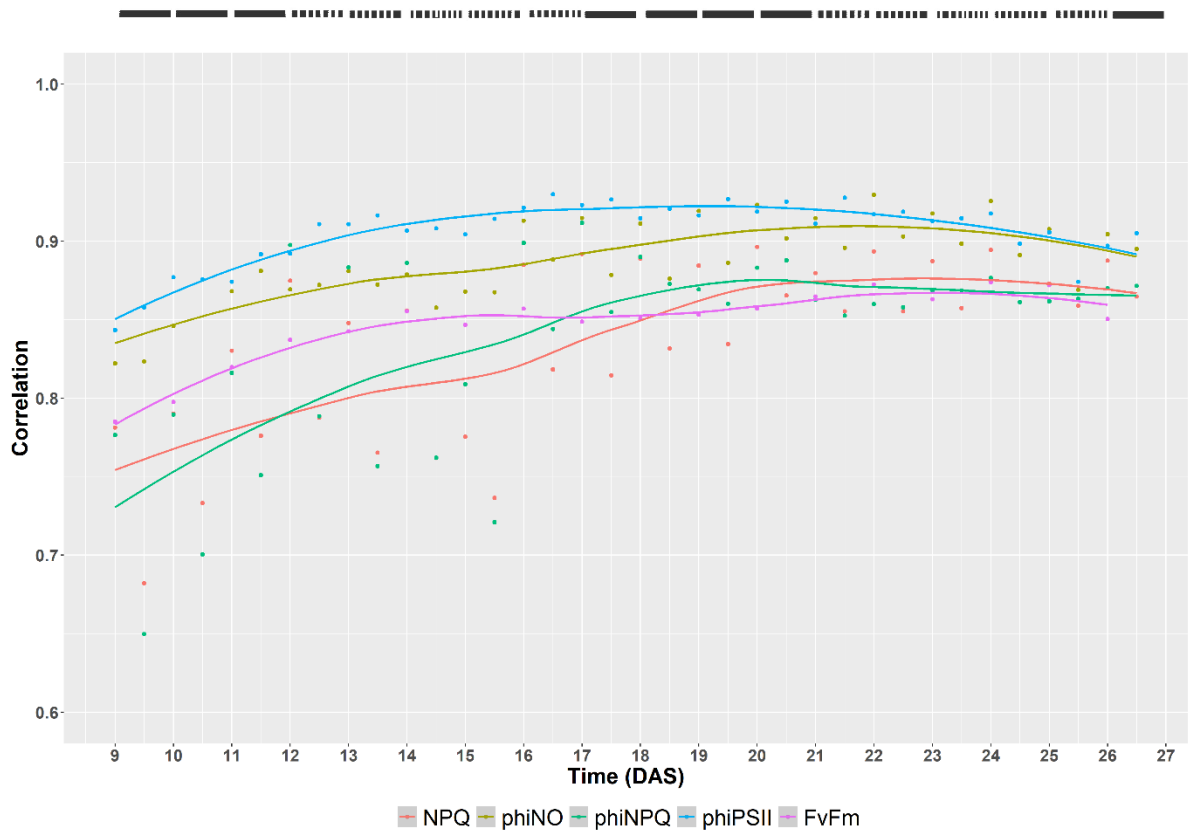


Figure 11: The correlation of the phenotypes NPQ, φNO, φNPQ, φPSII and Fv/Fm, between NWO22_01 and NWO22_03 was calculated on each timepoint in the experiment. Each dot represents a measurement. For NPQ, φNO, φNPQ and φPSII there are 2 measurements per day, and for Fv/Fm there is only one measurement per day. The graph includes a smooth line through the data points. The bar above the graph indicates whether or not both experiments received the same light treatment. A solid line means both experiments received the same light treatment, and a dotted line means the experiments received different light treatments, either FL15 or FL60.

Comparing QTLs identified in NWO22_01 and NWO22_03

The first comparison between NWO22_01 and NWO22_03, was done with all GWAS results of each experiment compiled, with no distinction between the two treatments or phenotype. This comparison resulted in quite some windows in the genome that were significantly associated with both experiments (figure 12, table 1). However, since this is all data compiled together, no distinction can be made on with what trait these QTLs are associated. This comparison mostly shows which QTLs might be involved in acclimation after periods of FL, in these experiments. In this comparison there were 19 windows significantly associated with both experiments.

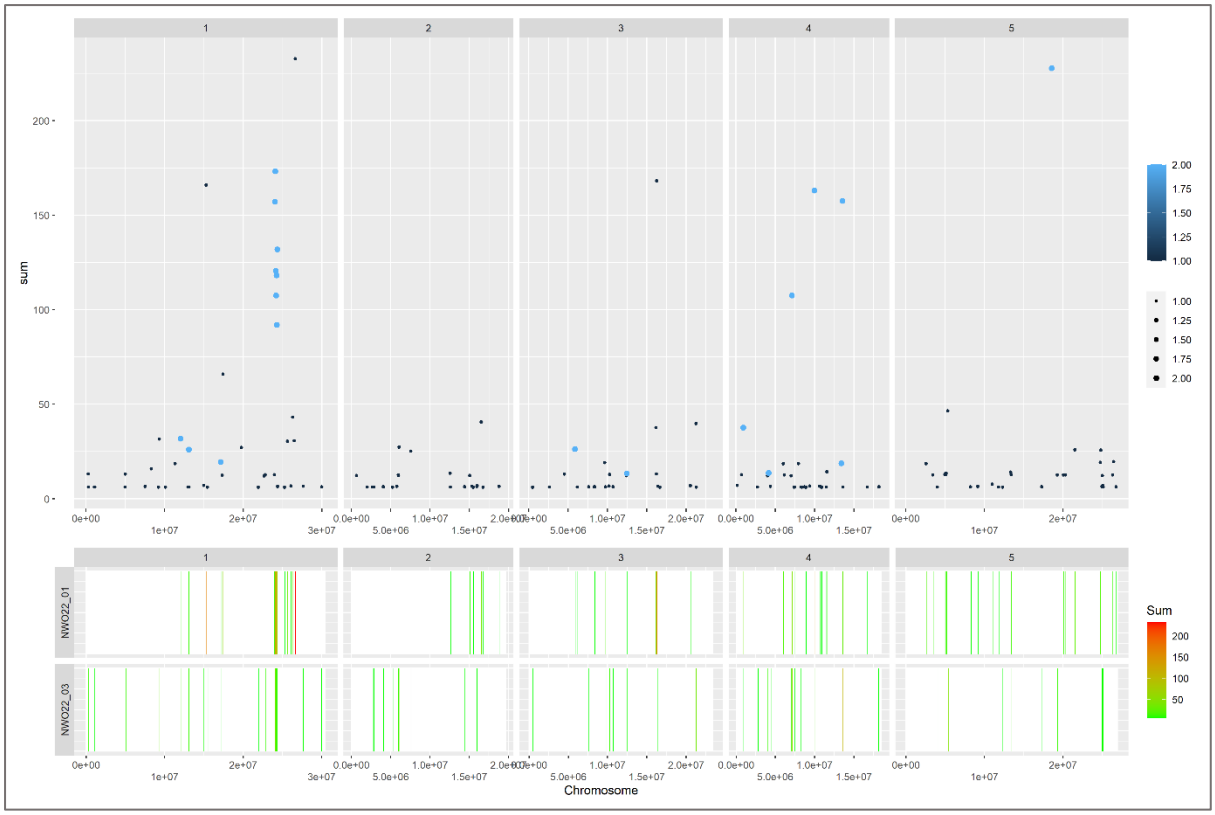


Figure 12: Manhattan plot and heatmap of the comparison between NWO22_01 and NWO22_03. Each dot in the Manhattan plot represents a window that is significantly associated with the traits φPSII, NPQ, φNO, φNPQ, and Fv/Fm, in either NWO22_01 or NWO22_03. Larger light blue dots are windows found significant in both experiments. The heatmap shows the same windows as the Manhattan plot, but separately for both experiments. The gradient is the sum of all LOD-scores above the threshold at that given window. The windows are ordered in physical position on the chromosome on the x-axis.

To detect the light treatment effect on phenotypic variation and thus underlying genetic factors, comparison of GWAS output for the phenotypes at each of the light treatments in the two experiments were compared. GWAS output from the FL15 period of NWO22_03 (21 to 24 DAS) and that from the FL15 period of NWO22_01 (12 to 15 DAS) were compiled in figure 13, showing only 3 windows found significant in both experiments under the same FL treatment.

The same comparison was done, but for FL60, so the data from 12 to 15 DAS from NWO22_01 and 21 to 24 DAS from NWO22_03. These results were compiled in figure 14 and showed 6 windows significantly associated with both experiments under the same FL treatment. Two of which, one on chromosome 1 and the other on chromosome 4, were also found significant in the FL15 comparison (table 1).

In total there were 88 windows significant for NWO22_01, and 82 for NWO22_03, 19 of which were overlapping in the compiled comparison. However, only 7 of these were actually significantly associated with both experiments when the same treatment was applied.

Additional interesting results are several windows on chromosome 4. In the overall comparison (figure 12) there were 3 windows significantly associated with both experiments, located in the middle of chromosome 4. But when cross referencing those windows with the two treatment comparisons (figures 13 &14, table 1), it appears that one of these windows is still significantly associated in both FL15 and FL60, while the other two windows are unique to one of the two treatments. However, one

of these windows is only 150kbp away from the common window. So, both these SNPs are likely linked to the same QTL.

Looking further into the windows that are significantly associated with the same treatment in both experiments will give more insight into what function these QTLs have within the plant. In the window chr1:12,000,000-12,050,000 there is the gene AT1G33230, which according to the gene ontology acts upstream or within response to light stimulus (Ashburner et al., 2000; Carbon et al., 2008; The Gene Ontology Consortium et al., 2023).

In the window chr4:9,950,000-10,000,000 a gene is found of which the knockout variant is involved in photosynthesis enhancement. But, as Medeiros et al. (2016) showed in their experiment, this change in photosynthesis can not be attributed to changes in photochemical systems. They measured Fv/Fm and ϕPSII for wildtypes and mutants of the gene, and there was no change in Fv/Fm and ϕPSII between the wildtype and mutant.

In the windows found significant on chromosome 4 and 5, genes could be found that are part of the RING/U-box superfamily protein. This superfamily of proteins is involved in the ubiquitination of proteins. In all the QTLs that were specific to one of the treatments, multiple genes with yet unknown functions could be found.

Chromosome	Window	FL15	FL60	Compiled
1	12,000,000		**X**	**X**
1	13,050,000	**X**	**X**	**X**
1	17,100,000			**X**
1	24,000,000			**X**
1	24,050,000			**X**
1	24,100,000			**X**
1	24,150,000			**X**
1	24,200,000			**X**
1	24,250,000			**X**
1	24,300,000			**X**
3	5,850,000			**X**
3	12,450,000		**X**	**X**
4	900,000			**X**
4	4,150,000			**X**
4	7,100,000			**X**
4	9,950,000		**X**	**X**
4	13,400,000	**X**		**X**
4	13,550,000	**X**	**X**	**X**
5	18,550,000		**X**	**X**

Table 1: Windows containing SNPs that are significantly associated with both NWO22_01 and NWO22_03 at LOD-threshold of 6.0, and a windowsize of 50,000. The 'Compiled' column indicates which windows are significant with all data compiled, disregarding the treatment. The columns 'FL15' and 'FL60' indicate which windows are significant in both experiments and the same treatment, fast fluctuating light (FL15) or slow fluctuating light (FL60).

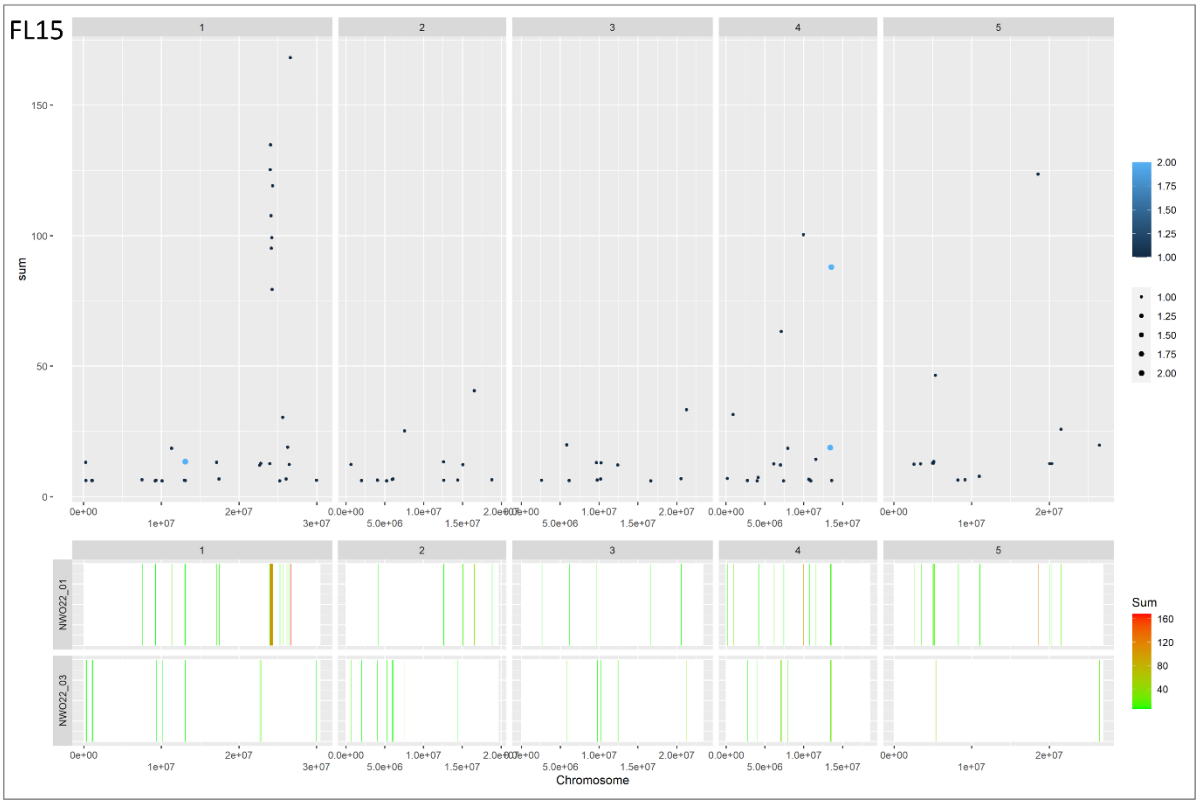


Figure 13: Manhattan plot and heatmap of the comparison between NWO22_01 and NWO22_03, but only using the data from the timepoints corresponding to the FL15 treatment. Each dot in the Manhattan plot represents a window that is significantly associated with the traits φPSII, NPQ, φNO, φNPQ, and Fv/Fm, in either NWO22_01 or NWO22_03. Larger light blue dots are windows found significant in both experiments. The heatmap shows the same windows as the Manhattan plot, but separately for both experiments. The gradient is the sum of all LOD-scores above the threshold at that given window. The windows are ordered in physical position on the chromosome on the x-axis.

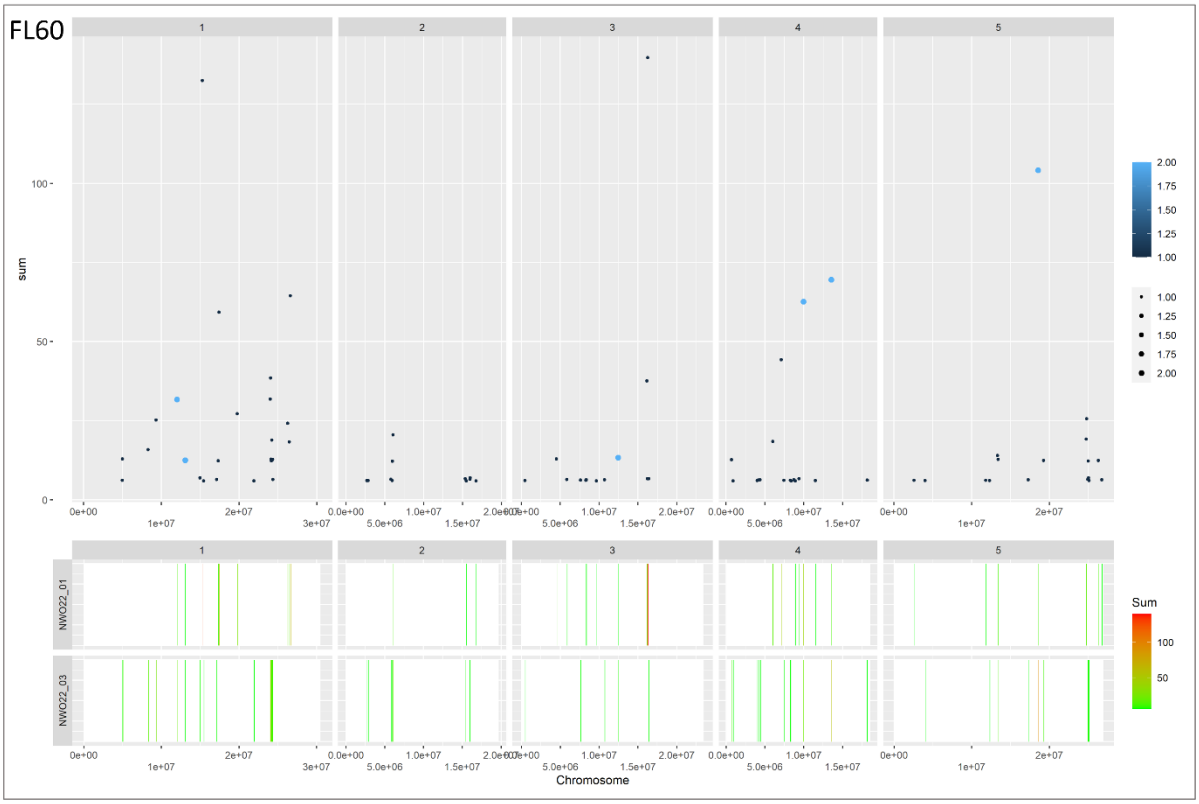


Figure 14: Manhattan plot and heatmap of the comparison between NWO22_01 and NWO22_03, but only using the data from the timepoints corresponding to the FL60 treatment. Each dot in the Manhattan plot represents a window that is significantly associated with the traits φPSII, NPQ, φNO, φNPQ, and Fv/Fm, in either NWO22_01 or NWO22_03. Larger light blue dots are windows found significant in both experiments. The heatmap shows the same windows as the Manhattan plot, but separately for both experiments. The gradient is the sum of all LOD-scores above the threshold at that given window. The windows are ordered in physical position on the chromosome on the x-axis.

Comparing individual leaf data to whole plant data of NWO22_01

The variation in plant development is acknowledged in the Dutch population: there is variation in flowering time and number of rosette leaves. Plant development is known to have an effect on photosynthesis performance. Therefore to minimize the effect of plant development on the phenotypic variation observed in the Dutch population, photosynthetic traits for individual leaves were obtained for experiment NWO22_01 (Jurado-Ruiz et al., 2022).

For the GWAS of the individual leaf data, only data corresponding to the FL60 treatment was used (16 to 25 DAS). This was because in most cases leaf 6 and 7 appeared around 15 DAS, or the leaves were too small for the algorithm to recognise the individual leaves. The final day was also removed for this analysis, because for most plants leaves 6 and 7 became overlapped nearing the end of the experiment. This analysis only included Fv/Fm and ϕPSII, since only Fv, Fm, Fp and Fmp were recorded for the leaf data.

First of all, the GWAS results of the leaf data showed much less significantly associated windows. Which is to be expected as there is much less data for the individual leaves. In total there were 59 windows significantly associated with the leaf data, 3 of which were significant for both Fv/Fm and ϕPSII (figure 15).

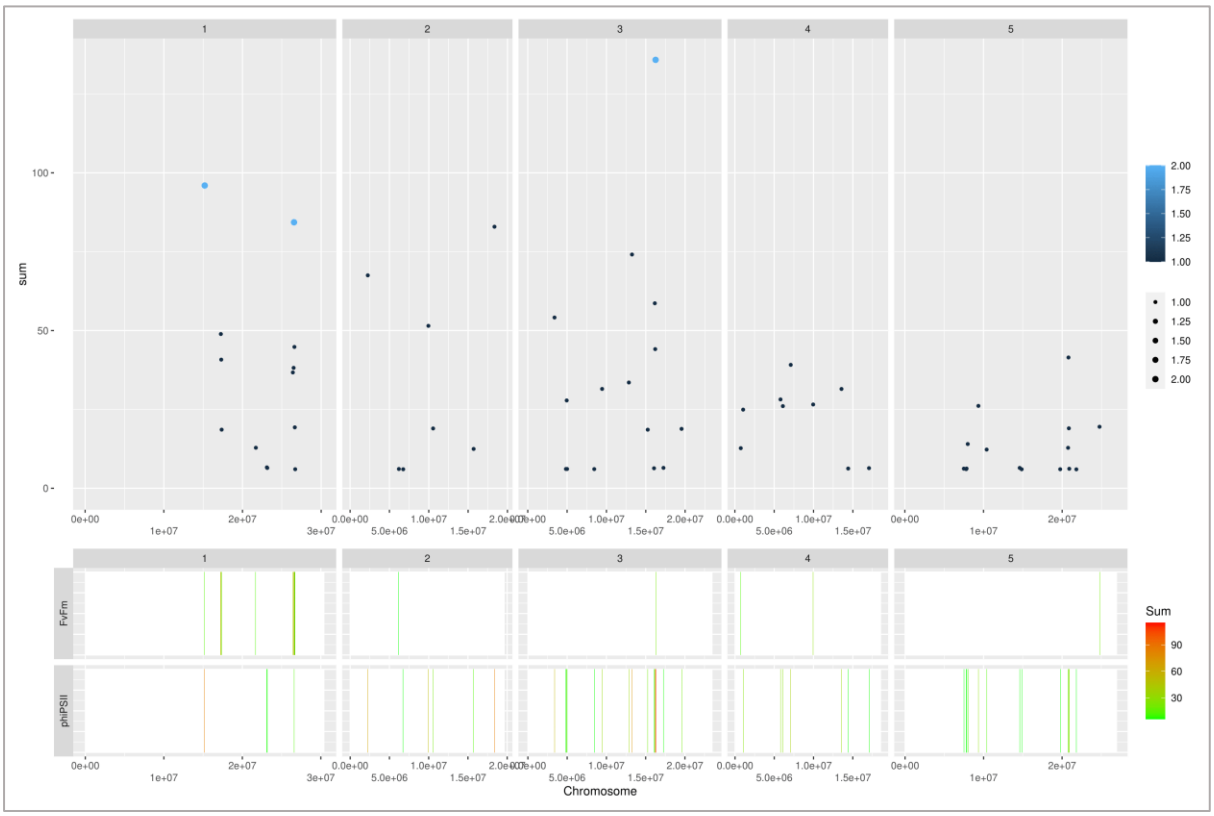


Figure 15: Manhattan plot and heatmap of the traits Fv/Fm and φPSII from the individual leaf data. The Manhattan plot, plots the sum of all LOD-scores above the threshold of 6.0, of that window. So if a window has a LOD-score above 6.0 for Fv/Fm and φPSII, the blue dot is the sum of those scores. If a window is significantly associated with both traits, the dot in the Manhattan plot is coloured light blue and larger than the others. The heatmap on the bottom, plots the sum of all LOD-scores above the threshold of 6.0 for each window. The windows are ordered in physical position on the chromosome on the x-axis.

The leaf data was compared to the overall data of NWO22_01, containing data of both FL15 and FL60, and to the data of only the FL60 part of NWO22_01 (21 to 24 DAS). Comparing the leaf data results to the FL15 results of NWO22_01 (12 to 15 DAS), resulted in the same windows being significant as the

comparison with the overall data. Further comparison of the leaf data to NWO22_01 will only be with the NWO22_01 data corresponding to the FL60 treatment.

In total there were 59 windows found significantly associated with the leaf data, 47 of which are unique to the leaf data, so 12 windows are significantly associated with both the FL60 data of NWO22_01 and the individual leaf data (figure 16). 3 of the 12 windows were also found significant in the comparison between NWO22_01 and NWO22_03. These windows were the 3 windows found significant for both datasets on chromosome 4, seen in figure 16 as the 3 light blue dots on chromosome 4. [chr4:7,100,000-7,150,000], [chr4:9,950,000-10,000,000], and [chr4:13,550,000-13,600,000]. However, the most left window [chr4:7,100,000-7,150,000] was only found significant in the overall NWO22_01 to NWO22_03 comparison, so not accounting for FL treatment, and most right window [chr4:13,550,000-13,600,000] was found significant in both the FL15 and FL60 comparison of NWO22_01 with NWO22_03. The middle window [chr4:9,950,000-10,000,000] is also significantly associated with the FL60 comparison of NWO22_1 with NWO22_03. This middle window is of interest, because the leaf data only contains data from the FL60 treatment.

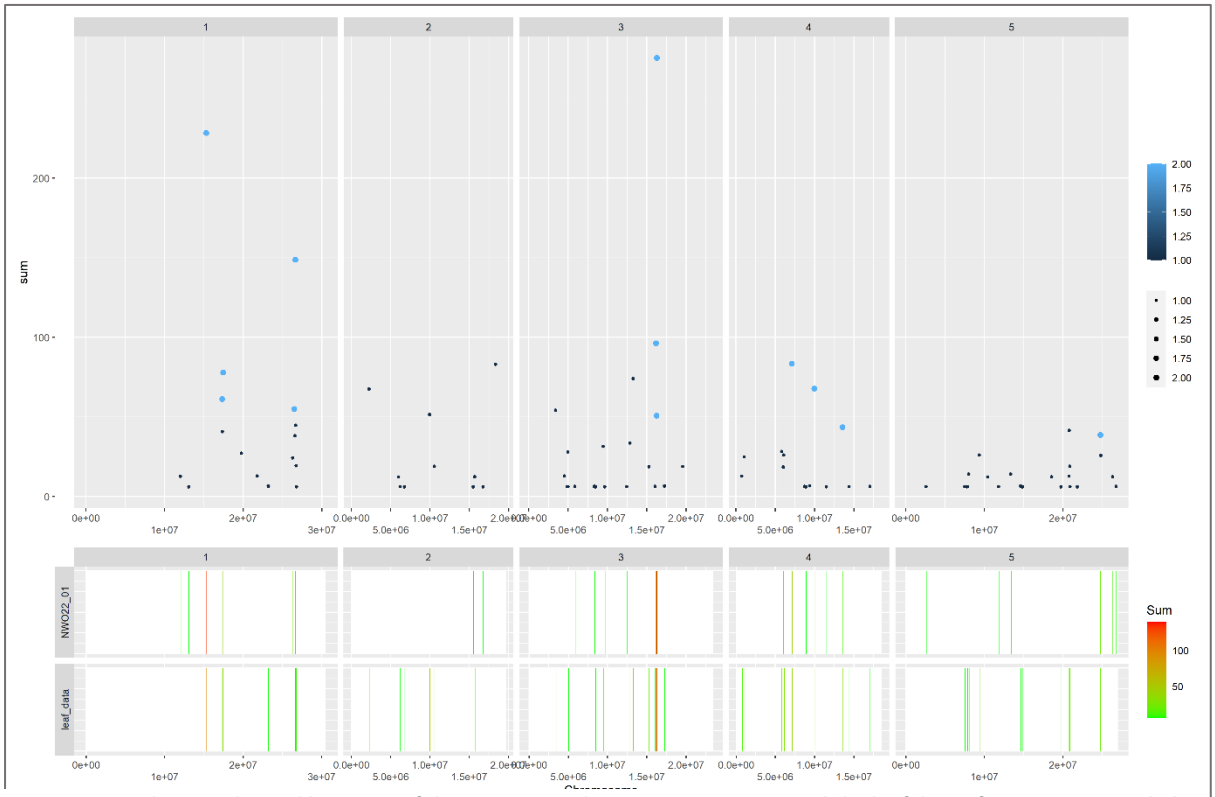


Figure 16: Manhattan plot and heatmap of the comparison between NWO22_01 and the leaf data of NWO22_01. Each dot in the Manhattan plot represents a window that is significantly associated with the traits φPSII, NPQ, φNO, φNPQ, and Fv/Fm, in either NWO22_01 or the leaf data. The heatmap shows the same windows as the Manhattan plot. The gradient is the sum of all LOD-scores above the threshold at that given window. The windows are ordered in physical position on the chromosome on the x-axis.

Discussion

From all these comparisons of GWAS data, the intention is to identify QTLs within the *Arabidopsis thaliana* genome that are involved in the acclimation of photosynthetic efficiency to fluctuating light. The goal of comparing two experiments of which the light treatments are swapped, and analysing the data of individual leaves, was to rule out or confirm any QTLs that were found to be associated with the acclimation to fluctuating light. In analysing the results of this research, a LOD significance threshold of 6.0 was used, and the window size in which the SNPs were grouped was set at 50,000 bp. The LOD-threshold is based on multiple permutation tests for several phenotypes in the DartMap population, done by René Boesten. From these permutation tests a consistent threshold between 6.3 and 6.8 resulted. Based on this, a threshold of 6.5 was set in the beginning of this research, but I did all analyses at LOD-thresholds of 6.5, 6.0 and 5.0. A threshold of 6.5 proved to be too high as almost no windows would be significant, but 5.0 was too low because there were too many windows found significant. The window size of 50kbp was decided on because in NWO22_01 the LD for some SNPs was calculated, which showed that the LD could be greater than 100kbp (Bos Calderó, 2022). With a window size of 50kbp enough resolution for these SNPs is given, but it is also not too small, so the results are easier to interpret.

From the phenotypic data (figure 6) could already be seen that development is a key factor in determining the rates of the photosynthetic traits. In both experiments the measured values increased over time, regardless of the FL treatment that was given. The importance of plant age in photosynthetic traits has already been confirmed in a previous research by Bielczynski et al. (2017), and their advise is to consider plant and leaf age when doing research on photosynthetic performance. That is why the two experiments were compared and an analysis including individual leaf data was done.

The expectation of the GWAS analysis was to find QTLs that were significantly associated with both experiments, but that most of these are associated with plant development rather than acclimation to fluctuating light. This is largely true from what we see in the results from comparing the two experiments (figures 12,13,14). There were 19 windows significantly associated for both NWO22_01 and NWO22_03 in the compiled comparison (table 1), but a comparison between treatments in the two experiments resulted in only 7 shared windows. So, 12 of these 19 windows are associated with the same developmental stage of the plants, rather than the FL treatment. Because the two datasets did correlate (figure 11), it is expected to find common genetic factors. However, that also includes QTLs involved in plant development, since it is known (Bielczynski et al., 2017), and shown in the phenotypic data (figure 6), that photosynthetic rates change during development. The heritability analysis further confirms this. In the plot (figure 7) it can be seen that heritability of the traits increases as the plants develop, meaning that in older plants that a greater part of the phenotypic variance can be explained by genetic factors (Wray; & Visscher, 2008). The younger plants also react more intense to the FL treatment, because in the beginning of both experiments there are great differences in measured trait values between morning and afternoon measurements. But this fluctuation in measured values between morning and afternoon decreases as the plants develop. The same pattern is visible in the correlation plot and the heritability plot.

So, to narrow down which QTLs are actually involved in the acclimation of photosynthetic efficiency after periods of fluctuating light, a GWAS analysis using individual leaves has been done. The vegetative state of *A. thaliana* can be divided into three sections, early juvenile stage, late juvenile stage, and adult stage (Clarke et al., 1999). So, the best choice for which leaf to track is a leaf that emerges in the adult stage. According to Clarke et al. (1999), around 18 DAS the first adult leaves emerge, which in most cases for NWO22_01 is the sixth or seventh leaf. For this reason, only the data of the sixth and seventh leaf were selected to use in the analysis. This also means that for the leaf data, only the data

from 16 to 25 DAS were used, as there was no data for most sixth and seventh leaves before 16 DAS and after 25DAS. This time frame corresponds to the FL60 treatment of NWO22_01, so the leaf data was compared to the FL60 treatment data of the whole plant.

The expectation from comparing the individual leaf data to the whole plant data of NWO22_01, was that almost all significant windows found with the leaf data, would also be significant for the whole plant data corresponding to the FL60 treatment. The windows that would be significant for the whole plant data, but not for the leaf data would then likely be genes associated with the developmental stage. However, the result shows that there is less than 21% overlap between the leaf data and whole plant. A reason for this difference could be that the leaf data is much more uniform, having only leaves that are in approximately the same developmental stage, where the whole plant data also included very young and old leaves. For example, if one accession still produces many young leaves in later stages of the experiment, then this accession likely has unique QTLs involved in plant development. The whole plant phenotype of this accession would be similar to a young plant even in later stages of the experiment, and the phenotype would be much more different than the other plants in this stage of the experiment. Then similar to the rare variants principle of GWAS (Korte & Farlow, 2013), the unique plant development QTLs of this accession will appear significantly associated with the photosynthetic traits, instead of the actual causal QTLs for photosynthetic processes. So, when the data of these young leaves is excluded from the analysis, the actual QTLs will be significantly associated. This could explain that an almost completely different set of QTLs is found significant for the individual leaf data analysis.

Genetic functions of identified QTLs

Looking at some of the genes found in the windows that were found significant for the same treatment in both experiments (table 1), there is some confirmation that these QTLs are involved in the acclimation to fluctuating light. Based on our current knowledge of the *A. thaliana* genome, most of the genes found within the QTLs are involved somewhere in the photosynthetic pathway and not necessarily acclimation after fluctuating light. The most interesting genes in these windows are the ones of which the function is yet unknown. Seeing as there is still little known on specifically the effect of fluctuating light on photosynthetic efficiency (Theeuwen et al., 2022; van Bezouw et al., 2019), these genes might be the missing links.

Another type of gene that was often found within the identified QTLs were genes that are part of a protein ubiquitination super family. Protein Ubiquitination is a post translational modification mechanism, that serves a variety of functions within cells. The most well-known function of protein ubiquitination is protein degradation (Guo et al., 2023). These genes likely came out of the GWAS analysis, because of the sudden change from low light to high light in the FL treatments. This high light probably still caused damages to the photosystems, in turn upregulating protein ubiquitination genes to repair or remove these damaged proteins.

The results from the comparison of the compiled data (figure 12), are also of interest. There are 12 out of 19 QTLs found to be significantly associated with the photosynthetic traits in both experiments, but not corresponding to the same treatment (table 1). According to my hypothesis, these genes are associated with changes in photosynthetic efficiency due to plant development rather than response to light fluctuations. This hypothesis is confirmed by one QTL in the backend of chromosome 1 [chr1:24,000,000-24,300,000]. Looking into the genes within this QTL, most of the genes are involved in the plant defence response, but in some cases there were genes that were involved in ABA signalling, which is an important signalling molecule in plant growth and development (Raghavendra et al., 2010).

Future research

In continuation of this research, I advise to perform a more targeted approach to identifying specific genes involved in the acclimation to fluctuating light. The 7 QTLs that were identified in the comparison between NWO22_01 and NWO22_03, and the 47 QTLs that were significantly associated uniquely with the individual leaf data are important to focus on. So instead of a GWAS, I would advise performing QTL characterization focusing on the aforementioned QTLs.

To identify more QTLs to focus on I would advise several things that I was not able to do in this research. The GWAS in this research was done using a univariate linear mixed model, implying that the traits Fv/Fm, ϕPSII, NPQ, ϕNPQ and ϕNO, are not influenced by each other. However, ϕPSII is very much dependant of the rate of NPQ, and NPQ is the ratio between ϕNPQ and ϕNO. So, for GWAS these traits should be taken together as variables in a multivariate linear mixed model. I also intended to do GWAS using response and recovery of the photosynthetic traits. By response I mean the ratio between the afternoon and morning measurements, and recovery the ratio between the morning and the afternoon of the previous day. These ratios might be a better representation of the acclimation, because a clear difference could be seen in figure 6 between morning and afternoon.

If another research with similar experiments will be done, I advise to increase the sample size, because 169 accessions is a rather small sample size for GWAS. That does not mean that these results are of no value, it is shown that smaller sample sizes still give significant results (Lee & Lee, 2021). The sample size of the leaf data was even smaller, because the algorithm wasn't perfect yet and not all data could be used. Also, to improve the quality of the leaf data, the used algorithm could be improved or the method of filtering faulty measurements. The method for filtering used now wasn't errorproof, because not every mask could be individually checked for quality.

Conclusion

An important conclusion from this research is that plant age has a great influence on photosynthetic traits. Seeing that of the identified QTLs, less than half of them could be involved in acclimation to fluctuating light, the rest is likely involved in plant development. So, in any future research on photosynthesis, plant age should be taken into account when designing the experiment.

Minimizing the influence of plant age on the results, by using data of individual leaves or redoing the same experiment on plants in a different developmental stage, will give better results. And the resulting QTLs that are most likely involved in acclimation of photosynthetic efficiency after periods of fluctuating light, are located on chromosomes 1, 3 and 4. On chromosome 1 these are the QTLs identified between 12,000,000 and 17,400,000, on chromosome 3 between 12,450,000 and 16,250,000, and on chromosome 4 the QTLs between 7,100,000 and 13,550,000. Especially the QTL within window 13,550,000 is an interesting QTL, since it was found for all three comparisons, FL15 and FL60 comparison between NWO22_01 and NWO22_03, and the leaf to whole plant comparison.

Acknowledgements

I want to thank Dr. Phuong Nguyen for supervising me throughout my thesis. You really helped me get outside of my data analysis bubble and look at the big picture.
René Boesten for providing feedback on my proposal and generally helping me understand the mechanisms of photosynthesis.
Prof.Dr. Mark Aarts for being my examiner and also providing feedback throughout my research.
Federico Jurado for supplying me with a very useful dataset, the individual leaf data.
And finally, the students of the laboratory of genetics, lab-rats for making working in Radix fun.

Bibliography

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25-29. https://doi.org/10.1038/75556

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*, 1-48. https://doi.org/10.18637/jss.v067.i01

Bielczynski, L. W., Łącki, M. K., Hoefnagels, I., Gambin, A., & Croce, R. (2017). Leaf and Plant Age Affects Photosynthetic Performance and Photoprotective Capacity. *Plant Physiol*, *175*(4), 1634-1648. https://doi.org/10.1104/pp.17.00904

Bos Calderó, L. (2022). *Identification of candidate genes for photosynthetic efficiency under fluctuating light* Wageningen University].

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Hub, t. A., & Group, t. W. P. W. (2008). AmiGO: online access to ontology and annotation data. *Bioinformatics*, *25*(2), 288-289. https://doi.org/10.1093/bioinformatics/btn615

Clarke, J. H., Tack, D., Findlay, K., Van Montagu, M., & Van Lijsebettens, M. (1999). The SERRATE locus controls the formation of the early juvenile leaves and phase length in Arabidopsis. *The Plant Journal*, *20*(4), 493-501. https://doi.org/https://doi.org/10.1046/j.1365-313x.1999.00623.x

De Souza, A. P., Burgess, S. J., Doran, L., Hansen, J., Manukyan, L., Maryn, N., Gotarkar, D., Leonelli, L., Niyogi, K. K., & Long, S. P. (2022). Soybean photosynthesis and crop yield are improved by accelerating recovery from photoprotection. *Science*, *377*(6608), 851-854. https://doi.org/doi:10.1126/science.adc9831

A Dictionary of Biology. (2019). (R. Hine, Ed. 8 ed.). Oxford University Press. https://doi.org/10.1093/acref/9780198821489.001.0001

Ehret, G. B. (2010). Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. *Curr Hypertens Rep*, *12*(1), 17-25. https://doi.org/10.1007/s11906-009-0086-6

Glen, S. (2020). *Wald Test: Definition, Examples, Running the Test*". Retrieved 13-06-2023 from StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/wald-test/

Green, M., Dunlop, E., Hohl-Ebinger, J., Yoshita, M., Kopidakis, N., & Hao, X. (2021). Solar cell efficiency tables (version 57). *Progress in Photovoltaics: Research and Applications*, *29*(1), 3-15. https://doi.org/https://doi.org/10.1002/pip.3371

Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). 19.3 Broad-Sense Heritability: Nature Versus Nurture. In *An Introduction to Genetic Analysis*. Macmillan Learning. https://books.google.nl/books?id=_AUnrgEACAAJ

Guo, H., Rahimi, N., & Tadi, P. (2023). Biochemistry Ubiquitination. In *StatPearls [Internet]*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK556052/

Hou, H. (2012). Manganese in Natural and Artificial Photosynthesis. In (pp. 27-50).
Huang, H. (2022, December 02). Population Stratification. In *Encyclopedia*. https://encyclopedia.pub/entry/37816
Jurado-Ruiz, F., P., N. T., Peller, J., Civit, M. J. A., Polder, G., & Aarts, M. G. M. (2022). *LeTra: A leaf tracking workflow based on convolutional neural networks and jaccard score.*
Kaiser, E., Morales, A., & Harbinson, J. (2017). Fluctuating Light Takes Crop Photosynthesis on a Rollercoaster Ride *Plant Physiology*, *176*(2), 977-989. https://doi.org/10.1104/pp.17.01250
Koornneef, M., & Meinke, D. (2010). The development of Arabidopsis as a model plant. *The Plant Journal*, *61*(6), 909-921. https://doi.org/https://doi.org/10.1111/j.1365-313X.2009.04086.x
Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*(1), 29. https://doi.org/10.1186/1746-4811-9-29
Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, *44*(9), 1066-1071. https://doi.org/10.1038/ng.2376
Kramer, D. M., Johnson, G., Kiirats, O., & Edwards, G. E. (2004). New Fluorescence Parameters for the Determination of QA Redox State and Excitation Energy Fluxes. *Photosynth Res*, *79*(2), 209. https://doi.org/10.1023/B:PRES.0000015391.99477.0d
Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K., & Long, S. P. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science*, *354*(6314), 857-861. https://doi.org/doi:10.1126/science.aai8878
Kromdijk, J., & Walter, J. (2023). *Relaxing non-photochemical quenching (NPQ) to improve photosynthesis in crops* (Burleigh Dodds Series in Agricultural Science, , Issue. B. D. S. Publishing. https://library.oapen.org/handle/20.500.12657/61520
Lee, T., & Lee, I. (2021). Genome-Wide Association Studies in Arabidopsis thaliana: Statistical Analysis and Network-Based Augmentation of Signals. In J. J. Sanchez-Serrano & J. Salinas (Eds.), *Arabidopsis Protocols* (pp. 187-210). Springer US. https://doi.org/10.1007/978-1-0716-0880-7_9
Medeiros, D. B., Martins, S. C., Cavalcanti, J. H., Daloso, D. M., Martinoia, E., Nunes-Nesi, A., DaMatta, F. M., Fernie, A. R., & Araújo, W. L. (2016). Enhanced Photosynthesis and Growth in atquac1 Knockout Mutants Are Due to Altered Organic Acid Accumulation and an Increase in Both Stomatal and Mesophyll Conductance. *Plant Physiol*, *170*(1), 86-101. https://doi.org/10.1104/pp.15.01053
Moser, B. K. (1996). 5- Least-Squares Regression. In B. K. Moser (Ed.), *Linear Models* (pp. 81-103). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012508465-9/50005-3

Müller, P., Li, X.-P., & Niyogi, K. K. (2001). Non-Photochemical Quenching. A Response to Excess Light Energy1. *Plant Physiology*, *125*(4), 1558-1566. https://doi.org/10.1104/pp.125.4.1558

Murchie, E. H., & Lawson, T. (2013). Chlorophyll fluorescence analysis: a guide to good practice and understanding some new applications. *Journal of Experimental Botany*, *64*(13), 3983-3998. https://doi.org/10.1093/jxb/ert208

Murchie, E. H., & Ruban, A. V. (2020). Dynamic non-photochemical quenching in plants: from molecular mechanism to productivity. *The Plant Journal*, *101*(4), 885-896. https://doi.org/https://doi.org/10.1111/tpj.14601

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 459-463. https://doi.org/10.1038/nrg2813

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. In (Version R version 4.2.2 (2022-10-31 ucrt)) R Foundation for Statistical Computing. https://www.R-project.org/

Raghavendra, A. S., Gonugunta, V. K., Christmann, A., & Grill, E. (2010). ABA perception and signalling. *Trends in Plant Science*, *15*(7), 395-401. https://doi.org/https://doi.org/10.1016/j.tplants.2010.04.006

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477-485. https://doi.org/10.1038/nrg2361

The Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., . . . Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1). https://doi.org/10.1093/genetics/iyad031

Theeuwen, T. P. J. M., Logie, L. L., Harbinson, J., & Aarts, M. G. M. (2022). Genetics as a key to improving crop photosynthesis. *Journal of Experimental Botany*, *73*(10), 3122-3137. https://doi.org/10.1093/jxb/erac076

Tietz, S., Hall, C. C., Cruz, J. A., & Kramer, D. M. (2017). NPQ(T): a chlorophyll fluorescence parameter for rapid estimation and imaging of non-photochemical quenching of excitons in photosystem-II-associated antenna complexes. *Plant, Cell & Environment*, *40*(8), 1243-1255. https://doi.org/https://doi.org/10.1111/pce.12924

Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software*. https://doi.org/10.21105/joss.00731

van Bezouw, R. F. H. M., Keurentjes, J. J. B., Harbinson, J., & Aarts, M. G. M. (2019). Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency. *The Plant Journal*, *97*(1), 112-133. https://doi.org/https://doi.org/10.1111/tpj.14190

Wijfjes, R. Y., Boesten, R., Becker, F. F. M., Theeuwen, T. P. J. M., Verheijen, J. J., Denkers, L. M., Koornneef, M., Eeuwijk, F. V., Smit, S., Ridder, D. D., & Aarts, M. G. M. (2021). Local adaptation of Arabidopsis thaliana in a small geographic region with mild environmental clines.

Wikimedia Commons. (2015). *Thylakoid membrane 3.svg* [File]. https://commons.wikimedia.org/wiki/File:Thylakoid_membrane_3.svg

Wray;, N., & Visscher, P. (2008). Estimating trait heritability. *Nature Education*, *1*(1), 29. https://www.nature.com/scitable/topicpage/estimating-trait-heritability-46889/

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*(4), 407-409. https://doi.org/10.1038/nmeth.2848

Zhu, X.-G., Long, S. P., & Ort, D. R. (2008). What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Current Opinion in Biotechnology*, *19*(2), 153-159. https://doi.org/https://doi.org/10.1016/j.copbio.2008.02.004

Zhu, X.-G., Long, S. P., & Ort, D. R. (2010). Improving Photosynthetic Efficiency for Greater Yield. *Annual Review of Plant Biology*, *61*(1), 235-261. https://doi.org/10.1146/annurev-arplant-042809-112206

Appendix

1) Used DartMap accessions

NWO22_01				
-	1396	1780	2206	2440
1060	1400	18	2214	2456
1070	1434	1869	2231	2458
1135	1459	1931	2244	2459
1147	1463	1947	2245	2474
1166	1466	1959	2251	2479
1206	1467	1961	2257	249
1210	1470	1964	2260	27
1212	1473	1968	2261	32
1214	1476	1970	2265	325
1217	1483	1977	2301	332
1222	1512	1979	2318	335
1223	1536	1983	2329	340
1224	1551	1988	2355	40703
1228	1587	1989	2357	557
1229	1594	1999	2360	567
1289	1612	2008	2364	687
1290	1630	2015	2372	712
1293	1637	2024	2373	72
1296	1658	2031	2375	736
1302	1672	2038	2378	764
1304	1675	2054	2381	768
1308	1683	2056	2382	774
1313	1687	2064	2385	81
1333	1689	2066	2390	816
1335	1692	2082	2391	822
1341	1696	2088	2392	826
1361	1704	2101	2393	828
1363	1716	2103	2394	837
1364	1724	2107	2396	85
1366	1726	2109	2399	88
1378	1728	2151	2400	886
1382	1774	2199	2405	890
1395	1775	2203	2406	92

NWO22_03				
1	1396	1780	2206	2440
1060	1400	18	2214	2456
1070	1434	1869	2231	2458
1135	1459	1931	2244	2459
1147	1463	1947	2245	2474
1166	1466	1959	2251	2479
1206	1467	1961	2257	249
1210	1470	1964	2260	27
1212	1473	1968	2261	32
1214	1476	1970	2265	325
1217	1483	1977	2301	332
1222	1512	1979	2318	335
1223	1536	1983	2329	340
1224	1551	1988	2355	40703
1228	1587	1989	2357	557
1229	1594	1999	2360	567
1289	1612	2008	2364	687
1290	1630	2015	2372	712
1293	1637	2024	2373	72
1296	1658	2031	2375	-
1302	1672	2038	2378	764
1304	1675	2054	2381	768
1308	1683	2056	2382	774
1313	1687	2064	2385	81
1333	1689	2066	2390	816
1335	1692	2082	2391	822
1341	1696	2088	2392	826
1361	1704	2101	2393	828
1363	1716	2103	2394	837
1364	1724	2107	2396	85
1366	1726	2109	2399	88
1378	1728	2151	2400	886
1382	1774	2199	2405	890
1395	1775	2203	2406	92

2) Selected outliers

	NWO22_01				
Accession no.	Fv/Fm	ϕPSII	NPQ	ϕNPQ	ϕNO
340	Removed	Removed	Removed	Removed	Removed
736	Removed	Removed	Removed	Removed	Removed
1166	Kept	Removed	Removed	Removed	Removed
1612	Kept	Removed	Kept	Kept	Removed
1675	Removed	Removed	Removed	Removed	Removed
2382	Removed	Kept	Kept	Kept	Kept

	NWO22_03				
Accession no.	Fv/Fm	ϕPSII	NPQ	ϕNPQ	ϕNO
340	Removed	Removed	Removed	Removed	Removed
736	Not in experiment	Not in experiment	Not in experiment	Not in experiment	Not in experiment
1166	Kept	Removed	Removed	Removed	Removed
1612	Kept	Removed	Kept	Kept	Removed
1675	Kept	Kept	Kept	Kept	Kept
2382	Removed	Kept	Kept	Kept	Kept

3) Used leaf data

Look for the file called “leaf_data_used_data.csv”

4) NWO22_03_correlation.R

```
library(Hmisc)
library(ggplot2)
library(plotly)
library(reshape2)
library(dplyr)

## Read in the data
NWO22_03_files <- "C:/Users/bwd/OneDrive - Wageningen University & Research/MSc thesis/NWO22-03_Bram/Data/Input_files"
NWO22_01_files <- "C:/Users/bwd/OneDrive - Wageningen University & Research/MSc thesis/NWO22-03_Bram/Data/Input_files"

setwd(NWO22_03_files)
NWO22_03 <- read.csv("NWO22_03_corr.txt", sep="\t", header = F)
FvFm_03 <- NWO22_03[,6:23]
```

```
phiPSII_03 <- NWO22_03[,24:59]
NPQ_03 <- NWO22_03[,60:95]
phiNO_03 <- NWO22_03[,96:131]
phiNPQ_03 <- NWO22_03[,132:167]

NWO22_01 <- read.csv("NWO22_01_corr.txt", sep="\t", header =F)
FvFm_01 <- NWO22_01[,7:24]
phiPSII_01 <- NWO22_01[,27:62]
NPQ_01 <- NWO22_01[,65:100]
phiNPQ_01 <- NWO22_01[,103:138]
phiNO_01 <- NWO22_01[,141:176]

## Make correlation matrices
## FvFm
colnames<-c(9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26)
colnames(FvFm_01)<-colnames
colnames(FvFm_03)<-colnames


cor_matrix_FvFm <- matrix(nrow = 18, ncol = 1)
rownames(cor_matrix_FvFm)<-colnames
for (i in 1:18) {
  cor_matrix_FvFm[i,]<-cor(FvFm_01[,i],FvFm_03[,i])
}
cor_matrix_FvFm <- data.frame(cor_matrix_FvFm)
cor_matrix_FvFm <- cor_matrix_FvFm %>% slice(rep(1:n(), each = 2))
cor_matrix_FvFm$cor_matrix_FvFm[c(FALSE,TRUE)] <- NA
png("FvFm_corr_Time_original.png",width = 15, height = 10, units =
'in', res = 300)
ggplot(data = cor_matrix_FvFm, aes(x=colnames,y=cor_matrix_FvFm,)) +
  geom_point() +
  geom_smooth() +
  labs(x="Time", y="Correlation", title = "Correlation FvFm") +
  ylim(c(0.65,1)) +
  scale_x_continuous(breaks = seq(1,18,1)) +
  theme(axis.text = element_text(face = "bold", size = 15),
axis.title = element_text(face = "bold", size = 20), title =
element_text(face = "bold", size = 20))
dev.off()

## phiPSII
colnames<-
c(9,9.5,10,10.5,11,11.5,12,12.5,13,13.5,14,14.5,15,15.5,16,16.5,17,1
7.5,18,18.5,19,19.5,20,20.5,21,21.5,22,22.5,23,23.5,24,24.5,25,25.5,
26,26.5)
colnames(phiPSII_01)<-colnames
colnames(phiPSII_03)<-colnames


cor_matrix_phiPSII <- matrix(nrow = 36, ncol = 1)
for (i in 1:36) {
```

```
  cor_matrix_phiPSII[i,]<-cor(phiPSII_01[,i],phiPSII_03[,i])
}
cor_matrix_phiPSII <- data.frame(cor_matrix_phiPSII)
png("phiPSII_corr_Time_original.png",width = 15, height = 10, units
= 'in', res = 300)
ggplot(data = cor_matrix_phiPSII,
aes(x=colnames,y=cor_matrix_phiPSII,)) +
  geom_point() +
  geom_smooth() +
  labs(x="Time", y="Correlation", title = "Correlation phiPSII") +
  ylim(c(0.65,1)) +
  scale_x_continuous(breaks = seq(1,36,1)) +
  theme(axis.text = element_text(face = "bold", size = 15),
axis.title = element_text(face = "bold", size = 20), title =
element_text(face = "bold", size = 20))
dev.off()

## NPQ
colnames<-
c(9,9.5,10,10.5,11,11.5,12,12.5,13,13.5,14,14.5,15,15.5,16,16.5,17,1
7.5,18,18.5,19,19.5,20,20.5,21,21.5,22,22.5,23,23.5,24,24.5,25,25.5,
26,26.5)
colnames(NPQ_01)<-colnames
colnames(NPQ_03)<-colnames

cor_matrix_NPQ <- matrix(nrow = 36, ncol = 1)
for (i in 1:36) {
  cor_matrix_NPQ[i,]<-cor(NPQ_01[,i],NPQ_03[,i])
}
cor_matrix_NPQ <- data.frame(cor_matrix_NPQ)
png("NPQ_corr_Time_original.png",width = 15, height = 10, units =
'in', res = 300)
ggplot(data = cor_matrix_NPQ, aes(x=colnames,y=cor_matrix_NPQ)) +
  geom_point() +
  geom_smooth() +
  labs(x="Time", y="Correlation", title = "Correlation NPQ") +
  ylim(c(0.65,1)) +
  scale_x_continuous(breaks = seq(1,36,1)) +
  theme(axis.text = element_text(face = "bold", size = 15),
axis.title = element_text(face = "bold", size = 20), title =
element_text(face = "bold", size = 20))
dev.off()

## phiNPQ
colnames<-
c(9,9.5,10,10.5,11,11.5,12,12.5,13,13.5,14,14.5,15,15.5,16,16.5,17,1
7.5,18,18.5,19,19.5,20,20.5,21,21.5,22,22.5,23,23.5,24,24.5,25,25.5,
26,26.5)
colnames(phiNPQ_01)<-colnames
colnames(phiNPQ_03)<-colnames
```

```
cor_matrix_phiNPQ <- matrix(nrow = 36, ncol = 1)
for (i in 1:36) {
  cor_matrix_phiNPQ[i,]<-cor(phiNPQ_01[,i],phiNPQ_03[,i])
}
cor_matrix_phiNPQ <- data.frame(cor_matrix_phiNPQ)
png("phiNPQ_corr_Time_original.png",width = 15, height = 10, units =
'in', res = 300)
ggplot(data = cor_matrix_phiNPQ,
aes(x=colnames,y=cor_matrix_phiNPQ,)) +
  geom_point() +
  geom_smooth() +
  labs(x="Time", y="Correlation", title = "Correlation phiNPQ") +
  ylim(c(0.65,1)) +
  scale_x_continuous(breaks = seq(1,36,1)) +
  theme(axis.text = element_text(face = "bold", size = 15),
axis.title = element_text(face = "bold", size = 20), title =
element_text(face = "bold", size = 20))
dev.off()

## phiNO
colnames<-
c(9,9.5,10,10.5,11,11.5,12,12.5,13,13.5,14,14.5,15,15.5,16,16.5,17,1
7.5,18,18.5,19,19.5,20,20.5,21,21.5,22,22.5,23,23.5,24,24.5,25,25.5,
26,26.5)
colnames(phiNO_01)<-colnames
colnames(phiNO_03)<-colnames

cor_matrix_phiNO <- matrix(nrow = 36, ncol = 1)
for (i in 1:36) {
  cor_matrix_phiNO[i,]<-cor(phiNO_01[,i],phiNO_03[,i])
}
cor_matrix_phiNO <- data.frame(cor_matrix_phiNO)
png("phiNO_corr_Time_original.png",width = 15, height = 10, units =
'in', res = 300)
ggplot(data = cor_matrix_phiNO, aes(x=colnames,y=cor_matrix_phiNO,))
+
  geom_point() +
  geom_smooth() +
  labs(x="Time", y="Correlation", title = "Correlation phiNO") +
  ylim(c(0.65,1)) +
  scale_x_continuous(breaks = seq(1,36,1)) +
  theme(axis.text = element_text(face = "bold", size = 15),
axis.title = element_text(face = "bold", size = 20), title =
element_text(face = "bold", size = 20))
dev.off()


## Comparison
comp_matrix <-
cbind(cor_matrix_NPQ,cor_matrix_phiNO,cor_matrix_phiNPQ,cor_matrix_p
hiPSII,cor_matrix_FvFm, colnames)
```

```
comp_matrix <- melt(comp_matrix, id.vars = "colnames")
png("corr_comp.png",width = 15, height = 10, units = 'in', res =
300)
ggplot(data = comp_matrix, aes(x = colnames, y = value,
col=variable)) +
  geom_point() +
  stat_smooth(level = 0) +
  labs(x="Time (DAS)", y="Correlation") +
  ylim(c(0.6,1)) +
  scale_x_continuous(breaks = seq(1,36,1)) +
  scale_color_discrete(labels =
c(cor_matrix_NPQ="NPQ",cor_matrix_phiNO="phiNO",cor_matrix_phiNPQ="p
hiNPQ",cor_matrix_phiPSII="phiPSII",cor_matrix_FvFm="FvFm"))+
  theme(axis.text = element_text(face = "bold", size = 15),
axis.title = element_text(face = "bold", size = 20), title =
element_text(face = "bold", size = 20), legend.position = "bottom",
legend.title = element_blank(), legend.text = element_text(size =
20))
dev.off()
```

5) Var_comp.R

```
library(lme4)

setwd("C:/Users/bwd/OneDrive - Wageningen University & Research/MSc
thesis/NWO22-03_Bram/Data/Heritability")


####################################heritability_Calculation

data <-
read.table("NWO22_03_input_for_heritability_outliers_removed.csv",
sep = ",", header = TRUE)
data$Block<- factor(data$Block)
data$X <- factor(data$X)
data$Y <- factor(data$Y)
data$Tray_ID <- factor(data$Tray_ID)
data$Genotype <- factor (data$Genotype)

varcom <- NULL
count <- 1
for(n in 8:ncol(data)){
  name <- colnames(data)[n]
  variable <- data[,n]


  fitlmer_rand <- lmer(variable ~  (1|Genotype) + (1|Tray_ID) +
(1|X) + (1|Y) +(1|Block) , data = data, REML = TRUE)
  sum <- summary(fitlmer_rand)
  out <- as.data.frame(VarCorr(fitlmer_rand))
```

```
  if (count == 1){
    header <- out$grp
    varcom <- rbind(varcom, header)
  }
  count <- count + 1
  varcom <- rbind(varcom, out$vcov)
  rownames(varcom)[count] <- name
}

write.csv(varcom,"NWO22_03_VarianceComponents_outliers_removed.csv")


##############################plotting_Heritability

library(ggplot2)
data <- read.csv("NWO22_03_Compiled_Heritability_nosize.csv",
header=T, check.names = F)
data$Trait <- as.factor(data$Trait)

# cbPalette <- c("#047d24","#02e840", "#1b30f2","#56B4E9",
"#000000", "#787777", "#d99702", "#fac348", "#ff0000","#ff5e5e")
cbPalette <- c("#047d24", "#1b30f2", "#000000", "#d99702",
"#ff0000")

p1 <- ggplot(data=data, aes(x=DAS, y=H2, fill= Trait)) +
  geom_point(aes(colour= Trait)) +
  geom_line(aes(colour= Trait)) +
  scale_colour_manual(values=cbPalette) +
  ylab("Heritability") +
  xlab("Days after sowing") +
  theme_bw()
p1

png("NWO22_03_Compiled_Heritability.png", width = 6.5, height = 3.5,
units = 'in', res = 300)
plot(p1)
dev.off()

pdf("NWO22_03_Compiled_Heritability.pdf", height = 10, width = 20,
useDingbats=FALSE)
plot(p1)
dev.off()
```

6) GEMMA loop script

```
#!/usr/bin/env python3
from sys import argv
from subprocess import run
import os.path

"""
```

```
Author: Bram Duurland
Date: 23/02/2023

Script: This script automatically runs GEMMA(1) on the commandline
for an
association test with a multiple linear mixed model. It repeats the
test for
each phenotype specified in the phenotype .fam file.

Options used in GEMMA:
    -bfile = input files in PLINK format
    -maf = Minor allele frequency threshold
    -k = kinship matrix file
    -lmm = specify frequentist analysis choice (default 1; valid
value 1-4; 1:
    Wald test; 2: likelihood ratio test; 3: score test; 4: all 1-3.)
    -n =  which column to use for phenotype. Default is 1 which
reads the sixth
    column of the .fam file
    -outdir = specify output directory path
    -o = prefix to use in the output folder

(1): Multivariate linear mixed models
Xiang Zhou and Matthew Stephens (2014). Efficient multivariate
linear mixed
model algo-rithms for genome-wide association studies. Nature
Methods. 11: 407-
409.

Update 2.0
    Added requirement to input the prefix for the output file. This
way the
    script is applicable on different phenotype datasets.

Update 2.1
    Added the option to specify the output directory path. This path
is based
    on the output prefix.
Update 2.1.1
    Prints the command before executing
Update 2.1.2
    Made it possible to specify the output directory in the initial
command. So
    an extra input is required for he output directory.
"""


def run_gemma(pheno_file_prefix, kinship_matrix, outdir_path,
              output_file_prefix, pheno_column):
    """This function runs GEMMA on the commandline with the
specified
```

```
    parameters.

    input
        pheno_file_prefix: type=string, the prefix of the .bim, .bed
and .fam
        files
        kinship_matrix: type=string, the filename or path to the
kinship matrix
        outdir_path: type=string, the path of the output directory
        output_file_prefix: type=string, the prefix to be used for
the output
        file that is created by GEMMA. Do not add the column number
of the
        analysed phenotype, that is added by the script.
        pheno_column: type=integer, the column number of the
phenotype to be
        used for the test

    output
        None
    """
    cmd = "./gemma-0.98.5-linux-static-AMD64 -bfile {} -k {} -maf
0.05 -lmm 1\
     -n {} -outdir {}/{}/ -o {}_{}"\
        .format(pheno_file_prefix, kinship_matrix, pheno_column,
outdir_path,
                output_file_prefix, output_file_prefix,
pheno_column)
    if os.path.exists("{}/{}/{}_{}.assoc.txt".format(outdir_path,

output_file_prefix,

output_file_prefix,
                                                    pheno_column)):
        return
    else:
        print(cmd)
        run(cmd, shell=True, check=True)
    return


def count_columns(input_fn):
    """This function determines from the .fam file how many
phenotypes need to
    be analysed.

    input
        input_fn: type=string, the prefix of the .fam file

    output
        columns: type=int, total number of columns with phenotype
```

```
data
        in the .fam file
    """
    with open("{}.fam".format(input_fn), "r") as file:
        line = file.read().splitlines()[0]
        columns = len(line.split(sep="\t"))-5
    return columns


def main():
    """The main function executes all of the above functions.

    input
        the input is taken from the commandline through argv
        argv[1] should be the prefix of the PLINK format files
        argv[2] should be the path to the kinship matrix file
        argv[3] should be the path of the output directory
        argv[4] should be the prefix of the output file, without the
column
        number of the phenotype.

    output
        None
        this script doesn't return anything, but the output of GEMMA
is stored
        in the 'output' folder.
    """
    try:
        pheno_prefix = argv[1]
        input_km = argv[2]
        outdirectory = argv[3]
        output_prefix = argv[4]
    except IndexError:
        exit("provide input prefix, kinship matrix file name and
output prefix"
             )

    pheno_columns = count_columns(pheno_prefix)
    for i in range(pheno_columns):
        run_gemma(pheno_prefix, input_km, outdirectory,
output_prefix, i + 1)
    return


if __name__ == "__main__":
    main()
```

7) PtChr_loop_manhattanplots.R

```
.libPaths(c('~/MSc_thesis/R_libs',.libPaths()))
library(qqman)

setwd("/lustre/BIF/nobackup/duurl001/leaf_data_new")
```

```
number<-1

for(number in 1:30){
  name <- paste("leaf_data_new_", number, ".assoc.txt", sep="")
  data <- read.table(name, sep="", header = TRUE)
  mylabs <- unique(data$chr)
  data$chr <- as.numeric(factor(data$chr, levels =  mylabs))

  new_name<- paste("plots/Manhattan_leaf_data_new_",number,".png",
sep="")
  png(new_name, height = 800, width = 1400)
  manhattan(data, chr="chr", bp="ps", snp="rs", p="p_wald",
suggestiveline = 6.5, genomewideline = FALSE,  ylim = c(2, 12))
  dev.off()

  new_name1<- paste("plots/qqPlot_leaf_data_new_",number,".png",
sep="")
  png(new_name1, height = 900, width = 900)
  qq(data$p_wald)
  dev.off()

  reduce <- subset(data, data$p_wald < 0.301 & data$chr != "Pt")
  newfilename <- paste("Reduce_leaf_data_new_", number,".assoc.csv",
sep="")
  write.csv(reduce, newfilename, row.names=FALSE)

}
```

8) Compatible_DutchPop169_Boxplot_multimapper_loop.R

```
# .libPaths(c('~/MSc_thesis/R_libs',.libPaths()))
library(reshape2)
library(ggplot2)
library(patchwork)

## load data
if(Sys.info()["user"] == "bwd"){
  raw.dir <- file.path("C:/Users/bwd/OneDrive - Wageningen
University & Research/MSc thesis/NWO22-
03_Bram/Data/BLUEs_parameters_input/")
  assoc.dir <- file.path("C:/Users/bwd/OneDrive - Wageningen
University & Research/MSc
thesis/SharedData/Bram/leaf_data/assoc_files")
  out.dir <- file.path("C:/Users/bwd/OneDrive - Wageningen
University & Research/MSc thesis/SharedData/Bram/leaf_data/plots")
}

phenotypes <- list("FvFm", "phiPSII", "NPQ", "phiNPQ", "phiNO")
start_pheno_number <- 1
for (name in phenotypes) {
```

```
  ### BOXPLOT ####
  setwd(raw.dir)
  data <- read.csv(paste("NWO22-03_BLUEs_",
name,"_outliers_removed.csv", sep=""), header = T, check.names = F)
  data <- data[data$Genotype != "Hun" & data$Genotype != "Col",]

  data_long <- melt(data, id.vars="Genotype")
  colnames(data_long) <- c("Genotype", "Time", name)

  boxplot <- ggplot(data_long, aes(x=Time, y=.data[[name]],
group=Time)) +
    geom_boxplot(fill='#A4A4A4', color="black") +
    theme_bw() + theme(text= element_text(size=15))+
    guides(x=guide_axis(n.dodge = 2))
  boxplot

  setwd(assoc.dir)

  ##########################THESE YOU CAN
CHANGE##############################
  start_chr <- 1
  sliding_size <- 50000 #25000
  number_of_traits <- ncol(data)-1
  LOD_threshold <- 6 ### This is something to consider
  column_with_sign_value <- 12

  ###########HARD CODED; DONT TOUCH BELOW
HERE#############################
  window_end <- sliding_size
  window_size <- sliding_size
  window_max <- NULL
  max_in_slide <- NULL
  sliding_window_max <- NULL
  sliding_window_collector <- NULL

  trait <- 1
  for(trait in
start_pheno_number:(start_pheno_number+number_of_traits-1)){
    filename <- paste("Reduce_leaf_data_new_", trait,".assoc.csv",
sep="")
    traitname <- paste(trait)
    data <- read.table(filename, sep = ",", header = TRUE)
    data[(column_with_sign_value+1)] <- -
log10(data[column_with_sign_value])

  ##add a row at the bottom!!
    max_chr_number <- max(as.numeric(data$chr),na.rm=T)
    data[(nrow(data)+1),1] <- max_chr_number+1
    data[(nrow(data)),3] <- 1
    data[(nrow(data)),column_with_sign_value+1] <- 0
    tail(data)
```

```
    for(i in 1:nrow(data)){
      if(data[i,1] == start_chr){
        chr <- data[i,1]
        pos <- data[i,3]
        if(pos < window_end){
          window_max <-
cbind(window_max,data[i,column_with_sign_value+1])
        } else{
          #i <- i - 1
          #print(i)
          window_start <- window_end - window_size
          window_end <- window_end + window_size
          max_in_slide <- max(window_max)
          max_in_slide_row <-
cbind(traitname,chr,window_start,max_in_slide)
          sliding_window_max <- rbind(sliding_window_max,
max_in_slide_row)
          window_max <- NULL
          window_max <-
cbind(window_max,data[i,column_with_sign_value+1])
        }
      } else {
        #print(chr)
        #End of chromosome
        window_start <- window_end - window_size
        window_end <- window_end + window_size
        max_in_slide <- max(window_max)
        max_in_slide_row <-
cbind(traitname,chr,window_start,max_in_slide)
        sliding_window_max <- rbind(sliding_window_max,
max_in_slide_row)
        window_max <- NULL
        window_max <-
cbind(window_max,data[i,(column_with_sign_value+1)])


        start_chr <- start_chr + 1
        window_end <- sliding_size
        window_size <- sliding_size
        window_max <- NULL
        max_in_slide <- NULL
      }
    }
    sliding_window_collector <-
rbind(sliding_window_collector,sliding_window_max)
    update <- paste("Finished with trait #",trait,sep="")
    print(update)
    window_end <- sliding_size
    window_size <- sliding_size
    window_max <- NULL
    max_in_slide <- NULL
```

```
    sliding_window_max <- NULL
    start_chr <- 1

  }

  setwd(out.dir)
  test <- sliding_window_collector

  mode(test) = "numeric"
  raw_LOD_data <- data.frame(test)

  write.csv(raw_LOD_data, paste("Summary_raw_LOD_", name,
"_",sliding_size,".csv", sep=""), row.names=F)
  data <- raw_LOD_data
  for(c in 1:nrow(data)){
    LOD <- data[c,4]
    if(LOD < LOD_threshold){
      data[c,4] <- NA
    }
  }

  data$chr <- factor(data$chr)
  data$traitname <- factor(data$traitname)
  data$LOD <- as.numeric(data$max_in_slide)
  data$Position <- as.numeric(data$window_start)
  write.csv(data,  paste("Summary_threshold_LOD_", name,"_",
sliding_size,".csv", sep=""), row.names=F)

  dd <- seq(min(as.numeric(as.character(data$window_start))),
max(as.numeric(as.character(data$window_start))), by = 10000000)

  qtl_plot <- ggplot(data=data, aes(x=traitname, y=window_start)) +
    geom_tile(aes(fill=LOD, width=1)) +
    facet_grid(vars(chr), scales="free_y", space= "free_y", switch =
"y") +
    theme_bw() + theme(panel.grid.major = element_blank(),
                       panel.grid.minor = element_blank(), axis.line
= element_line(colour="black"),
                       axis.text =
element_text(colour="black"),text= element_text(size=15),
                       panel.spacing = unit(0, "lines")) +
    ylab("Position (bp)") +
    xlab("Time (DAS)") +
    scale_fill_gradient(low="#22FF00", high="#FF0000", limits
=c(LOD_threshold, max(data$LOD)), na.value = NA) +
    scale_y_discrete(breaks=dd, position="right")

  qtl_plot
  plot(boxplot/qtl_plot + plot_layout(heights = c(1,5)))

  pdf(paste("Compatible_Boxplot_multimapper", name, window_size,
```

```
LOD_threshold, "leaf_data_new.pdf", sep="_"), width= 10, height=15,
useDingbats=FALSE)
  plot(boxplot/qtl_plot + plot_layout(heights = c(1,5)))
  dev.off()

  png(paste("Boxplot_multimapper", name, window_size, LOD_threshold,
"leaf_data_new.png", sep="_"), width = 10, height = 15, units =
'in', res = 300)
  plot(boxplot/qtl_plot + plot_layout(heights = c(1,5)))
  dev.off()

  #Signal averaging script
  signal_data <- reshape(data=data,
idvar=c("chr","window_start"),v.names="max_in_slide",
timevar="traitname",direction="wide")
  end <- number_of_traits+5-1
  signal_data[is.na(signal_data)] <- 0
  signal_data$count <- apply(signal_data[,5:end], 1, function(x)
length(which(x!="0")))
  head(signal_data)
  signal_data$sum <- rowSums(signal_data[,6:end])
  signal_new <- ggplot(data=signal_data, aes(x=window_start, y=sum))
+
    geom_point(aes(color=count)) +
    scale_colour_continuous(limits =c(0.1, max(signal_data$count)))+
    facet_grid(cols = vars(chr),scales = "free_x", space = "free_x",
switch="y")+
    geom_hline( yintercept = 6, linetype="dashed", color="black") +
    guides(x=guide_axis(n.dodge = 2))

  signal_new
  File_Output_name <-
paste("Select_Promising_Peak_",name,sliding_size,".pdf", sep="")
  pdf(File_Output_name, height = 10, width = 20, useDingbats=FALSE)
  signal_new
  dev.off()
  signal_new
  signal_data <- subset(signal_data, select = -c(LOD))
  write.csv(signal_data, paste("Signal_",name,"_",LOD_threshold,"_",
sliding_size,"_leaf_data_new.csv", sep=""), row.names = F)
  start_pheno_number <- start_pheno_number + number_of_traits
}
```

9) Compiled_turbo_multimapper.R

```
library(reshape2)
library(ggplot2)
library(patchwork)

setwd("C:/Users/bwd/OneDrive - Wageningen University & Research/MSc
thesis/SharedData/Bram/NWO22_01_NLpop169_outliers_removed/NWO22_01_6
.0_50000")
```

```
########################THESE YOU CAN
CHANGE#############################
start_chr <- 1
sliding_size <- 50000
start_pheno_number <- 1
number_of_traits <- 2
LOD_treshold <- 6

##########HARD CODED; DONT TOUCH BELOW
HERE############################

window_end <- sliding_size
window_size <- sliding_size
window_max <- NULL
max_in_slide <- NULL
sliding_window_max <- NULL
sliding_window_collector <- NULL


##NWO22_01
# 1=fvfm
# 2=NPQ
# 3=phiNO
# 4=phiNPQ
# 5=phiPSII
# 6=sizefvfm


for(trait in
start_pheno_number:(start_pheno_number+number_of_traits-1)){
  filename <- paste("Signal_",
trait,"_",LOD_treshold,"_",sliding_size,"_NWO22_01_FL60.csv",
sep="")
  traitname <- paste(trait)
  data <- read.table(filename, header = TRUE, sep=",")
  data <- data[-2386:-nrow(data),]
  column_with_sign_value <- ncol(data)
  data[(column_with_sign_value+1)] <-
1*(data[column_with_sign_value])
  ##add a row at the bottom!!
  max_chr_number <- max(as.numeric(data$chr),na.rm=T)
  data[(nrow(data)+1),1] <- max_chr_number+1
  data[(nrow(data)),3] <- 1
  data[(nrow(data)),column_with_sign_value+1] <- 0
  tail(data)
  i <- 1
  for(i in 1:nrow(data)){
    if(data[i,1] == start_chr){
      chr <- data[i,1]
      pos <- data[i,3]
      if(pos < window_end){
```

```
        window_max <-
cbind(window_max,data[i,column_with_sign_value+1])
      } else{
        #i <- i - 1
        #print(i)
        window_start <- window_end - window_size
        window_end <- window_end + window_size
        max_in_slide <- max(window_max)
        max_in_slide_row <-
cbind(traitname,chr,window_start,max_in_slide)
        sliding_window_max <- rbind(sliding_window_max,
max_in_slide_row)
        window_max <- NULL
        window_max <-
cbind(window_max,data[i,column_with_sign_value+1])
      }
    } else {
      #print(chr)
      #End of chromosome
      window_start <- window_end - window_size
      window_end <- window_end + window_size
      max_in_slide <- max(window_max)
      max_in_slide_row <-
cbind(traitname,chr,window_start,max_in_slide)
      sliding_window_max <- rbind(sliding_window_max,
max_in_slide_row)
      window_max <- NULL
      window_max <-
cbind(window_max,data[i,(column_with_sign_value+1)])


      start_chr <- start_chr + 1
      window_end <- sliding_size
      window_size <- sliding_size
      window_max <- NULL
      max_in_slide <- NULL
    }
  }
  sliding_window_collector <-
rbind(sliding_window_collector,sliding_window_max)
  update <- paste("Finished with trait #",trait,sep="")
  print(update)
  window_end <- sliding_size
  window_size <- sliding_size
  window_max <- NULL
  max_in_slide <- NULL
  sliding_window_max <- NULL
  start_chr <- 1

}
test <- sliding_window_collector
```

```
mode(test) = "numeric"
raw_LOD_data <- data.frame(test)

write.csv(raw_LOD_data, "Summary_LOD.csv")
data <- subset(raw_LOD_data, select = -c(X))

for(c in 1:nrow(data)){
  LOD <- data[c,4]
  if(LOD < LOD_treshold){
    data[c,4] <- NA
  }
}

data$chr <- factor(data$chr)
data$trait <- factor(data$traitname, labels =
c("FvFm","NPQ","phiNO","phiNPQ","phiPSII"))
data$LOD <- as.numeric(data$max_in_slide)
data$Position <- as.numeric(data$window_start)
write.csv(data, "Summary_LOD.csv")
limit_max <- max(data$LOD,na.rm=T)
new <- ggplot(data=data, aes(x=window_start, y=1, fill = LOD)) +
  facet_grid(cols = vars(chr), rows= vars(trait),scales = "free_x",
space = "free_x", switch="y") +
  scale_fill_gradient(low="green", high="red", limits =
c(LOD_treshold,limit_max), space = "Lab", na.value = "white", name =
"Sum") +
  geom_raster() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        strip.text.y = element_text(angle = 180))+
  guides(x=guide_axis(n.dodge = 2)) +
  xlab("Chromosome")
new
height <- 1*number_of_traits
File_Output_name <-
paste("Output_NWO22_01_NLpop169_outliers_removed_FL60_",LOD_treshold
,"_",sliding_size,".pdf", sep="")
pdf(File_Output_name, height = height, width = 20,
useDingbats=FALSE)
new
dev.off()

#Signal averaging script
signal_data <- reshape(data=data,
idvar=c("chr","window_start"),v.names="max_in_slide",
timevar="traitname",direction="wide")
end <- number_of_traits+6-1
signal_data[is.na(signal_data)] <- 0
signal_data$count <- apply(signal_data[,6:end], 1, function(x)
```

```
length(which(x!="0")))
head(signal_data)
signal_data$sum <- rowSums(signal_data[,6:end])
signal_new <- ggplot(data=signal_data, aes(x=window_start, y=sum)) +
  geom_point(aes(color=count, size=count)) +
  scale_colour_continuous(limits =c(1, max(signal_data$count)))+
  scale_size(range=c(1,2), limits = c(1, max(signal_data$count))) +
  facet_grid(cols = vars(chr),scales = "free_x", space = "free_x",
switch="y")+
  guides(x=guide_axis(n.dodge = 2))+
  theme(axis.title.x = element_blank(),
        legend.title = element_blank())
signal_new

File_Output_name <-
paste("Compiled_Select_Promising_Peak_NWO22_01_NLpop169_outliers_rem
oved_FL60_",LOD_treshold,"_",sliding_size,".pdf", sep="")
pdf(File_Output_name, height = 10, width = 15, useDingbats=FALSE)
signal_new/new+plot_layout(heights = c(5,2))
dev.off()
File_Output_name <-
paste("Compiled_Select_Promising_Peak_NWO22_01_NLpop169_outliers_rem
oved_FL60_",LOD_treshold,"_",sliding_size,".png", sep="")
png(File_Output_name, height = 10, width = 15,units = "in", res =
300)
signal_new/new+plot_layout(heights = c(5,2))
dev.off()
signal_new
signal_data <- subset(signal_data, select = -c(LOD))
write.csv(signal_data,"Compiled_Signal_6_50000_NWO22_01_NLpop169_out
liers_removed_FL60.csv", row.names = F)
```

10)Compiled_turbo_multimapper_comparison.R

```
library(reshape2)
library(ggplot2)
library(patchwork)

setwd("C:/Users/bwd/OneDrive - Wageningen University & Research/MSc
thesis/Comparison/6.0_50000")

#########################THESE YOU CAN
CHANGE##############################
start_chr <- 1
sliding_size <- 50000
start_pheno_number <- 1
number_of_traits <- 2
LOD_treshold <- 6

###########HARD CODED; DONT TOUCH BELOW
HERE############################

window_end <- sliding_size
```

```
window_size <- sliding_size
window_max <- NULL
max_in_slide <- NULL
sliding_window_max <- NULL
sliding_window_collector <- NULL

for(trait in
start_pheno_number:(start_pheno_number+number_of_traits-1)){
  filename <-
paste("Compiled_Signal_",LOD_treshold,"_",sliding_size,"_",trait,"_F
L60.csv", sep="")
  traitname <- paste(trait)
  data <- read.table(filename, header = TRUE, sep=",")
  column_with_sign_value <- ncol(data)
  data[(column_with_sign_value+1)] <-
1*(data[column_with_sign_value])
  ##add a row at the bottom!!
  max_chr_number <- max(as.numeric(data$chr),na.rm=T)
  data[(nrow(data)+1),1] <- max_chr_number+1
  data[(nrow(data)),3] <- 1
  data[(nrow(data)),column_with_sign_value+1] <- 0
  tail(data)
  i <- 1
  for(i in 1:nrow(data)){
    if(data[i,1] == start_chr){
      chr <- data[i,1]
      pos <- data[i,3]
      if(pos < window_end){
        window_max <-
cbind(window_max,data[i,column_with_sign_value+1])
      } else{
        #i <- i - 1
        #print(i)
        window_start <- window_end - window_size
        window_end <- window_end + window_size
        max_in_slide <- max(window_max)
        max_in_slide_row <-
cbind(traitname,chr,window_start,max_in_slide)
        sliding_window_max <- rbind(sliding_window_max,
max_in_slide_row)
        window_max <- NULL
        window_max <-
cbind(window_max,data[i,column_with_sign_value+1])
      }
    } else {
      #print(chr)
      #End of chromosome
      window_start <- window_end - window_size
      window_end <- window_end + window_size
      max_in_slide <- max(window_max)
      max_in_slide_row <-
```

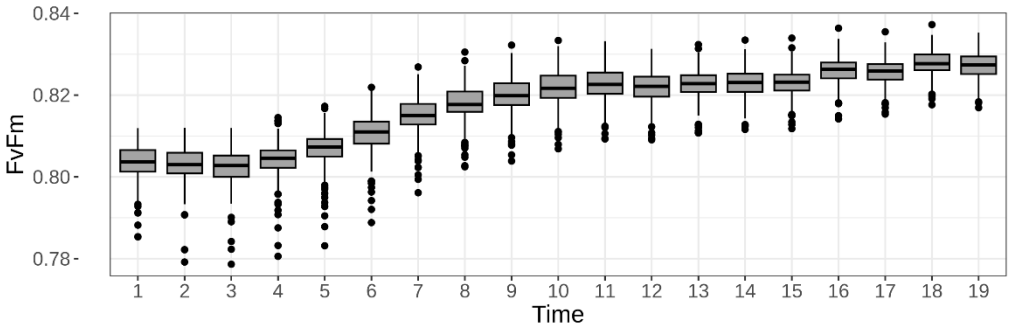
```
cbind(traitname,chr,window_start,max_in_slide)
      sliding_window_max <- rbind(sliding_window_max,
max_in_slide_row)
      window_max <- NULL
      window_max <-
cbind(window_max,data[i,(column_with_sign_value+1)])


      start_chr <- start_chr + 1
      window_end <- sliding_size
      window_size <- sliding_size
      window_max <- NULL
      max_in_slide <- NULL
    }
  }
  sliding_window_collector <-
rbind(sliding_window_collector,sliding_window_max)
  update <- paste("Finished with trait #",trait,sep="")
  print(update)
  window_end <- sliding_size
  window_size <- sliding_size
  window_max <- NULL
  max_in_slide <- NULL
  sliding_window_max <- NULL
  start_chr <- 1

}
test <- sliding_window_collector

mode(test) = "numeric"
raw_LOD_data <- data.frame(test)

write.csv(raw_LOD_data, "Summary_LOD.csv")
data <- raw_LOD_data

for(c in 1:nrow(data)){
  LOD <- data[c,4]
  if(LOD < LOD_treshold){
    data[c,4] <- NA
  }
}

data$chr <- factor(data$chr)
data$trait <- factor(data$traitname, labels = c("NWO22_01",
"leaf_data"))
data$LOD <- as.numeric(data$max_in_slide)
data$Position <- as.numeric(data$window_start)
write.csv(data, "Summary_LOD_comp_leaf_01_FL60.csv")
data <- read.csv("Summary_LOD_comp.csv")
data <- subset(data, select = -c(X))
limit_max <- max(data$LOD,na.rm=T)
```

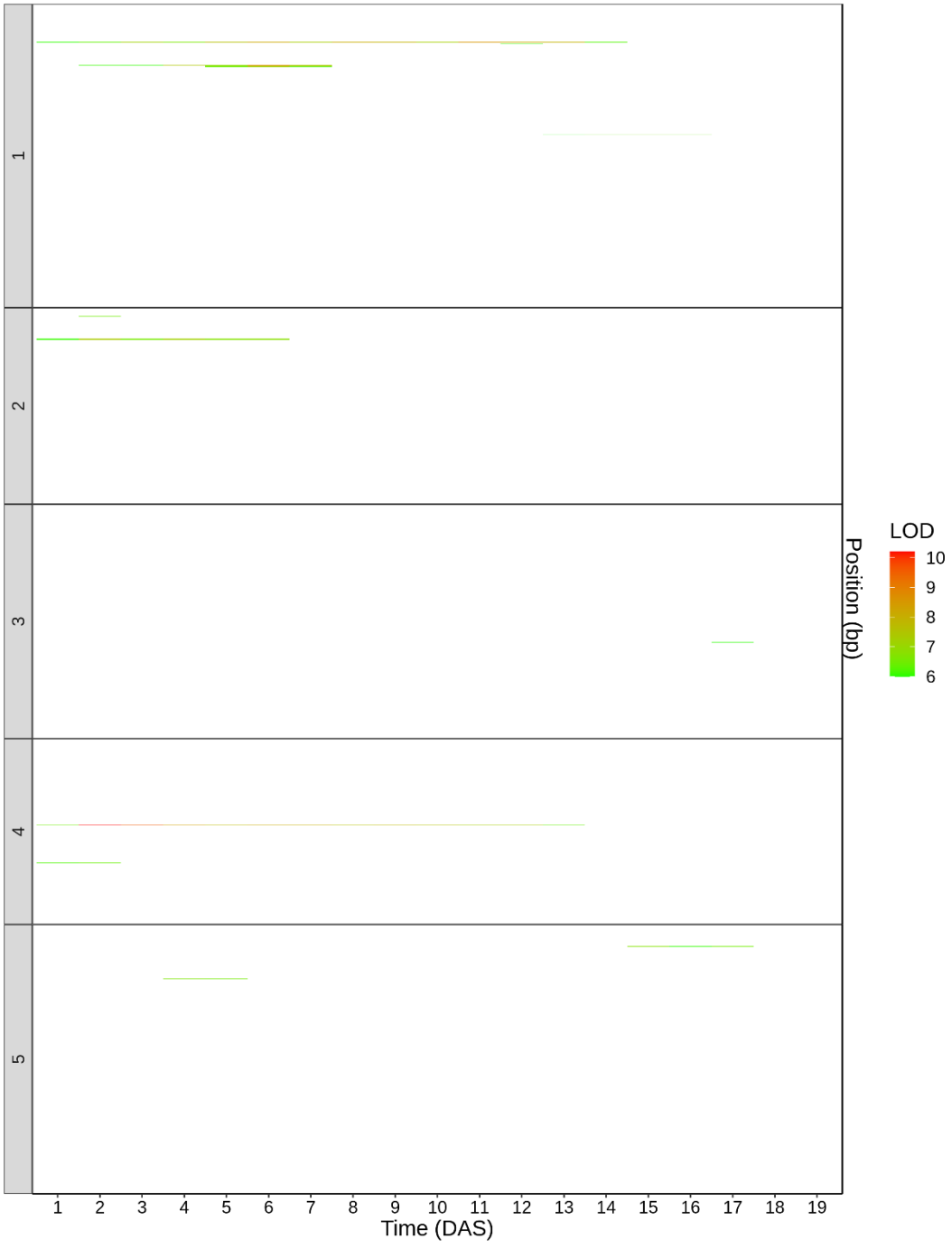
```
new <- ggplot(data=data, aes(x=window_start, y=1, fill = LOD)) +
  facet_grid(cols = vars(chr), rows= vars(trait),scales = "free_x",
space = "free_x", switch="y") +
  scale_fill_gradient(low="green", high="red", limits =
c(LOD_treshold,limit_max), space = "Lab", na.value = "white", name =
"Sum") +
  geom_raster() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        strip.text.y = element_text(angle = 180)) +
  guides(x=guide_axis(n.dodge = 2)) +
  xlab("Chromosome")
new
height <- 1*number_of_traits
File_Output_name <-
paste("Output_",sliding_size,"_comp_leaf_01_FL15.pdf", sep="")
pdf(File_Output_name, height = height, width = 20,
useDingbats=FALSE)
new
dev.off()

#Signal averaging script
signal_data <- reshape(data=data,
idvar=c("chr","window_start"),v.names="max_in_slide",
timevar="traitname",direction="wide")
end <- number_of_traits+6-1
signal_data[is.na(signal_data)] <- 0
signal_data$count <- apply(signal_data[,6:end], 1, function(x)
length(which(x!="0")))
head(signal_data)
signal_data$sum <- rowSums(signal_data[,6:end])
signal_data$sum[signal_data$sum == 0] <- NA
signal_new <- ggplot(data=signal_data, aes(x=window_start, y=sum)) +
  geom_point(aes(color=count, size=count)) +
  scale_colour_continuous(limits =c(1, max(signal_data$count)))+
  scale_size(range=c(1,2), limits = c(1, max(signal_data$count))) +
  facet_grid(cols = vars(chr),scales = "free_x", space = "free_x",
switch="y") +
  guides(x=guide_axis(n.dodge = 2))+
  theme(axis.title.x = element_blank(),
        legend.title = element_blank())
signal_new

File_Output_name <-
paste("Compiled_Select_Promising_Peak_comp_leaf_01_FL60_count_",LOD_
treshold,"_",sliding_size,".pdf", sep="")
pdf(File_Output_name, height = 10, width = 15, useDingbats=FALSE)
signal_new/new+plot_layout(heights = c(5,2))
dev.off()
File_Output_name <-
```

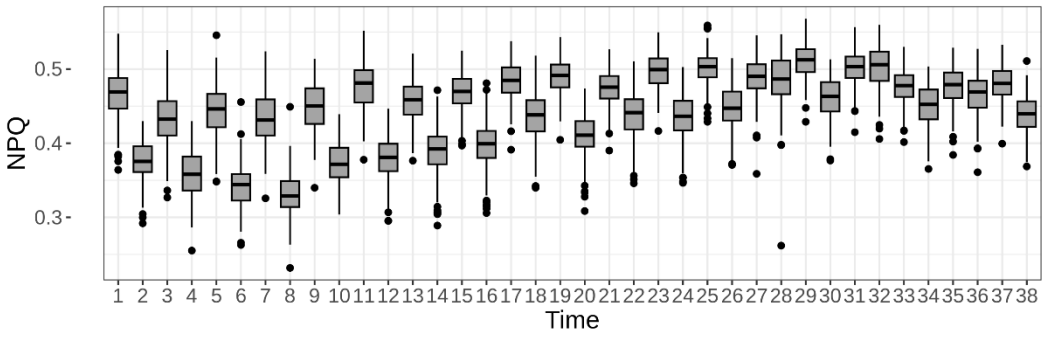
```
paste("Compiled_Select_Promising_Peak_comp_leaf_01_FL60_count_",LOD_
treshold,"_",sliding_size,".png", sep="")
png(File_Output_name, height = 10, width = 15,units = "in", res =
300)
signal_new/new+plot_layout(heights = c(5,2))
dev.off()
signal_new
write.csv(signal_data,paste("Compiled_Signal_comp_leaf_01_FL60_count
_",LOD_treshold,"_",sliding_size,".csv",sep=""))
```

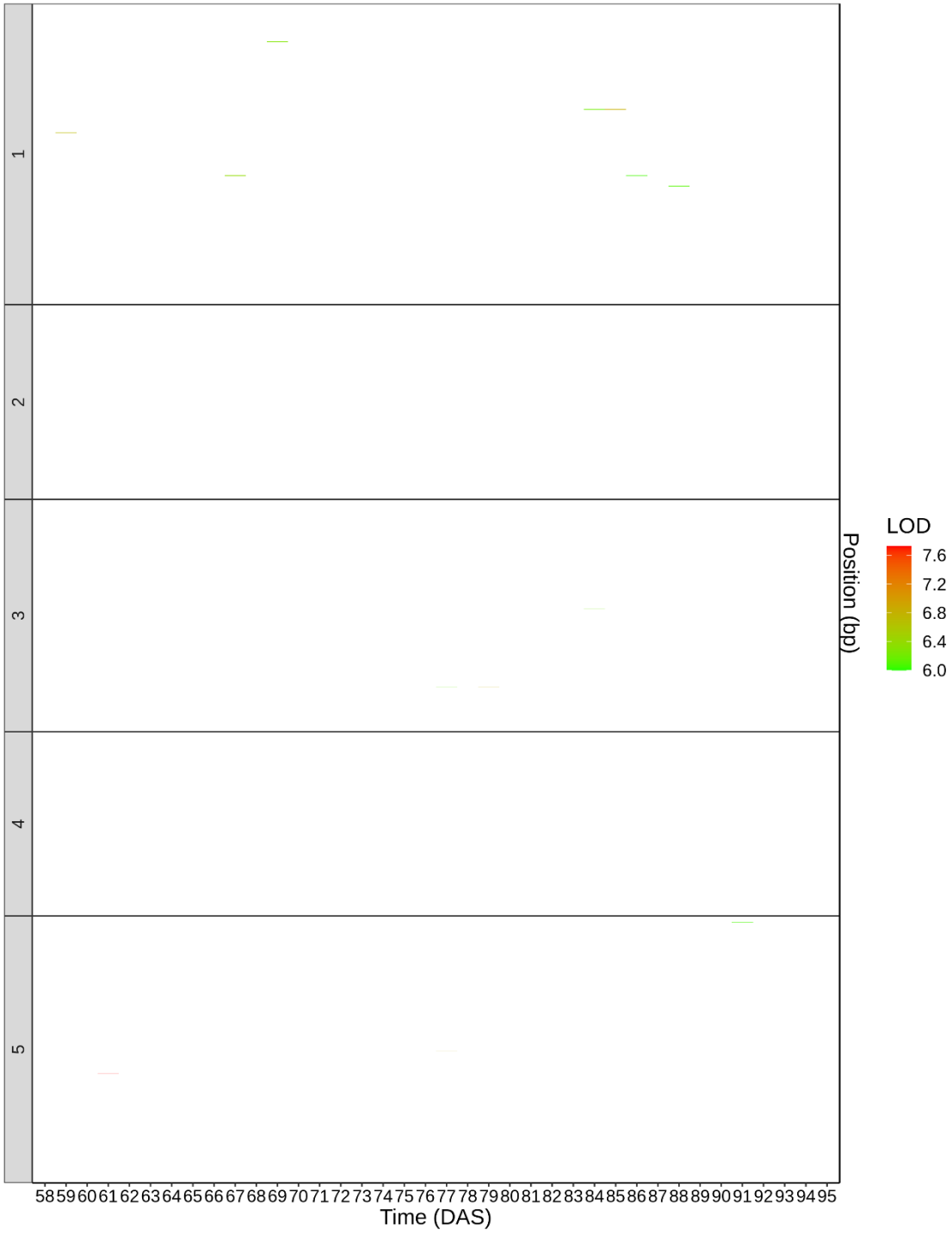
11) Boxplots and heatmaps

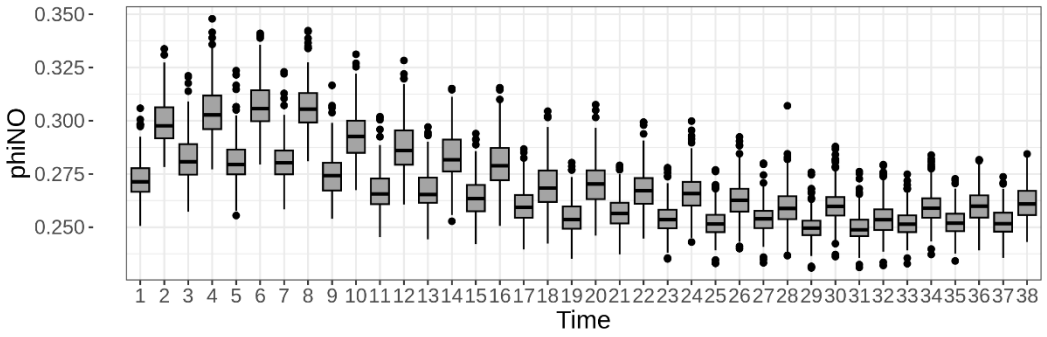
NWO22_01

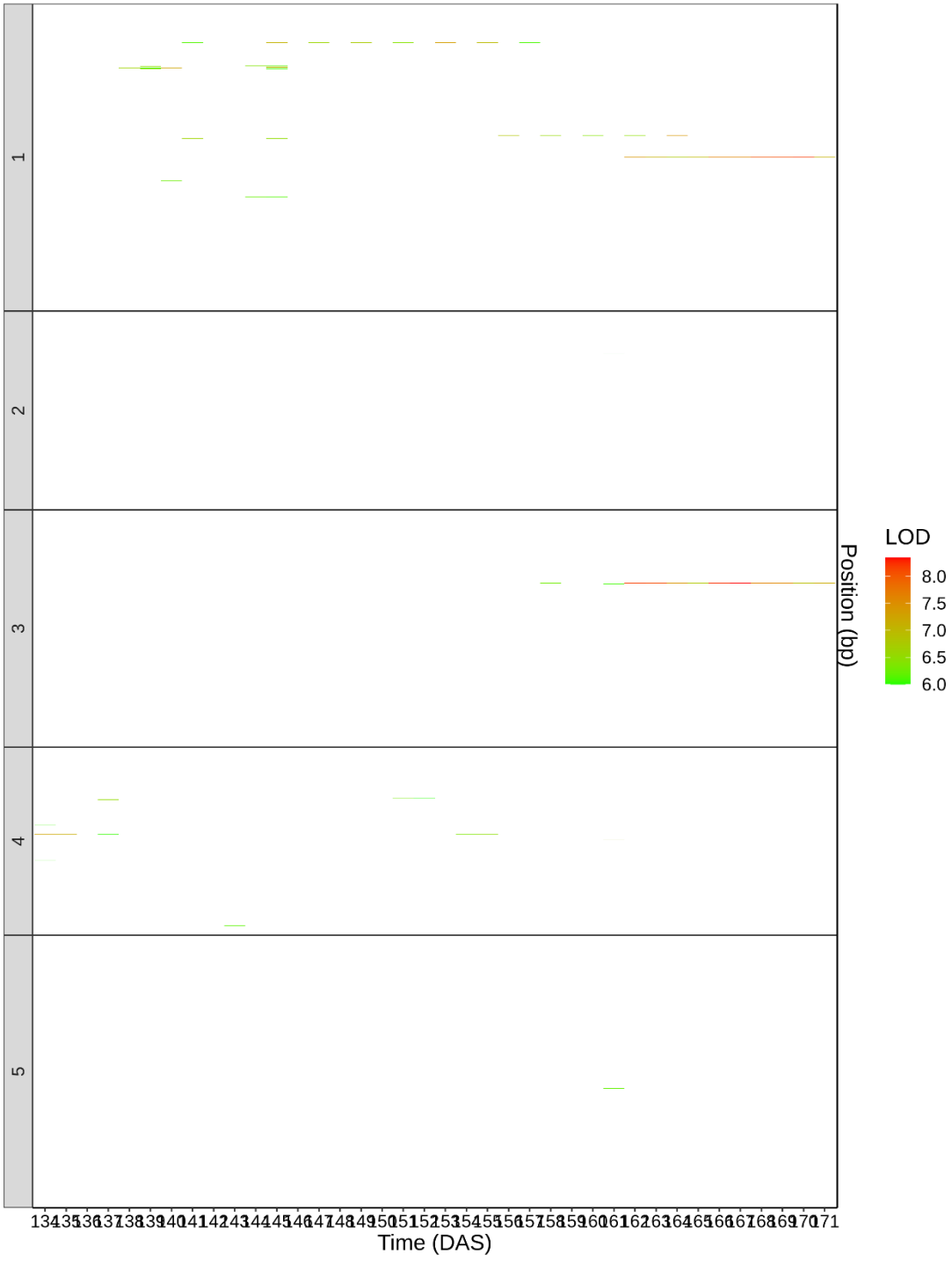


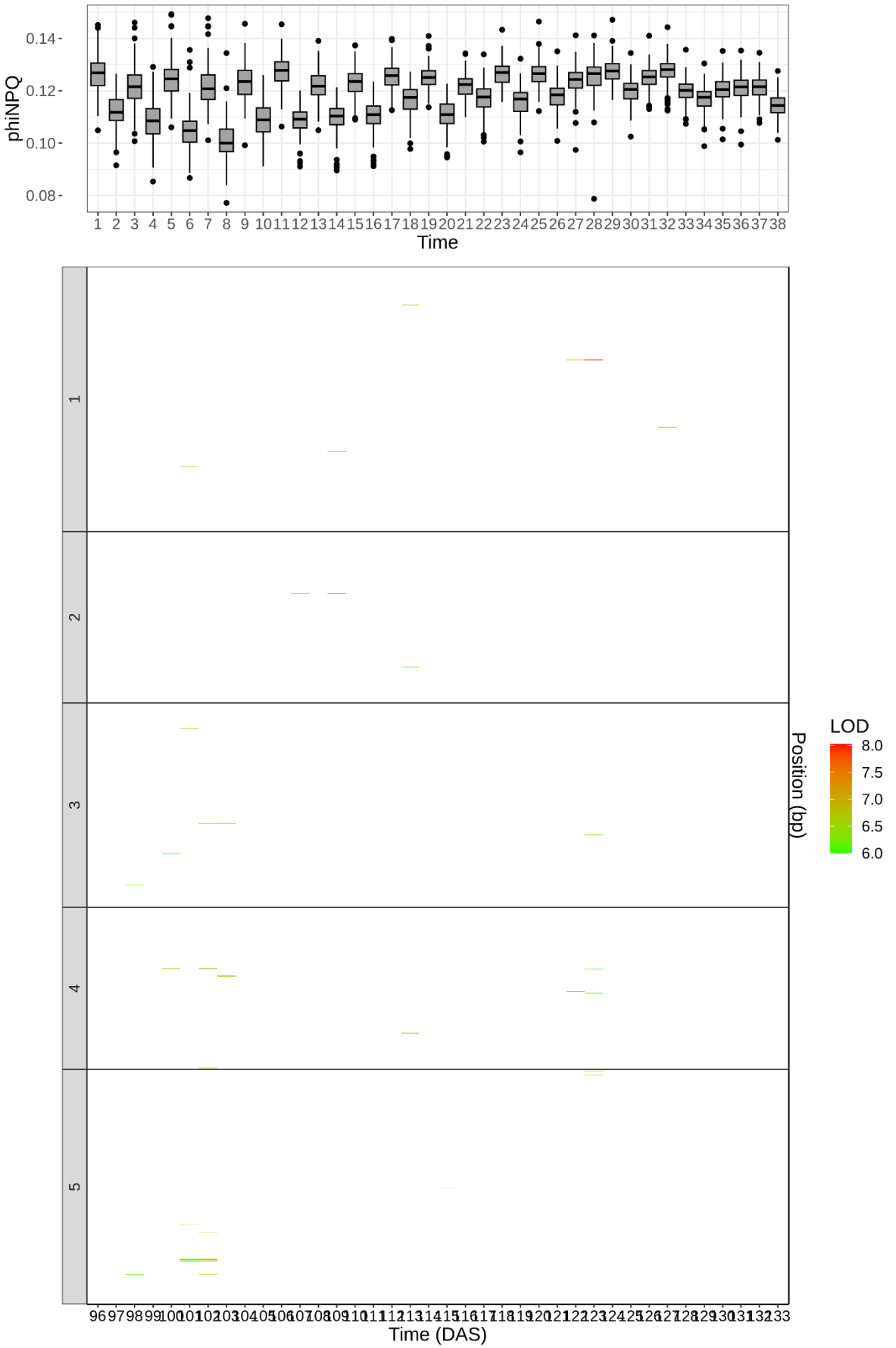


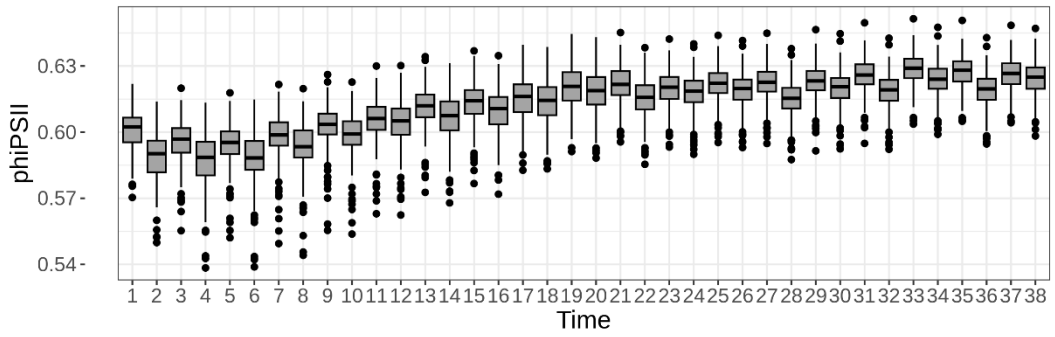

NPQ
0.5
0.4
0.3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Time


1
2
3
4
5
Position (bp)
LOD
7.6
7.2
6.8
6.4
6.0
58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
Time (DAS)


phiNO
0.350
0.325
0.300
0.275
0.250
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Time

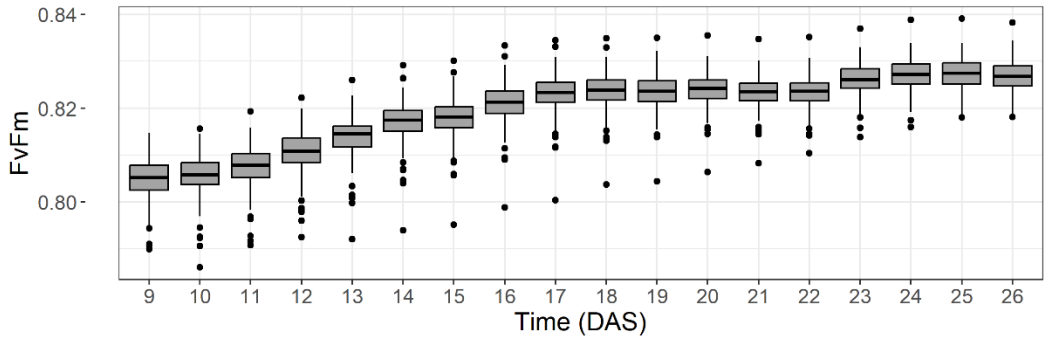

1
2
3
4
5
Position (bp)
LOD
8.0
7.5
7.0
6.5
6.0
Time (DAS)

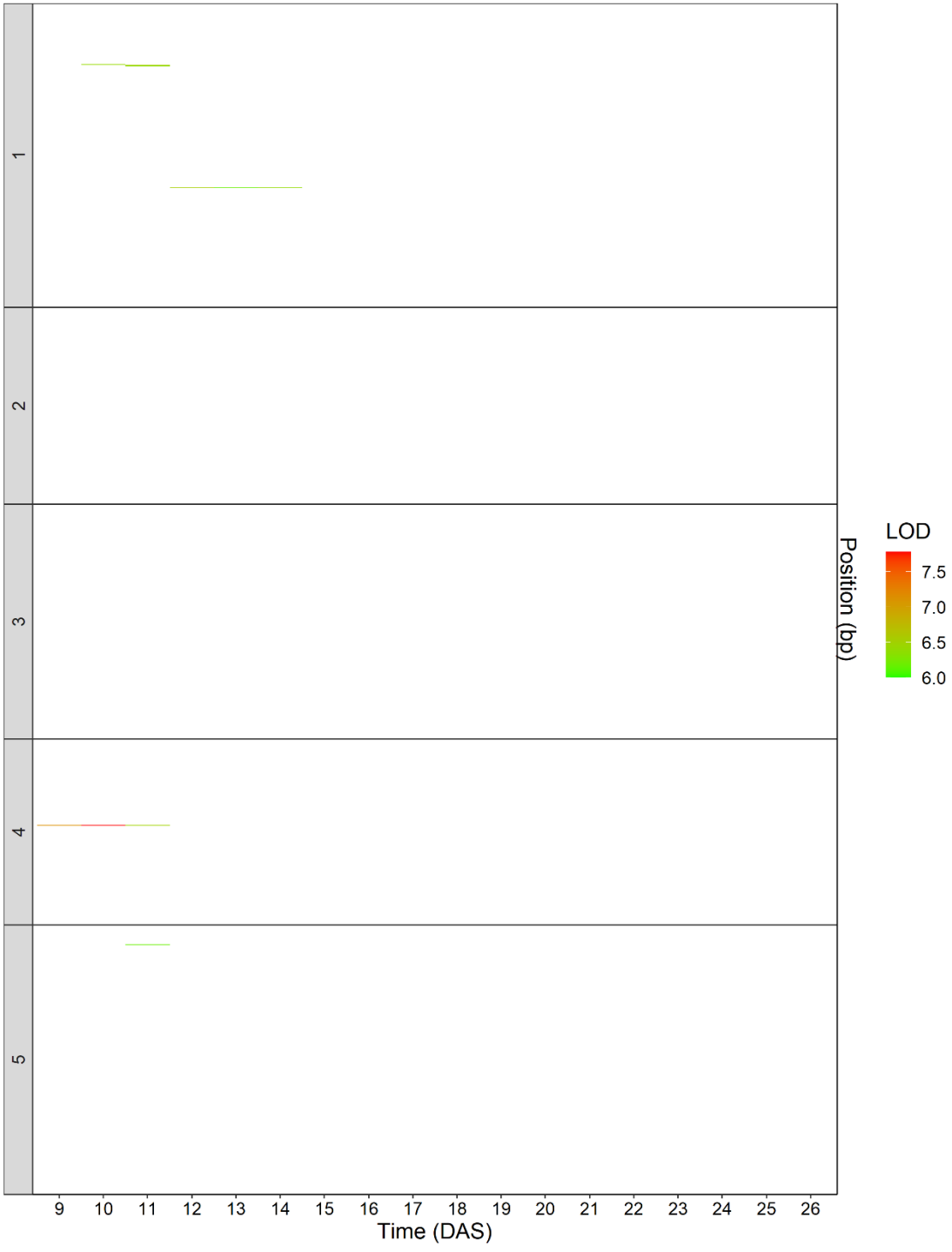

phiNPQ
0.14
0.12
0.10
0.08
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Time
1
2
3
4
5
Position (bp)
LOD
8.0
7.5
7.0
6.5
6.0
96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133
Time (DAS)

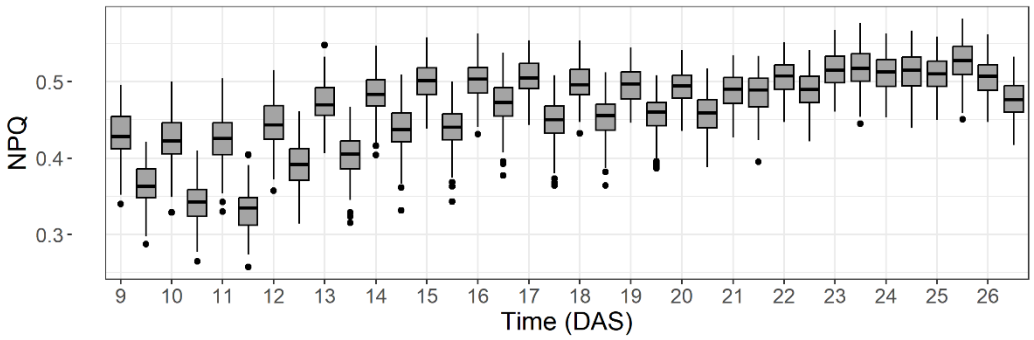

phiPSII
0.63
0.60
0.57
0.54
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Time

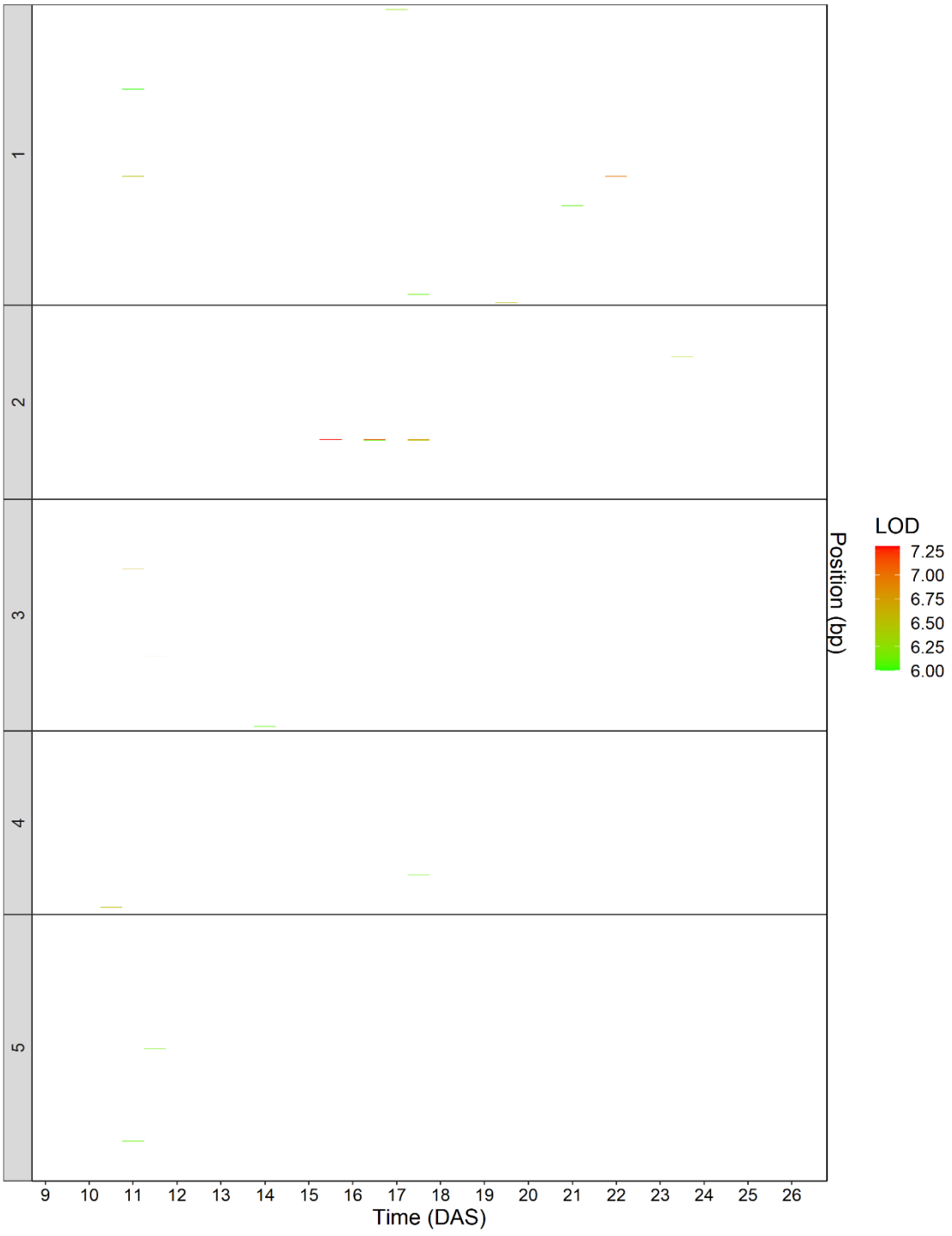

1
2
3
4
5
Position (bp)
LOD
10
9
8
7
6
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
Time (DAS)

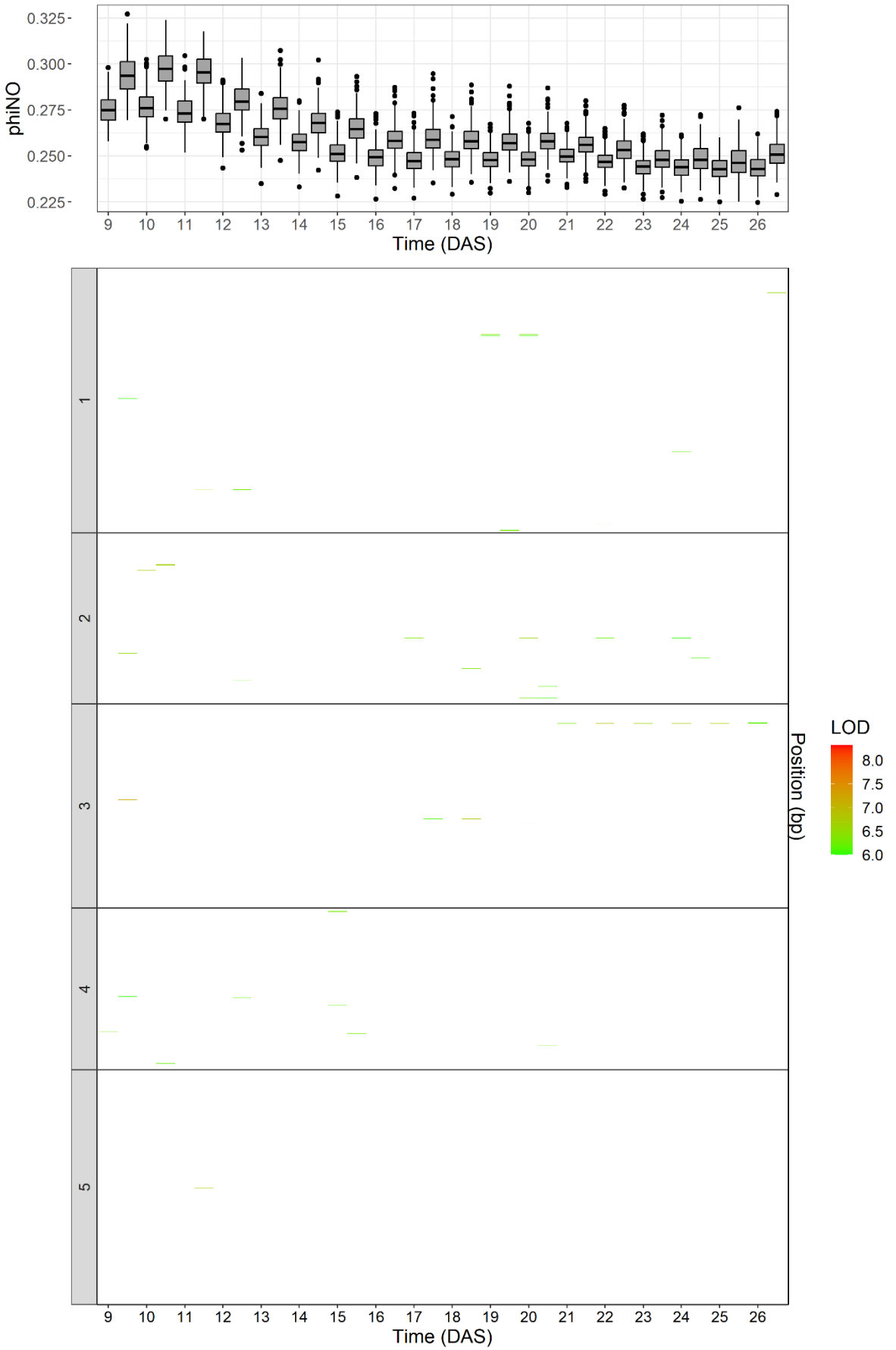
NWO22_03

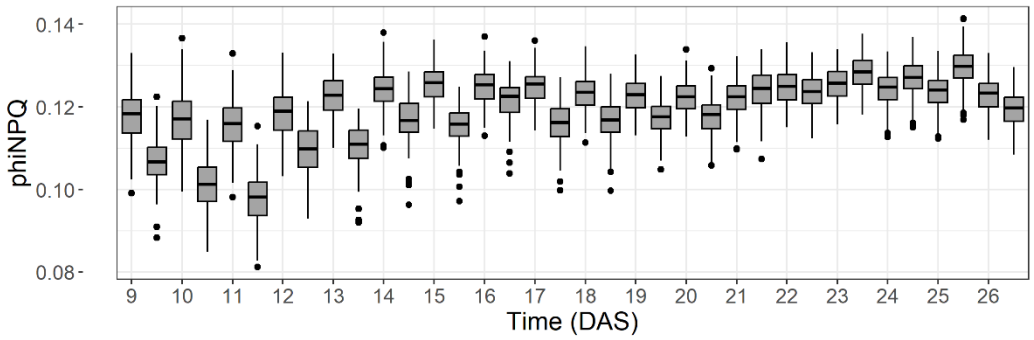

FvFm
0.84
0.82
0.80
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)

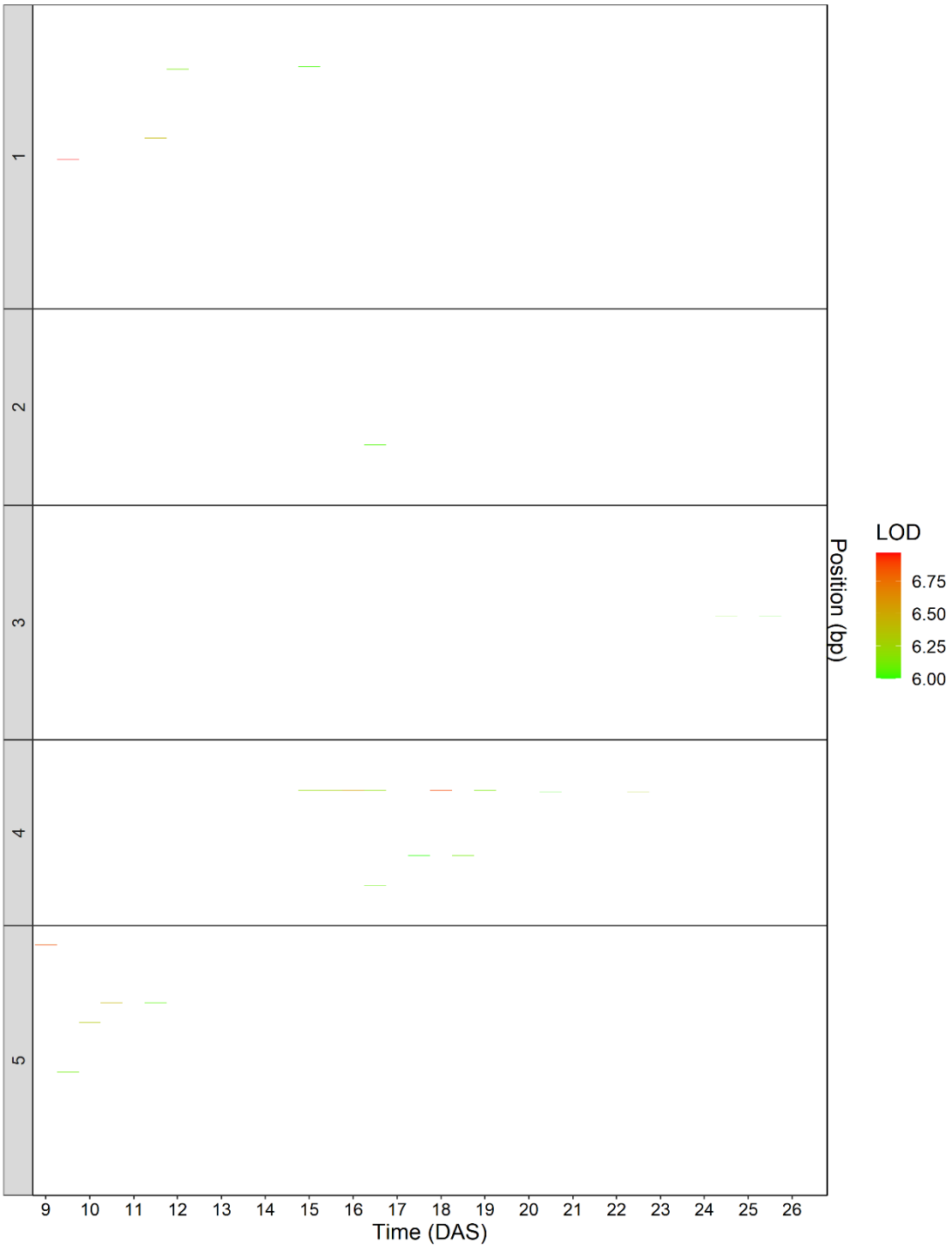

1
2
3
4
5
Position (bp)
LOD
7.5
7.0
6.5
6.0
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)

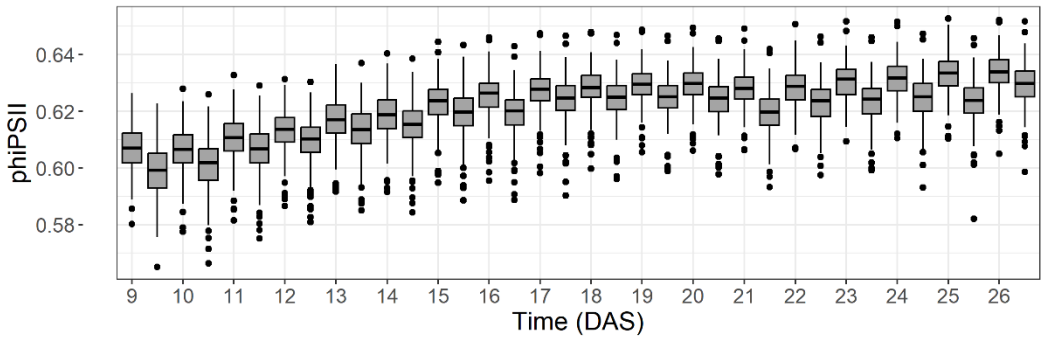

NPQ
0.5
0.4
0.3
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)

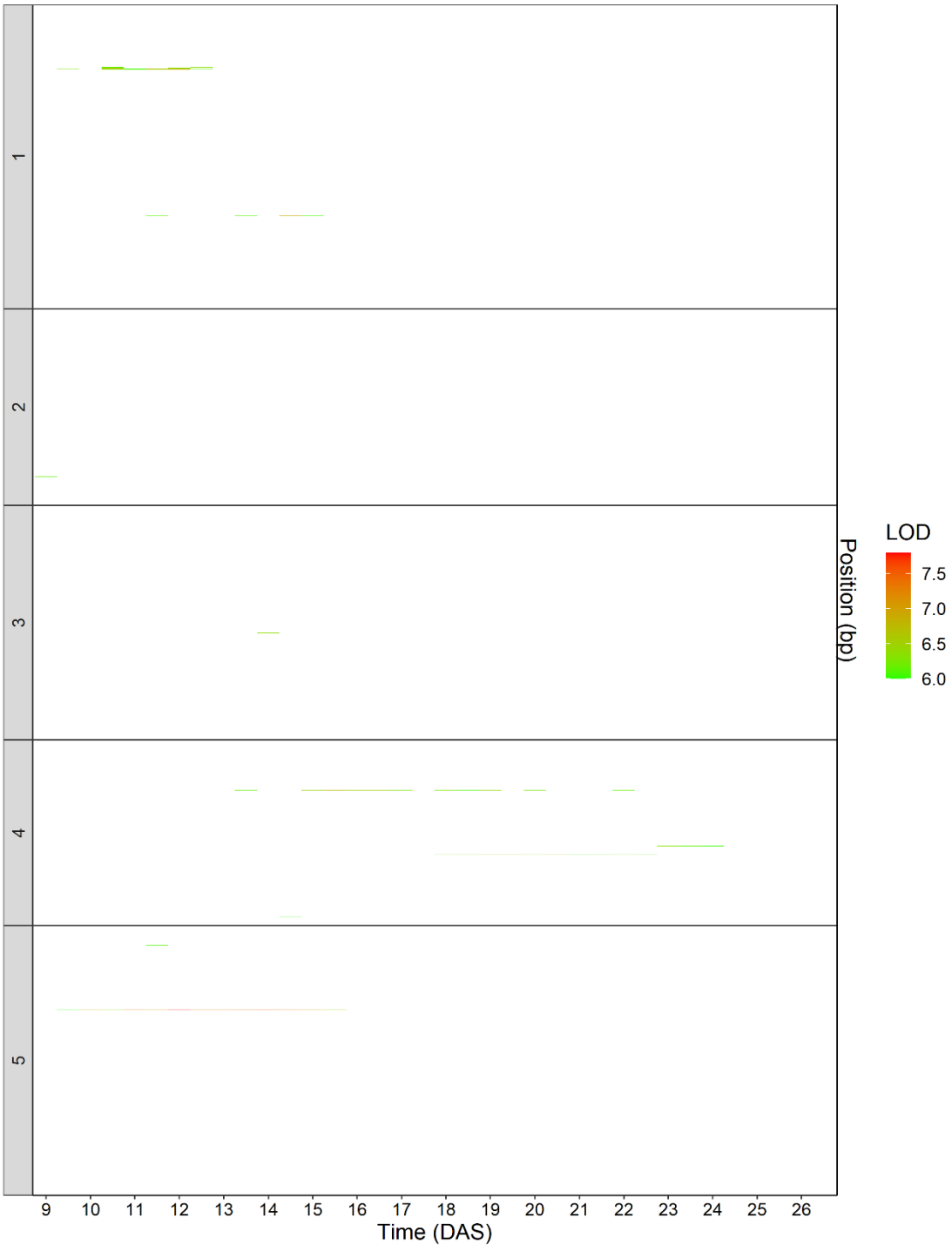

1
2
3
4
5
Position (bp)
LOD
7.25
7.00
6.75
6.50
6.25
6.00
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)


phiNO
0.325
0.300
0.275
0.250
0.225
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)
1
2
3
4
5
Position (bp)
LOD
8.0
7.5
7.0
6.5
6.0
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)


phiNPQ
0.14
0.12
0.10
0.08
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)


1
2
3
4
5
Position (bp)
LOD
6.75
6.50
6.25
6.00
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)


phiPSII
0.64
0.62
0.60
0.58
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)


1
2
3
4
5
Position (bp)
LOD
7.5
7.0
6.5
6.0
9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Time (DAS)