Forschung hautnah:

Wissenschaftliches Schülerpraktikum vergeben durch den Förderverein der BiologieOlympiade e.V.

Am Max-Planck-Institut
für Molekulare Pflanzenphysiologie

in Potsdam-Golm

Arbeitsgruppe um Prof. Dr. Alisdair Fernie

Betreuer: Dr. Saleh Alseekh

Thema: Genome-wide association studies (GWAS) revealed the genetic architecture of lipid metabolism in common bean seeds.

Konrad Frahnert
Klassenstufe 11/12
Schule: Weinberg-Gymnasium Kleinmachnow

Zeitraum: 13.07.2020 – 07.09.2020

Index

1. Introduction

1.1 Personal Introduction

Nature and living forms always fascinated and interested me. In school, I was curious in any nature science and my interests then focused on biology since class 7. Being zoologists themselves, my parents supported my interests and showed me their work. At first, I mostly was fascinated by animals and classic zoology, not much in botanic or other fields. I still like observing insects or any other animal and am excited about the variety of nature. But I then also wanted to understand how the organisms themselves and biological processes work, leading me to physiology and molecular biology. By now, my interests are widespread within biology, but my passion for zoology and physiology remained. I plan to study biology and become a scientist, I just do not know yet which sub-discipline interests me most.

I just finished class 11 with a mathematics and biology as my advanced courses. My school has a nature-scientific profile and tries to put the focus on STEM-subjects. That's why my teachers also motivated me to participate in several student-Olympiads like the *LandesBiologieolympiade Brandenburg* or the *Internationale Biologieolympiade.* I reached the third round of the *Internationale Biologieolympiade* 2019/20. The *Förderverein der BiologieOlympiade e.V* appreciated my efforts by giving me the opportunity to do an internship in the Max-Plank-Institute of molecular plant physiology in Potsdam Golm. This internship provides the great opportunity to me to discover modern research fields of botany, molecular biology and genetics as well as plant physiology, allowing me to experience the work of a scientist and do lab-work.

1.2 Institutional Introduction

The Max-Plank-Institute (MPI) of molecular plant physiology is located in the Science Park Golm, together with the MPI of Colloids and Interfaces and the MPI for Gravitational Physics. The research of the institute is focused on analysis of central metabolic pathways combined with the analysis of gene function and the development and implementation of phenotyping technologies and system approaches. The overall goal is to understand how growth and metabolism are regulated, to learn how they respond to environmental factors, and to unravel genetic factors that underlie these processes and responses.

I worked in the department Molecular Physiology (Prof. Dr. Dr. h.c. Lothar Willmitzer), in the group "Central Metabolism" of Prof. Dr. Alisdair Fernie. This group focuses on identifying factors involved in metabolic regulation of primary metabolism within both photosynthetic and heterotrophic tissues. Thale cress (*Arabidopsis thaliana)*, tomato (*Solanum lycopersicum*), tobacco (*Nicotiana tabacum*) or beans (*Phaseolus vulgaris L.*) and related species like lupine (*Lupinus albus L.*) are often used for the experiments.

Figure 1.1: The Max-Planck-Institute for molecular plant physiology

1.3 Abstract

The aim of the project I worked on during my internship was to identify the genetic architecture of lipid profiling in common beans (*Phaseolus vulgaris*) and the genes involved in lipid biosynthesis in the seeds. In 202 different breeding lines, 161 different lipid-features were found and used for the further analyzation. I extracted the lipids from the seeds and analyzed the content and composition of the extracts by UHPLC-MS. The lipid profiles were used in a genome-wide association studies (GWAS) approach to identify candidate genes affecting the accumulation of lipid specific features. For several features candidate genes were identified. In this report, I will only show a few examples. In addition I created a heatmap and a PCA to explore the lipid diversity in the seeds of common bean. The result of this project provide a general overview about genetic control of lipid profiling in common bean and a framework for understanding the metabolic processes to improve quality of common bean seeds.

2. Project referring information

2.1 Background information

Common beans are one of most important grain legumes for direct human consumption worldwide. They are a major source of protein and micronutrients and provide health benefits that are related to their consumption.

Humans domesticated crops since their transition from hunting-gathering to agriculture (Rendón-Anaya et al., 2017). Being a nutritional very important legume, especially beans were domesticated and selected on special traits (Quiroz-Sodi et al., 2018). There were firstly two species of beans independently domesticated at least twice - in Mesoamerica and in the Andes (Rendón-Anaya et al., 2017). Later cultivated lines were also brought to Europe, which has a very different climate and biotic and abiotic factors, leading to a further domestication. Beans are important as food, forage and nitrogen fixers. They were selected based on apparent qualities like seed size, growth and gain (Singh et al., 2018). But the domestication-focus did not include other helpful traits for example resistance against diseases and pests or tolerance for changing environment (Quiroz-Sodi et al., 2018). Traits like nutrient, vitamin or mineral content where also overlooked. Because of its importance, beans are still part of breeding-efforts. These projects focus on development of improved varieties with tolerance to stresses, resistance against diseases, nutritional value and good growth (Lobaton et al., 2018). The general aim is to breed an optimized or improved bean. Using genetic tools for this breeding holds the potential to be faster and cheaper than the common breeding. But therefore it is needed to know which genes encode for which traits.

In my project I examined the lipid profiling in the seeds of, 202 different lines (varieties) of common beans which were collected from South America and Europe (Bellucci et al. unpublished). The kinships between the lines are not completely known. The project was part of the INCREASE study, an EU-wide project to implement a new approach to conserve, manage and characterize genetic resources leading to benefits on different levels, focusing on chickpea, common bean, lentil and lupin (The INCREASE consortium, 2020). Lipids are mostly non-polar, organic molecules. They are very important for the bean itself as well as a nutrient for humans. Lipids are essential for all living cells, because they are building blocks of the membranes, which enclose the cell and the internal organelles and also function as energy storage or signaling molecules (Hummel et al., 2011). Lipids can also be essential nutrients or work as antioxidants. Because of this importance of lipids for living forms, a complete new branch in the metabolomics, namely the field of lipidomics, emerged (Hummel et al., 2011).

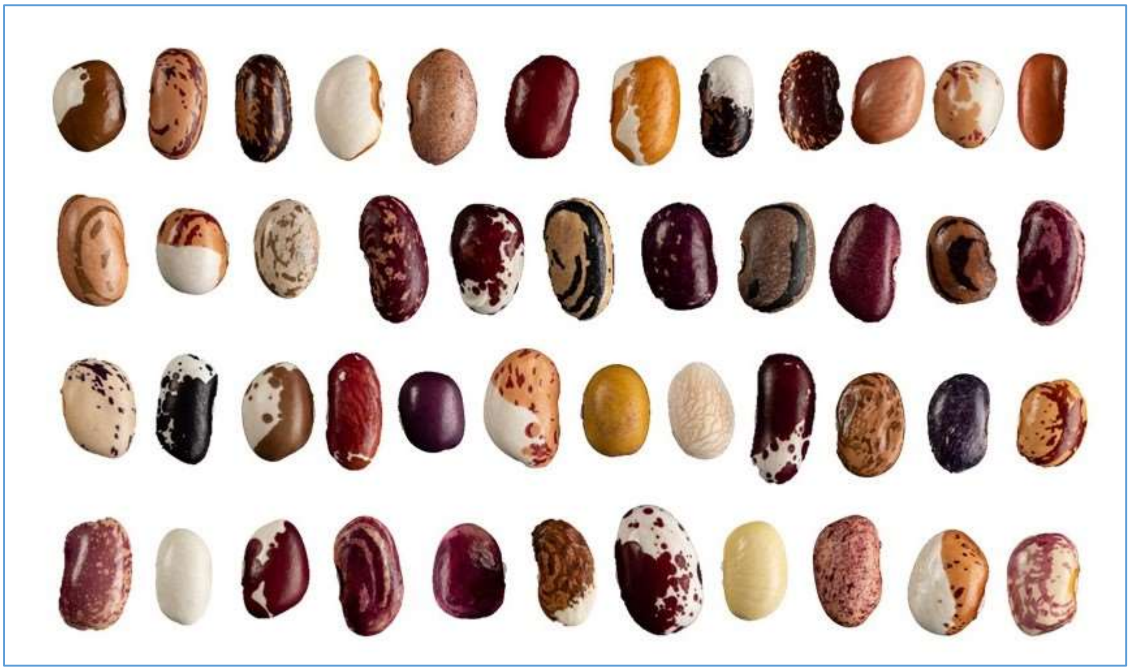

Figure 2.1: phenotype of the seeds of some of the studied beans illustrating the high variability within P. vulgaris

2.2. Methods to identify genes associated with the lipom

The lipid content and composition of 202 bean lines serves as phenotype for the following genome-wide association studies (GWAS)-analysis. The initial steps were to extract, measure and quantified the lipids in bean seeds. To analyze which lipids were contained and in which quantities, I used the ultra-high performance liquid chromatography-mass spectrometry (UHPLC-MS)-method. The ultra-high performance liquid chromatography (UHPLC) is a physiox separation-method, in which the analytes distribute between two non-mixable phases. The stationary phase is the solid material inside the packed column. The mobile phase is the liquid-mixture running through the column. Based on their chemical characteristics, the analytes bind differently to the two phases. Analytes binding better to the stationary phase move slower than analytes binding better to the mobile phase, leading to a separation of the analytes. The time it takes the molecules to pass through the column from injection till detection is the retention-time (RT), which is characteristic for the molecule. After being separated, the analytes are characterized with the mass spectrometer (MS). The molecules are ionized and fragmented in an ion source and then speed up in an electrical field. The speed they reach depends on their charge and mass, so the molecules and fragments separate by mass, creating a characteristic mass-spectrum. The data-output of the UHPLC-MS-method are UHPLC-chromatograms with mass-spectra associated with every peak.

2.2. Genome-wide association studies (GWAS) approach

To identify the region in the genome associated with the accumulation of to certain lipid class, I mapped the genes by GWAS. This approach uses single nucleotide polymorphisms (SNPs) to map genome regions associated with a specific trait expression. In this case, the identification was reached by mathematical correlation of single nucleotide polymorphisms (SNPs) across 200 accessions and the changes in the lipid composition of these accessions.

Single nucleotide polymorphisms SNPs are a genetic variation in form of single base-pair substitutions. SNPs are the most common type of genetic variation among individuals. These variations can be harmless, harmful or latent.

GWAS is a method that involves scanning across the complete set of DNA of many individuals to find genetic variations associated with a particular phenotype. For a GWAS hundreds to thousands of individuals are phenotyped for one or several traits. These individuals need to have phenotypic differences in the trait of interest, so a population or several groups with different phenotypes can be formed and compared. The individuals then get genotyped by extracting and sequencing their DNA. Especially the varieties in the genotypes (different allele frequencies) are important (mostly SNPs with different allele frequencies). Then the varieties in the genome sequence are correlated with the phenotypic traits to find out which SNP is associated with the trait of interest.

The GWAS-results were displayed by a Manhattan-plot with -log10(p-value) plotted against the position in the genome. The p-Value is an indicator for the significance of the difference in frequency of the allele tested between the groups. In a Manhattan-plot each dot represents a SNP and the height of the dots show the strength of difference of the allele frequencies between the different groups is. A significant difference in the frequency distribution of the nucleotides indicates that a SNP could be actual responsible for the phenotype trait, but more often, they are just near the casual differences. So mostly the correlated SNPs are more or less just a hint on the genome region, where the gene encoding for the trait of interest is located (often the SNP is just near the associated gene and does not affect the trait itself). Furthermore the gene found with GWAS could be a regulatory gene, which just regulates the expression of the gene encoding for the trait of interest.

Correlations between the individuals and the analytes are often visualized by a heatmap. A heatmap displays the correlation between two groups with a color-code (for example high-correlated traits red, low-correlated traits blue), giving a good overview over the correlations and enabling to create clusters or relations.

To find the gene, it is needed to go through all genes encoded in an area around the SNPs correlated with the trait of interest. Therefore, the single genes of the genome must be known. For model organisms there is often a databank with the single genes, their position in the genome and often also the function of the genes. In this case it is possible to guess a gene candidate for the trait of interest by searching in the surrounding of the correlated SNP for a gene with a function related to the trait. To prove, that this gene itself is correlated to the trait of interest, it could be knocked out, silenced or overexpressed in another experiment, to see, if this affects the trait significantly.

It is possible, that several genes, affecting the trait of interest are near to each other. The significant SNPs of these genes would also be near each other (but not linked) and could be displayed as one significant gene region. To discuss the results of the GWAS, it is therefore useful to calculate the Linkage Disequilibrium (LD) of candidate genes and to display the correlation between SNP-haplotype and expression of the trait in a boxplot. A LD correlates the single significant SNPs for a single trait, making it possible to check whether these SNPs are linked. If they are not linked, the nearby SNPs could represent different loci, meaning that several genes could be involved.

3. Material and methods

3.1 Extraction

Lipids were extracted from seeds of 202 common bean accessions. For each line, three independent replicates were used. For each replicate 5 – 6 dry seeds were grinded to fine powder using a Retsch mill (MM301, 30 seconds). 50 mg of this powder was weighed into a 2 ml Eppendorf tube. 10 ml of a homogenous, precooled (-20°C) methyl-tert-butyl-ether:methanol (3:1) mixture was added into the tubes. The mixture was homogenized using a Vortex-mixer. The samples were kept on ice to prevent chemical reactions. The samples then were shaked in a shaker (Eppendorf Mixer 5432) for 20 minutes at 4°C, followed by 15 minutes sonication in an ice cooled ultra-sonication bath. After adding 500 µl of ice cooled water:methanol (3:1) and mixing by a Vortex-shaker for 15 seconds, the mixture was centrifuged for 5 minutes with 11500 rpm in a table top centrifuge (Eppendorf Centrifuge 5417C, 15 cm diameter). The addition of the water:methanol leads to a phase separation producing an upper organic phase, containing the lipids and a lower phase containing the polar and semi-polar metabolites (Hummel et al., 2011). 300 µl of the upper phase were transferred into fresh 1.5 ml Eppendorf tubes under the fumehood. In the case of one of the three replicates missing in a line, it was replaced by taking once more 300 µl from another replicate of this line. The tubes with the extracted lipids were dried in a speedvac, closed and stored at -20°C till the resuspension. In total about 650 samples were extracted for this experiment.

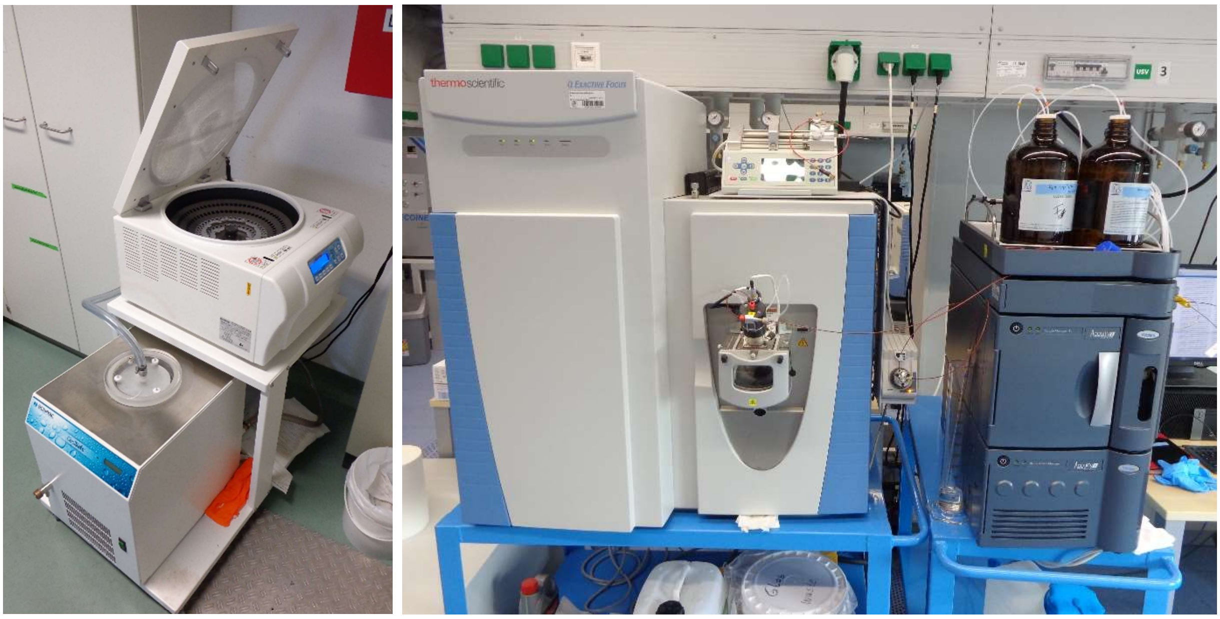

Figure 3.1: The speedvac and the UHPLC-MS

3.2 Quality Control

Due to technical varieties of the HPLC-MS during and between runs, shifts in the retention-time or intensity of analytes may appear. In order to minimize these variations across several days of analysis, we prepared a standard sample. Such a quality control (QC) allows to detect the HPLC-MS-shifts and normalize affected samples if needed. To create this pooled quality sample (QC), 300 µl of the leftover of the first set of extracted samples was mixed in a 50 ml Sarstedt tube and an exact volume of 300 µl was distributed into 1.5 ml Eppendorf 40 tubes labeled as quality control. These tubes were dried in the speedvac and stored at -20°C till resuspension. Per analysis a set of 60-90 experimental samples, and 3 QC-samples were prepared (see section 3.2 resuspension) and analyzed as one batch (one set). In total 10 sets of samples with about 650 samples in total were analyzed in 2 weeks.

3.3 Resuspension

The dried lipid extracts were re-suspended in the HPLC-medium for the analyzation. Therefore, 200 µl of a acetonitrile:isopropanole (7:3) mixture was added to each sample (tube). The tubes then were closed and kept for for 3-5 minutes at room-temperature and were then vortexed 5-10 seconds. After sonicating the tubes in a sonication bath and mixing by a Vortex-shaker again, the tubes were centrifuged for 3 minutes at 11500 rpm in a table top centrifuge (Eppendorf Centrifuge 5417C, 15 cm diameter). A small pellet precipitated on the bottom of the tube. 100 µl of the sample without pellet were transferred into labeled HPLC-MS glass-vials (1.5 ml vials with 300 µl inlet).

3.4 Analysis of lipid profile using HPLC-MS

The samples were analyzed with a UHPLC-MS (UHPLC: Waters, ACQUITY UPLC System; MS: Thermo Scientific, Q Exactive Plus) on a C_8 reverse-phase column (100 mm x 2.1 mm x 1.7 µm particle size, Waters) at 60°C. The mobile phases consisted of 1% 1 M NH4CH3COO and 0.1% acetic acid in water (buffer A) and acetonitrile/isopropanol (7:3, UPLC grade BioSolve) supplemented with 1 M NH_4Ac and 0.1% acetic acid (buffer B). The following gradient profile was applied: 1 min 45% A, 3 min linear gradient from 45% A to 35% A, 8 min linear gradient from 25% to 11% A, 3 min linear gradient from 11% to 1% A. Finally, after washing the column for 3 min with 1% A the buffer was set back to 45% A and the column was re-equilibrated for 4 min, leading to a total run time of 22 min. The flow rate of the mobile phase was 400 µl/min. With the MS, all the spectra were recorded using altering full-scan and all-ion fragmentation scan mode, covering a mass range from 100–1500 m/z at a capillary voltage of 3.0 kV, with a sheath gas flow value of 60 mL/minute and an auxiliary gas flow of 35 mL/minute. The resolution was set to 10 000 with 10 scans per second, restricting the Orbitrap loading time to a maximum of 100 ms with a target value of 1E6 ions. The capillary temperature was set to 150°C, while the drying gas in the heated electrospray source was set to 350°C. The skimmer voltage was held at 25 V while the tube lens was set to a value of 130 V.

3.5 Data-processing

At first, the relevant data must be extracted from the chromatograms with MS-spectra. Therefore the data were processed with the program *GeneData*. The processing-pipeline included a noise-filter, a RT-alignment, peak detection and an isotope cluster. The noise-filter filtered out all background-signals, which were not from the analytes of interest. The RT-alignment removed RT-shifts and aligned the RTs of the samples. The peak-detection-module determined the process of finding and detecting peaks to quantify their intensities and to validate the identity of the analytes by the spectra. The isotope cluster module was needed because the MS is so exact, that it detects molecules with isotopes in their structure as other molecules. The isotope cluster assembled these molecules, because for my analyzation differences in the isotope composition of the lipid species are not important. After this processing, the data was exported as a table or matrix with the sample number and its quantities for all target peaks (see appendix). The program does not identify the single analytes, often a mass-feature could just be identified as a group of derivates of a basic lipid.

The aim of the project was the identification of genomic regions associated with the variability in lipid accumulation and ultimately to map genes involved in lipid biosynthesis. Large amount of data was generated by the analysis and data-processing, given that the limitation for my report, I only will show some examples from this analysis. I used GWAS to do this (see method information). The data was uploaded into an R-program that matched a p-Value to each SNP, in order of the chromosomal position and created Manhattan-plots out of them. For each detected lipid (used as measurable trait), an own plot was created. The program also identified the genome locations of the significant correlated SNPs. Then I searched for possible candidate genes associated with the expression of the lipid species, looking around the position of significant SNPs in *Phaseolus vulgaris* genome.

The online-tool *ClustVis* was used to generate the heatmap and a cluster-tree for finding relations between the different breeding lines based on their lipid profile.

To extract the haplotypes from significant SNPs and create a Linkage Disequilibrium (LD), the program *Tassel 5* was used. Extracting the haplotypes is important to correlate the lipid-expression with the special SNP-haplotype. With the haplotypes a Boxplot matching the phenotype-feature expression and the SNP-haplotypes was made with *Excel*.

4. Results

In our experiment, we were able to measure 161 lipid derivatives across all the lines. The lipid profiling showed large differences between the accessions. Particularly, large changes were found between the genotypes which were collected from Europe and America (Figure 4.1 cluster analysis of a PCA). The left and the right part of the PCA (Andean origin beans and Mesoamerican origin beans) are clearly separated (Figure 4.1). There is also a clear but weaker separation between the European and American beans, with the European beans in the upper part of the PCA and the American beans in the lower part of the PCA. Differences also appear by clustering the data in a cluster-tree (Figure 4.2) and a heatmap (Figure 4.3).

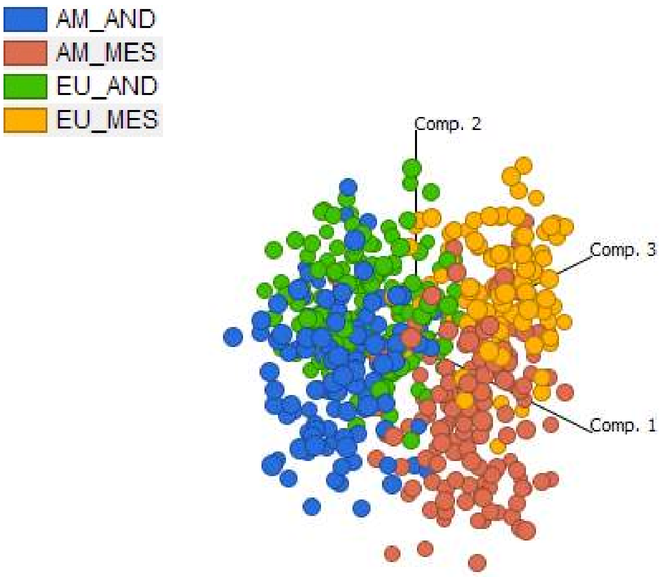

Figure 4.1: Cluster Analysis of a principle component analysis (PCA) reduced data set across 202 accessions. EU_AND (Green dots, Europe Andean origin), AM_AND (blue dots, American Andean origin), EU_MES (yellow dots; Europe Mesoamerican origin), EU_MES (orange dots, American Mesoamerican origin).

The heatmap (Figure 4.3) displays 4 clear clusters. An additional narrow stripe at the top of the heatmap contains lipid species that are more or less homogenous present throughout all bean species. The correlation-cluster in the top left corner shows a correlation between the first lipid-cluster and the left big cluster in the heatmap-cluster-tree, including both, American and European beans. A second correlation-cluster in the middle of the heatmap displays the correlation between the American cluster of the right part of the heatmap-cluster-tree and several lipid-clusters. A third correlation-cluster is slightly visible in the bottom-right corner of the heatmap. It shows the correlation of the European beans of the right part of the heatmap-cluster-tree and a strongly delimited lipid-cluster.

In the heatmap the bean lines cluster differently than in the cluster-tree.

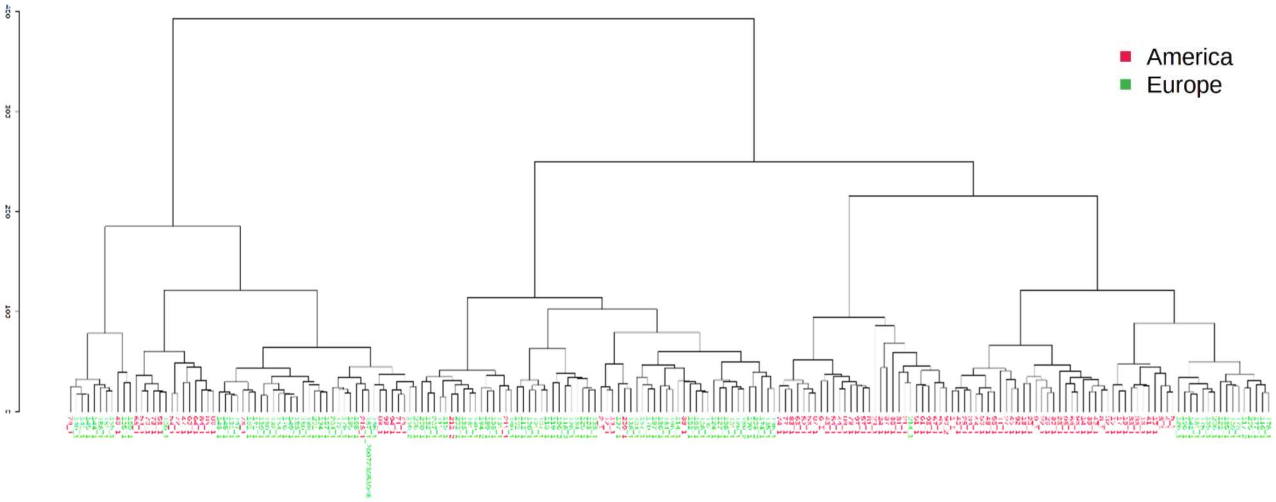

Figure 4.2: Cluster-tree of the lines based on the differences in the content of different lipids (only for first replicate per line)

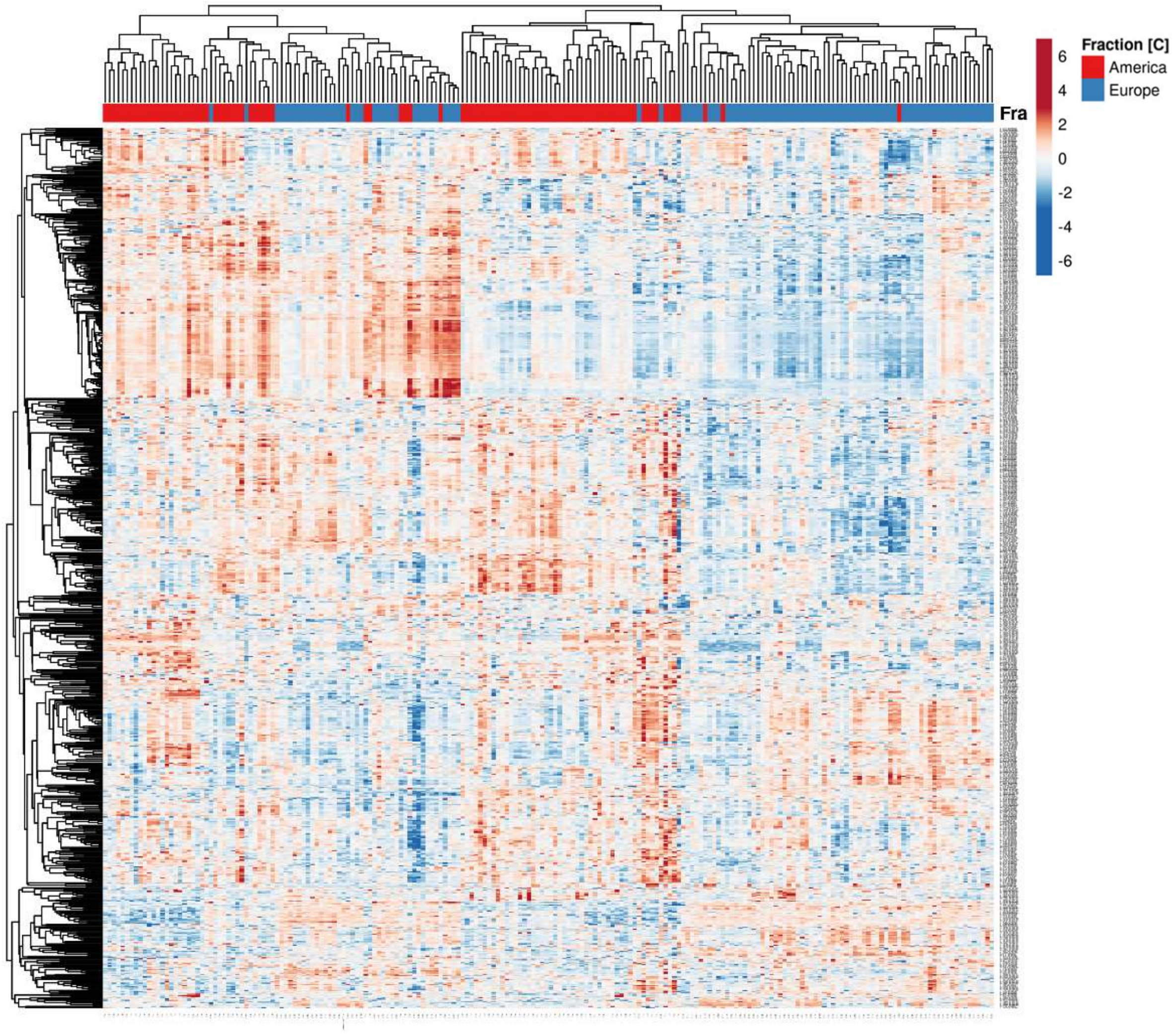

Figure 4.3: Heatmap of the lipid-features and the replicates (only one replicate per line)

In order to explore the genetic architecture of lipid metabolism in common bean seeds, GWAS analysis was performed on all 161 lipid mass features. The GWAS resulted in 8.500 Manhattan-plots of lipids within 1.200 ones with significant SNP correlations. For this report, I have chosen four plots as an example. In the first case, a significant correlation between the lipid-feature 379.151 (Figure 4.4), with several SNPs at the end-region of chromosome 2 (around 40 megabases (Mb)) was found. In the genomic region of the SNP correlated with the lipid-feature 379.151 (Figure 4.4), a gene which encodes for the lipoxygenase3 is located. This protein is known to involve in lipid metabolism and could be behind the regulation of the expression of this lipid derivates. For the lipid-feature 996.705 (Figure 4.5), which includes Triacylglyceride (TAG) derivatives, a significant correlation between the lipids expression and a SNP-region in the end region of chromosome 2 (around 42 Mb) was found. Within this region, there is a gene, encoding for acyl-CoA oxidase 2, which is involved in long chain fatty acid biosynthesis, making it a potential candidate gene. For the lipid-feature 898.726 (Figure 4.6), which includes Triacylglyceride (TAG) derivatives, a significant correlation between the lipid expression and a SNP-region in the mid-region of chromosome 9 (around 14 Mb), was found. Within this region, there are three genes, all encoding for phospholipase D alpha 2, which functions in the phospholipase D activity and highly appear to be the candidate genes for changes in lipid accumulation across the accessions. A significant correlation was also found between the lipid-feature 698.629 (Figure 4.7), which includes Monogalactosyldiacylglycerol (MGDG) derivatives and a SNP-region in the beginning-region of chromosome 7 (around 2 Mb). Within this region, there is a gene, encoding for lipase class 3 family protein, which is involved in triglyceride lipase activity and lipid metabolic processes, making it a potential candidate gene for this lipid trait.

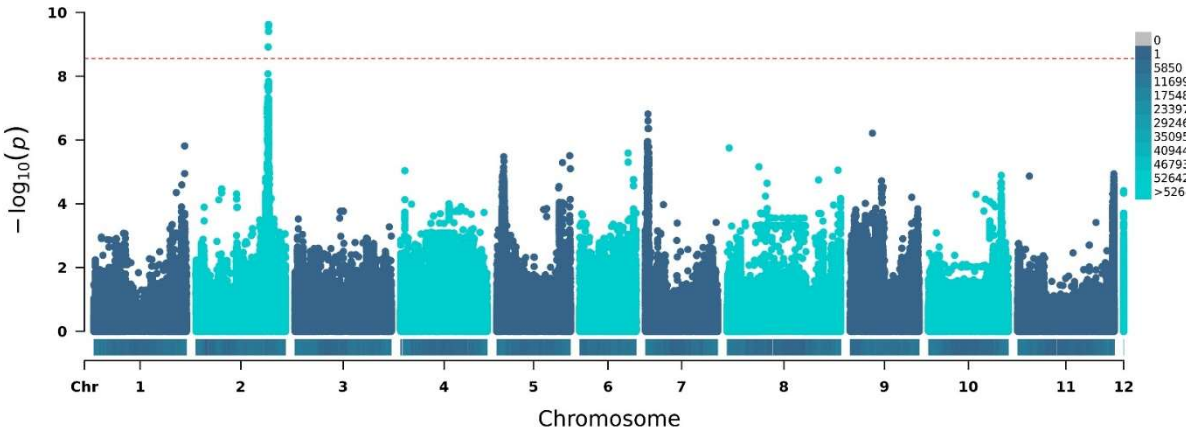

Figure 4.4: Manhattan-plot of the lipid-feature 379.151 (lipid derivatives)

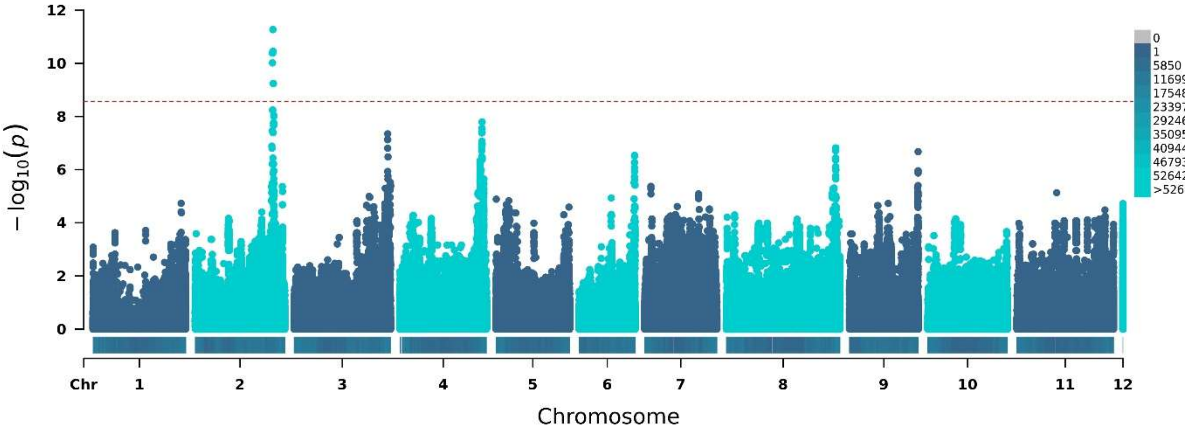

Figure 4.5: Manhattan-plot of the lipid-feature 996.705 (Triacylglyceride (TAG) derivatives)

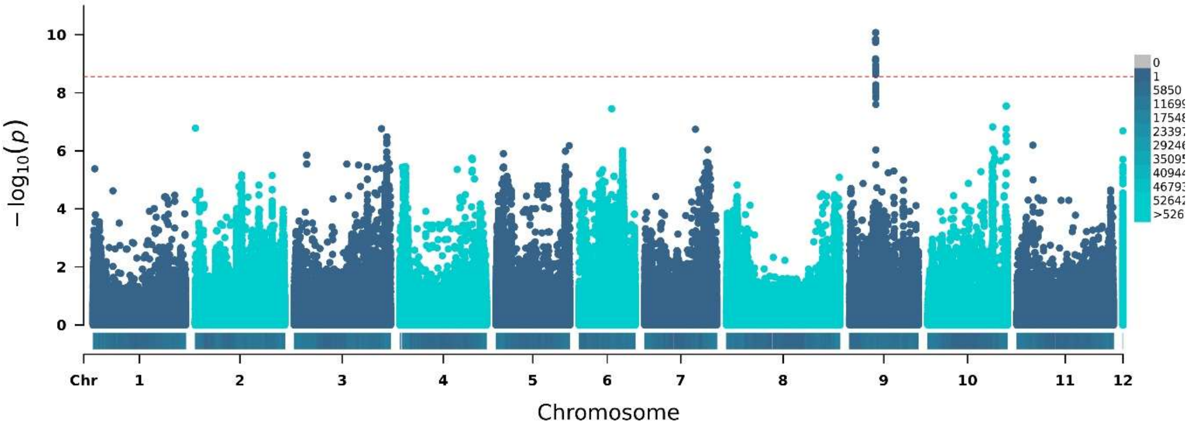

Figure 4.6: Manhattan-plot of the lipid-feature 898.726 (Triacylglyceride (TAG) derivatives)

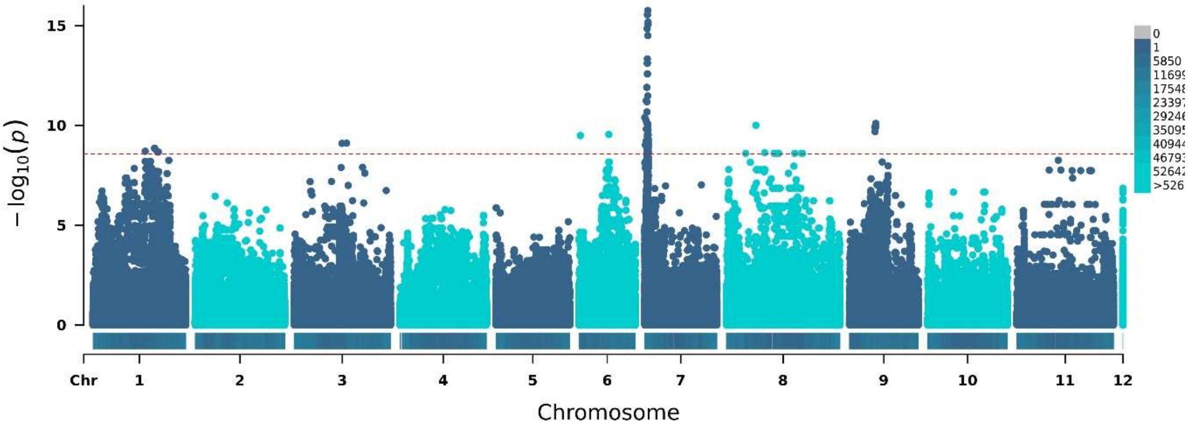

Figure 4.7: Manhattan-plot of the lipid-feature 698.629 (Monogalactosyldiacylglycerol (MGDG) derivatives)

5. Discussion

Seed lipid is a major carbon resource during germination and sapling growth. In this study, we analyzed the lipid composition in 202 common bean accessions. Data showed that the bean accessions represented high phenotypic variability for lipid composition (heatmap (Figure 4.3)). The influence of geographical origin on these lipid contents was clear (PCA figure 4.1). In several plant species it has been shown that the biochemical compositions influence quality traits including lipid composition and containment (Correa et al., 2020). In addition, accessions with different lipid and content serve as a source of energy and influence development of plants. Therefore, the results of the present study can contribute to common bean breeding-programs to deliver high-quality bean varieties according to the consumer market demands.

The accessions differentiate in their lipid profiles. But the lipids do not variate completely free, there are some correlations resulting in clearly visible clusters in the heat map (Figure 4.3).

Beans differentiate also regarding the region of primary cultivation. The beans with Andean origin differentiate clearly from the beans with Mesoamerican origin in their lipid profiles (Figure 4.1). This separation is even stronger than the separation between the European and the American beans. This implies, that the beans with different origins differentiated in their lipid compositions and these differences remained through the process of domestication. Differences in the lipid compositions of American and European beans also appear, meaning that the lipid compositions of the European beans changed in adaption to the new habitat. This adaption was similar within the beans with Andean and Mesoamerican origins, visible in the proximity of the European beans in the cluster analysis.

Also in the clustering tree (Figure 4.2), the groups of the American and European beans are mostly, but not strictly separated, meaning, that there are some differences between them. There are some exceptions and the groups are partially mixed, but the differences are clear. The European beans originate from the American beans. So the differences could be a result of adaption and domestication to the European environment and also show differences in lipid-metabolism. It is unclear, weather the adaption of the lipid-metabolism is about the different climate or another aim of cultivation. The different clustering of the clustering tree and the clustering in the heatmap could be because of turning some clusters by their branch circuit.

The data can be further analyzed, especially to find out whether the trait-expression is associated with a special SNP-allele-frequency. Therefore a Boxplot (Figure 5.2) was created. Before that, a LD (correlation of single significant SNPs for a single trait; Figure 5.1) was created, to check, if the significant correlating SNP-region is really just one region and not two nearby regions with more than one gene affecting the trait.

The LD shows that the significant SNPs of the lipid-feature 898.726 are mostly correlated. The border-regions of SNPs are also correlated, so that the few exceptions (non-correlating stripes in the LD) could be random. In the genomic region of significant SNP for this feature, there are three possible gene-candidates which are all copies of the same gene (see results). It is possible, that the non-correlating SNPs in the LD represent several significant loci because of the different but nearby possible genes. In both cases, the SNPs would represent more or less the same candidate genes so that it is unlikely, that another gene is correlated with this feature here.

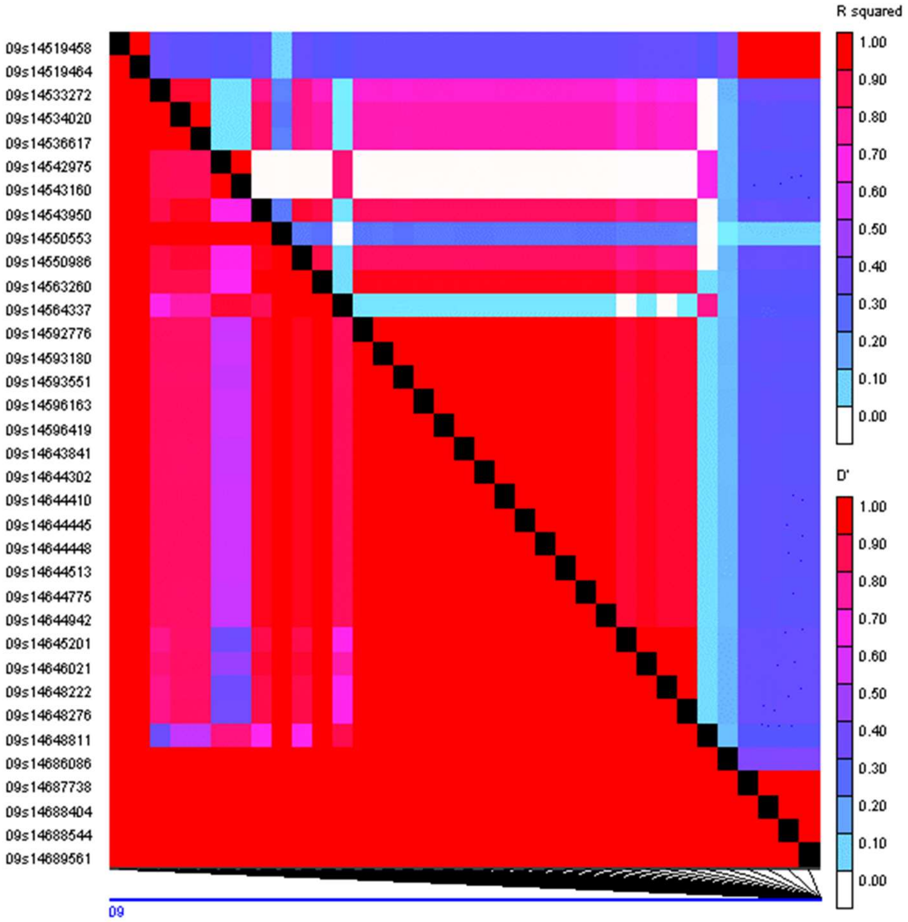

Figure 5.1: LD of the feature 898.726 (Triacylglyceride (TAG) derivatives)

The expression of the traits m/z 898.726 and m/z 698.629 depends on the SNP-haplotype (Figure 5.2). For the lipid-feature 898.726, the lipid expression is likely correlated with the SNP-haplotype C. So for the haplotype C, a higher expression of Triacylglycerides (TAG) is expectable. TAG are neutral lipids, in organisms used for energy and carbon storage and a major component of seed oil (Liu et al.,2015). So the lines with haplotype C will produce more of these TAGs, probably making the seeds containing more of them and so the beans would be more nutritious and a better crop. The associated gene could therefore be important for further breeding programs. The expression of the lipid-feature 698.629 is associated with the SNP-haplotype G, so that for plants with the SNP-haplotype G a higher expression of Monogalactosyldiacylglycerol (MGDG) derivates is expectable. MGDG is a predominant and essential plant lipid required for biogenesis and integrity of plastids and for photosynthetic activity, which require a galactolipid-rich environment (Yamaryo et al., 2006). MGDG is also a precursor of digalactosyldiacylglycerol (DGDG), which is transferred to the extraplastidic membrane under phosphate deprivation (Yamaryo et al., 2006). So all in all it is crucial for plants. The lines with the haplotype G so probably are more resistant as others, making the related gene also interesting for further breeding projects.

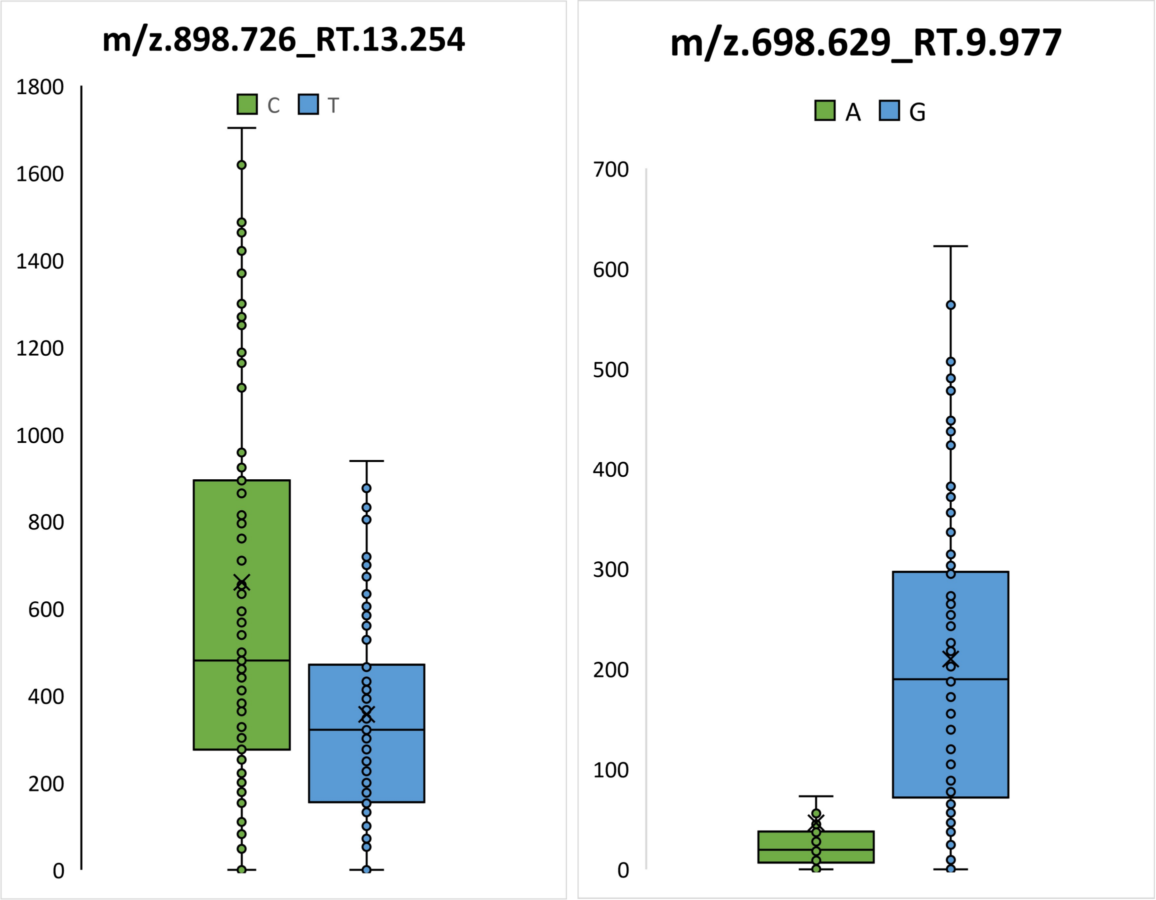

Figure 5.2: Boxplots of the feature-expression for the haplotypes for the two features 898.726 (Triacylglyceride (TAG) derivatives) and 698.629 (Monogalactosyldiacylglycerol (MGDG) derivatives)

The Boxplots show, that in these cases the trait is correlated with a special SNP-haplotype. This does not mean that the SNP itself affects the expression. It is possible that a SNP could affect the expression because it is located in the related gene and so changes the gene-sequence. But it is also possible, that the SNP is just nearby the gene and so correlated with it and the gene expression changed because of a mutation of the gene.

In conclusion, several possible candidate genes for different traits (lipid features) were identified. The next step would be to go on with reversed genetics to experimentally validate these results. This would mean to knock out, silence or overexpress the candidate gene to check, whether this gene affects significantly the trait of interest. Another step would be to completely identify the lipids related features, to helpful understand the functions for the candidate genes. The results finally could be used in breeding projects to create an optimized bean or to adapt beans to different conditions.

6. Personal Conclusion

During my internship, I learned a lot about molecular plant physiology, lab-work and scientific work in general. The lab-work was really interesting and even if at some point it became a bit repetitive, it was still a great experience. The whole process of reaching new knowledge by scientific work fascinated me. It is really exciting to understand how clever modern methods are designed and how these powerful tools can be used to unravel the many secrets of physiological networks and their genetic basis. Processing and analyzing the data was also very interesting and I learned, that there are unexpected many possibilities to do so. I really had to pick just some examples because I had no time and space to use all of my data for an exhaustive analysis. On the other side, I also experienced that science is hard work. It needs much knowledge and time to plan the experiments and even more time and knowledge to analyze the data properly. Then all this also needs to be comprehensible documented and published to the scientific community, including an accurate literature research for other relevant publications in this scientific field. It seems to me that writing down the whole process as a scientific paper is the hardest part of scientific work. But all this makes science so varied and exciting. Overall it is never boring because the tasks change relatively quickly and you never have to repeat just one step or task the whole time. I really enjoyed going through the steps of scientific work, allowing me to experience the work of a scientist especially in a field I did not knew well before. I planned to study biology before the internship anyway, but it really confirmed me with this plan. Now I am sure, that I actually want to become a scientist and that I really enjoy this work. I like exploring and am curious, what really fits with science. Finding out new things and always do a little progress in knowledge is very motivating. I am still not sure which sub-discipline interests me most, but I now know, that molecular plant physiology interests me, broadening my horizons. Now I think, that I am more likely to do science related lab-work, than for example ethology. All in all this internship helped me to find my career aspiration by giving me the chance to experience new fields and tasks and ensured me with my plan to study biology and become a scientist.

7. Acknowledgement

I want to thank my supervisor Dr. Saleh Alseekh for his help and guiding me through this internship and my project. He explained me everything, so that I was able to do some real experiments and scientific work as one of the lab members. He prepared a project for me so that I was able to work relatively autonomous. Also thanks to Mustafa Bulut who helped me with and explained me the data processing and analyzation. Thanks to the whole group of Prof. Dr. Fernie for integrating and supporting me. I also would like to thank Prof. Roberto Papa from Università Politecnica delle Marche Ancona for providing the SNPs data and seeds for this study. Finally, I want to thank the *Förderverein der BiologieOlympiade e.V* for making this internship possible and supporting me during this time. This Internship really helped me to orientate myself in view of choice of study and was a great experience.

8. Bibliography

S. M. Correa, A. R. Fernie, Z. Nikoloski, Y. Brotman; “Towards model-driven characterization and manipulation of plant lipid metabolism”; Progress in Lipid Research, Volume 80; 2020

J. Hummel, S. Segu, Y. Li, S. Irgang, J. Jueppner, P. Giavalisco; “Ultra performance liquid chromatography and high resolution mass spectrometry for the analysis of plant lipids”; frontiers in Plant Science; October 2011.

The INCREASE consortium; “INCREASE – Intelligent Collections of Food Legumes Genetic Resources for European Agrofood Systems”; URL:https://www.increase-h2020.eu, retrieved: 15.08.2020

F. Liu, Y. Xia, L. Wu, D. Fu, A. Hayward, J. Luo, X. Yan, X. Xiong, P. Fu, G. Wu, C. Lu; “Enhanced seed oil content by overexpressing genes related to triacylglyceride synthesis “; Gene 557; (2015)

J. D. Lobaton, T. Miller, J. Gil, D. Ariza, J. Fernando de la Hoz, A. Soler, S. Beebe, J. Duitama, P. Gepts, B. Raatz; “Resequencing of Common Bean Identifies Regions of Inter–Gene Pool Introgression and Provides Comprehensive Resources for Molecular Breeding”; Plant Genome; 2018

M. Quiroz-Sodi, S. Mendoza-Díaz, L. Hernández-Sandoval, I. Carrillo-Ángeles; “Characterization of the secondary metabolites in the seeds of nine native bean varieties (Phaseolus vulgaris and P. coccineus) from Querétaro, Mexico”; Botanical Sciences; 2018

M. Rendón-Anaya, J. M. Montero-Vargas, S. Saburido-Álvarez, A. Vlasova, S. Capella-Gutierrez, J. J. Ordaz-Ortiz, O. M. Aguilar, R. P. Vianello-Brondani, M. Santalla, L. Delaye, T. Gabaldón, P. Gepts, R. Winkler, R. Guigó, A. Delgado-Salinas, A. Herrera-Estrella; “Genomic history of the origin and domestication of common bean unveils its closest sister species”; Genome Biology; 2017

J. Singh, J. Zhao, C. E. Vallejos; “Differential transcriptome patterns associated with early seedling development in a wild and a domesticated common bean (Phaseolus vulgaris L.) accession”; Plant Science; 2018

Y. Yamaryo, K. Motohashi, K. Takamiya, T. Hisabori, H. Ohta; “In vitro reconstitution of monogalactosyldiacylglycerol (MGDG) synthase regulation by thioredoxin”; FEBS Letters 580; 2006

9. Appendix

The complete data tables can be requested from Dr. Saleh Alseekh (Alseekh@mpimp-golm.mpg.de).