Max-Planck-Institut für Molekulare Pflanzenphysiologie
Arbeitsgruppe Prof. Dr. Alisdair Fernie

Canalization of plant metabolism and yield

Dissertation

**zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Molekulare Pflanzenphysiologie"**

**eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Institut für Biologie und Biochemie
der Universität Potsdam**

von

Micha Wijesingha Ahchige

Potsdam, Oktober 2021

Datum der Disputation: 15.02.2022

Summary

Plant metabolism is the main process of converting assimilated carbon to different crucial compounds for plant growth and therefore crop yield, which makes it an important research topic. Although major advances in understanding genetic principles contributing to metabolism and yield have been made, little is known about the genetics responsible for trait variation or canalization although the concepts have been known for a long time. In light of a growing global population and progressing climate change, understanding canalization of metabolism and yield seems ever-more important to ensure food security. Our group has recently found canalization metabolite quantitative trait loci (cmQTL) for tomato fruit metabolism, showing that the concept of canalization applies on metabolism. In this work two approaches to investigate plant metabolic canalization and one approach to investigate yield canalization are presented.

In the first project, primary and secondary metabolic data from *Arabidopsis thaliana* and *Phaseolus vulgaris* leaf material, obtained from plants grown under different conditions was used to calculate cross-environment coefficient of variations or fold-changes of metabolite levels per genotype and used as input for genome wide association studies. While primary metabolites have lower CV across conditions and show few and mostly weak associations to genomic regions, secondary metabolites have higher CV and show more, strong metabolite to genome associations. As candidate genes, both potential regulatory genes as well as metabolic genes, can be found, albeit most metabolic genes are rarely directly related to the target metabolites, suggesting a role for both potential regulatory mechanisms as well as metabolic network structure for canalization of metabolism.

In the second project, candidate genes of the *Solanum lycopersicum* cmQTL mapping are selected and CRISPR/Cas9-mediated gene-edited tomato lines are created, to validate the genes role in canalization of metabolism. Obtained mutants appeared to either have strong aberrant developmental phenotypes or appear wild type-like. One phenotypically inconspicuous mutant of a pantothenate kinase, selected as candidate for malic acid canalization shows a significant increase of CV across different watering conditions. Another such mutant of a protein putatively involved in amino acid transport, selected as candidate for phenylalanine canalization shows a similar tendency to increased CV without statistical significance. This potential role of two genes involved in metabolism supports the hypothesis of structural relevance of metabolism for its own stability.

Zusammenfassung

In the third project, a mutant for a putative disulfide isomerase, important for thylakoid biogenesis, is characterized by a multi-omics approach. The mutant was characterized previously in a yield stability screening and showed a variegated leaf phenotype, ranging from green leaves with wild type levels of chlorophyll over differently patterned variegated to completely white leaves almost completely devoid of photosynthetic pigments. White mutant leaves show wild type transcript levels of photosystem assembly factors, with the exception of ELIP and DEG orthologs indicating a stagnation at an etioplast to chloroplast transition state. Green mutant leaves show an upregulation of these assembly factors, possibly acting as overcompensation for partially defective disulfide isomerase, which seems sufficient for proper chloroplast development as confirmed by a wild type-like proteome. Likely as a result of this phenotype, a general stress response, a shift to a sink-like tissue and abnormal thylakoid membranes, strongly alter the metabolic profile of white mutant leaves. As the severity and pattern of variegation varies from plant to plant and may be effected by external factors, the effect on yield instability, may be a cause of a decanalized ability to fully exploit the whole leaf surface area for photosynthetic activity.

Zusammenfassung

Der pflanzliche Stoffwechsel ist der Hauptprozess, der assimilierten Kohlenstoff in unterschiedliche Stoffe umwandelt, die wichtig für das Pflanzenwachstum und somit den Ertrag sind, weswegen es ein wichtiges Forschungsthema ist. Obwohl große Fortschritte beim Verständnis der genetischen Prinzipien, die zum Stoffwechsel und Ertrag beitragen, gemacht wurden, ist noch relativ wenig über die genetischen Prinzipien bekannt, die für die Variation oder Kanalisierung von Eigenschaften verantwortlich sind, obwohl diese Konzepte schon lange bekannt sind. In Anbetracht einer wachsenden Weltbevölkerung und des fortschreitenden Klimawandels, scheint es immer wichtiger zu sein, Kanalisierung von Metabolismus und Ertrag zu verstehen, um Ernährungssicherheit zu garantieren. Unsere Gruppe hat kürzlich metabolisch kanalisierte quantitative Merkmalsregionen für den Stoffwechsel von Tomatenfrüchten gefunden und damit gezeigt, dass sich das Konzept der Kanalisierung sich auf den Stoffwechsel anwenden lässt. In dieser Arbeit werden zwei Ansätze zu Untersuchung von Kanalisierung des pflanzlichen Stoffwechsels und ein Ansatz zur Untersuchung von Ertragskanalisierung präsentiert.

Im ersten Projekt, wurden Daten von Primär- und Sekundärmetaboliten von *Arabidopsis thaliana* und *Phaseolus vulgaris*, gewonnen von Pflanzen, die unter unterschiedlichen

Zusammenfassung

Bedingungen wuchsen, verwendet, um den Variationskoeffizient (VarK) oder die relative Änderung von Stoffgehalten umweltübergreifend für jeden Genotyp zu berechnen und als Eingabe für genomweite Assoziationsstudien verwendet. Während Primärmetabolite über unterschiedliche Umweltbedingungen einen geringeren VarK haben und nur wenige eher schwache Assoziationen zu genomischen Regionen zeigen, haben Sekundärstoffe einen höheren VarK und zeigen mehr und stärkere Assoziationen zwischen Metabolit und Genom. Als Kandidatengene können sowohl potenziell regulatorische, als auch metabolische Gene gefunden werden, jedoch sind metabolische Gene selten direkt zu den Zielmetaboliten verbunden, was für eine Rolle von sowohl regulatorischen Mechanismen als auch metabolischer Netzwerkstruktur für die Kanalisierung des Stoffwechsels spricht.

Im zweiten Projekt wurden Kandidatengene aus der *Solanum lycopersicum* cmQTL-Kartierung, ausgewählt und CRISPR/Cas9-vermittelte, genomeditierte Tomatenlinien erschaffen, um die Rolle dieser Gene in der Kanalisierung des Metabolismus zu validieren. Erhaltene Mutanten zeigten entweder starke Fehlentwicklungsphänotypen oder erschienen wildtypähnlich. Eine phänotypisch unauffällige Mutante einer Pantothensäurekinase, die als Kandidat für die Kanalisierung von Apfelsäure gewählt wurde, zeigte einen signifikanten Anstieg des VarK über unterschiedliche Bewässerungsbedingungen. Eine andere solche Mutante eines Proteins, welches mutmaßlich im Aminosäuretransport involviert ist, welches als Kandidat für die Kanalisierung von Phenylalanin gewählt wurde, zeigt eine ähnliche Tendenz zu einem erhöhten VarK ohne statistische Signifikanz. Diese potenzielle Rolle von zwei Genen, die im Stoffwechsel involviert sind, unterstützt die Hypothese einer strukturellen Relevanz des Metabolismus für seine eigene Stabilität.

Im dritten Projekt wurde eine Mutante einer mutmaßlichen Disulfid-Isomerase, welche wichtig für die Thylakoidbiogenese ist, durch einen Multiomik Ansatz charakterisiert. Die Mutante wurde vorher in einer Ertragsstabilitäts-Selektierung charakterisiert und zeigte einen panaschierten Blattphänotyp, welcher von grünen Blättern mit Wildtyp Chlorophyllgehalt über unterschiedlich gemustert panaschierte Blätter bis zu komplett weißen Blätter reichte, die fast gar keine photosynthetischen Pigmente enthielten. Weiße Blätter der Mutante zeigen Wildtyp Transkriptlevel von Photosystem-Aufbaufaktoren, mit der Ausnahme von ELIP und DEG Orthologen, was indikativ für eine Stagnation in einer Etioplast-zu-Chloroplast-Übergangsphase ist. Grüne Blätter der Mutante zeigen eine Hochregulierung dieser Aufbaufaktoren, was möglicherweise als Überkompensation für eine partiell defekte Disulfid-Isomerase wirkt und letztlich ausreichend für Chloroplastenentwicklung zu sein scheint, was

wiederum durch ein wildtyp-ähnliches Proteom bestätigt wird. Wahrscheinlich als Effekt dieses Phänotyps ändern, eine generelle Stressantwort, eine Umschaltung zu einem Senke-ähnliches Gewebe und abnormale Thylakoidmembranen, stark das metabolische Profil von weißen Blättern der Mutante. Da der Schweregrad und das Muster der Panaschierung von Pflanze zu Pflanze unterschiedlich ist und durch äußere Faktoren beeinflusst sein könnte, könnte der Effekt auf die Ertragsstabilität eine Folge einer dekanalisierten Fähigkeit sein die ganze Blattoberfläche für photosynthetische Aktivität zu nutzen.

Acknowledgments

Abbreviations

ABA	abscisic acid
ADK	adenosine kinase
ANOVA	analysis of variance
APCI	atmospheric pressure chemical ionization
APPI	atmospheric pressure photoionization
ARF	ADP-ribosylation factors
ARLB	ARF-like GTPase
ATP	adenosine triphosphate
bHLH	basic helix-loop-helix
BIL	backcross inbred line
bp	base pair
CANA1	canalization1
Cas	CRISPR associated
CBC	Calvin-Benson-cycle
CDKB	cell division protein kinase B
cmQTL	canalized metabolic QTL
CO_2	carbon dioxide
CoA	coenzyme A
CRISPR	clustered regularly interspaced short palindromic repeat
crRNA	CRISPR RNA
CV	coefficient of variation
cv.	cultivar
DEG	differentially expressed genes
DEG	degradation of periplasmic proteins
DEPC	diethyl pyrocarbonate
DGDG	Digalactosyldiacylglycerol
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
ELIP	early light inducible protein
EMMA	efficient mixed model association
EMS	ethylmethanesulfonate
ESI	electrospray ionization

Abbreviations

FDR	false discovery rate
FT-MS	fourrier-transform mass spectrometer
GBS	genotyping-by-sequencing
GC	gas chromatography
GDU1	glutamine dumper1
GO	gene ontology
gRNA	guide RNA
GWAS	genome wide association studies
GxE	genotype-by-environment effect
H2O	water
HCl	hydrochloric acid
IL	introgression line
indel	insertion/deletion
kb	kilobases
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	liquid chromatography
L-DOPA	levodopa
LEA	late embryogenesis abundant
LiCl	lithium chloride
LOD	logarithm of odds
LOG2	loss of GDU2
LSTRAP	large-scale transcriptome analysis pipeline
LUL2	LOG2-like uqiquitin ligase
MBF	multiprotein bridging factor
MeJA	methyl jasmonate
MEOH	methanol
MGDG	Monogalactosyldiacylglycerol
MGL	methionine gamma-lyase
MLM	mixed linear model
MPIMP	Max Planck Institute of Molecular Plant Physiology
MPIPZ	Max Planck Institut für Pflanzenzüchtungsforschung
mRNA	messenger RNA
MS	mass spectrometry

Abbreviations

MS	Murashige & Skoog
MTBE	methyl tert-butyl ether
NaCl	sodium chloride
NADH	nicotinamide adenine dinucleotide
NADP	nicotinamide adenine dinucleotide phosphate
NaOAc	sodium acetate
NGS	next-generation sequencing
NMR	nuclear magnetic resonance
NPP	net primary production
PANK	pantothenate kinase
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
Pg	peta (= 10^9) gram
QC	quality control
QRILC	quantile regression imputation of left-censored data
QTL	quantitative trait loci
RFLP	restriction fragment length polymorphism
RH	relative humidity
RNA	ribonucleic acid
RNAseq	RNA sequencing
SCO2	snowy cotyledons2
SNP	single nucleotide polymorphism
TAE	tris-acetate EDTA
TAG	triacylglycerol
TCA	tricarbocylic acid
T-DNA	transfer DNA
TIP	transposable element insertion
TOF	time-of-flight
tpm	transcripts per kilobase million
tracrRNA	transactivating crRNA
UPLC	ultra performance liquiq chromatography

Declaration

I hereby declare, that this work has been submitted to no other university and has been produced autonomously by me and only with the stated means.

Potsdam,

Erklärung

Hiermit erkläre ich, dass die Arbeit an keiner anderen Hochschule eingereicht sowie selbständig von mir und nur mit den angegebenen Mitteln angefertigt wurde.

Potsdam, den

Table of contents

Table of contents

Table of contents

Table of contents

Chapter 1: General Introduction

1.1. Plant metabolism

As autotrophs, plants, green algae and cyanobacteria are all capable to use oxygenic photosynthesis, to convert solar energy into chemical energy and assimilate carbon dioxide (CO_2) into organic matter (Knapp et al., 2014; Nelson & Ben-Shem, 2004). The sum of all energy fixed by autotrophs minus the energy needed for their own respiration is called net primary production (NPP) (Knapp et al., 2014). Terrestrial and marine net primary production, is estimated to be 56.4 Pg of C and 48.5 Pg of C respectively, and therefore each contribute about half to global net primary production (Field et al., 1998). Given that producers make up roughly 95% of terrestrial biomass and the fact that the terrestrial environment is considered to be dominated by plants, likely makes plants the most important primary producers on the planet (Bar-On et al., 2018; Delwiche & Cooper, 2015).

The complex process which facilitates this production is photosynthesis, which can be divided into electron-driven light reactions, that generate ATP and NADPH and the Calvin-Benson cycle (CBC), which is fueled by these compounds to assimilate CO_2 (Baslam et al., 2020). This process is arguably the most important biochemical reaction on the planet, given that it supports the growth of almost all organisms (Y. Wang et al., 2015). The reactions and metabolites involved in the CBC are just a few of the thousands, which together make up the network of plant metabolism (de Oliveira Dal'Molin et al., 2010).

The core part of this network, called primary or central metabolism, gives rise to important compounds, which are crucial for plant survival (Pott et al., 2019). Branching from primary metabolite precursors, such as sugars and amino acids, plants also produce so-called secondary or specialized metabolites, which are more related to the plants capability to adapt to or interact with the environment (Maeda, 2019; Pott et al., 2019). The functions of specialized metabolites are extremely diverse, but they can roughly be attributed to protection against abiotic stresses and interaction with other organisms (Kessler & Kalske, 2018). So far 200000 specialized metabolites have already been described but the true number is estimated to be up to 1 million (S. Wang et al., 2019).

Several methods and machines can be used to detect and qualitatively or quantitatively measure metabolites in plant tissue extracts. Methods are broadly separated into systems based on nuclear magnetic resonance (NMR) or mass spectrometry (MS) (Alseekh & Fernie, 2018). NMR measures the absorbed and re-emitted electromagnetic radiation of atoms in a strong

magnetic field and uses it to identify atoms in molecules (Fernie & Tohge, 2017). Mass spectrometry on the other hand uses their mass to charge-ratio to identify compounds (Nalbantoglu, 2019).

There are three basic steps to the most common MS-based workflows, namely separation, ionization and detection, for each of which different approaches are available (Nalbantoglu, 2019). Although separation-free methods also exist, gas chromatography (GC) or liquid chromatography (LC) is often used as a first step to make up for the lack of separation in the mass spectrometers and to eliminate interactions of the analyzed compounds with the background (Ren et al., 2018). Different ionization methods can be used, depending on the aim and separation method. Electron impact ionization can be used to connect gas chromatography with mass spectrometry (Fernie & Tohge, 2017). For LC-MS for example, the electrospray ionization source (ESI), atmospheric pressure chemical ionization (APCI) and atmospheric pressure photo-ionization (APPI) can be used (Pitt, 2009). Finally, ionized metabolites are captured by different detectors, like "time of flight" (TOF), quadrupole, ion trap and orbitrap mass analyzers (Nalbantoglu, 2019).

Both central and specialized metabolism of plants have been studied well for differing reasons. As central metabolism is directly connected to plant growth, understanding and engineering it promises the possibility to improve crop yield (Sweetlove et al., 2017). Specialized metabolites on the other hand are not only beneficial for the plant, but many so-called phytonutrients are suggested to help in the prevention of diseases (Martin, 2013), while other compounds are important flavor components (Tieman et al., 2006).

While, early metabolomics experiments dealt with profiling of transgenic plants or trying to find biomarkers for plant performance, recent research has shifted towards establishing metabolite-to-gene associations (Alseekh & Fernie, 2018). These associations were studied, through knock-out studies, quantitative trait loci (QTL) mapping and genome wide association studies (GWAS) (Alseekh & Fernie, 2018). As causal genes for the level of certain metabolites, often enzymes, relevant in their biosynthesis or degradation, or transporters were discovered (S. Wu et al., 2016; Zhao et al., 2019). Other mapping studies sometimes also found transcription factors, involved in the regulation of metabolites (Wen et al., 2014; Ye et al., 2019).

Despite this knowledge, engineering primary metabolism has proven difficult so far, likely due to its multilevel regulation and high interconnectivity of metabolites (Sweetlove et al., 2017).

In specialized metabolic pathways however, some successes could already be obtained. Recently, it has been achieved to significantly increase the level of anthocyanins in tomato fruit, by expressing two snapdragon transcription factors (Butelli et al., 2008). Another engineering effort, resulted in tomato fruit accumulating L-DOPA, when expressing a beetroot enzyme, which is normally part of the pathway of betalain production (Breitel et al., 2021).

1.2. QTL mapping and GWAS

Genetic and phenotypic variance within a species, can be used to find a genotype to phenotype relationship. The two main methods used to establish this connection are quantitative trait loci (QTL) mapping and genome wide association studies (GWAS) (Korte & Farlow, 2013). QTL mapping uses introduced genetic variation from recombination, which results from crossing two parental lines followed by selfing or sibling mating the resulting population to create an F_2 generation, which can be directly used or further propagated to create recombinant inbred lines (Takuno et al., 2012). GWAS on the other hand uses genetic variation, which is already present in the form of single nucleotide polymorphisms (SNPs), which makes it also very suitable for human population genetics (Bush & Moore, 2012). Both GWAS and QTL mapping work through similar principles. In GWAS a linear model or $\chi 2$-like model is used to calculate an association between phenotypic variation and genomic variants (Burghardt et al., 2017). In QTL mapping, quantitative trait variation is associated with markers, by interval mapping combined with multiple regression (Zeng, 1994).

Both methods have widely been used in plant biology. In *Arabidopsis thaliana* for example, QTL mapping has helped to discover candidate genes related to flowering time and plant growth (Koornneef et al., 2004). Similarly GWAS has for example been used to investigate primary and secondary metabolism (S. Wu et al., 2016, 2018). Recently, the two methods have also been used in conjunction, which facilitates a more accurate candidate gene selection (Asekova et al., 2021; He et al., 2017). A prerequisite for conducting these experiments, however is the availability of suitable model plants, which will be discussed in the next section.

1.3. Model plants for metabolomics

The self-pollinating, flowering plant *Arabidopsis thaliana*, which can easily be cultivated under controlled conditions, has 5 chromosomes, a relatively small genome (~28000 genes) and a generation time of 6-8 weeks, making it an ideal model species for plant science (Serino & Marzi, 2018). The continued collaborative effort of the *Arabidopsis thaliana* research community, has given rise to a large pool of resources, including stock centers and databases

(Diaz, 2019; Koornneef & Meinke, 2010). To date, more than a thousand wild ecotypes from around the world have been fully sequenced, which largely facilitates comparative genetic experiments (Alonso-Blanco et al., 2016). These available genetic resources have for example been exploited in genome wide association studies for both primary and secondary metabolites, which identified novel associations (S. Wu et al., 2016, 2018). By using enzyme activity as input, GWAS has also identified trans-QTL which are suggested to have regulatory roles (Fusari et al., 2017). A library of T-DNA insertions across the whole genome, which contains insertional mutants of a large portion of all genes, has been generated, which has played an important role in *Arabidopsis* research (Alonso et al., 2003; O'Malley et al., 2015).

Another plant, which has widely been used for metabolomics experiments is the cultivated tomato *Solanum lycopersicum* (G. Zhu et al., 2018). With roughly 35000 genes (Hosmani et al., 2019) and a regular plant size of more than one meter and a roughly 4 month generation time, of most cultivars (Campos et al., 2010), may make tomato appear as a less desirable, model plant. Biologically speaking, a distinct feature of tomato, however is that it produces fleshy fruit, which is why it is a suitable model plant for fruit bearing crops (Kimura & Sinha, 2008). Fresh tomato fruits, or products thereof, like tomato ketchup, paste or sauce, make up a major part of the daily diet in many countries, which has made tomato the most highly produced vegetable worldwide with a large economic importance (Anwar et al., 2019). As arable land per person decreases, new vertical farming technologies are employed (Benke & Tomkins, 2017) and new tomato varieties are being developed, to grow under these space-restricting conditions (Kwon et al., 2020), which could lead to another increase in overall tomato production.

Within the Solanum genus, the cultivated tomato *Solanum lycopersicum* belongs to the Lycopersicon section together with 12 other wild relative species and 4 more species in the sections Lycopersicoides and Juglandifolia (Peralta et al., 2008). In order to regain some of the genetic diversity that tomato varieties have lost during their domestication, cultivated varieties are commonly crossed back to wild varieties (Menda et al., 2014). However, this availability of wild relatives has not only been key for exploiting genetic diversity for the breeding of modern varieties (Mata-Nicolás et al., 2020), but it has also facilitated the development of different mapping populations (Canady et al., 2005; Eshed & Zamir, 1994; Ofner et al., 2016).

By crossing the cultivated tomato *Solanum lycopersicum* to its wild green-fruited relative *Solanum pennellii* followed by several backcrosses and selfings, a so-called introgression line

(IL) population was created, originally consisting of 50 lines with single, marker-defined, homozygous substitutions of the cultivated tomato genome background, by homologous *Solanum pennellii* genomic content (Eshed & Zamir, 1994). Another mapping population, has been generated in a similar fashion by several rounds of backcrosses and selfing, leading to 446 lines with several but much smaller genomic introgressions, resulting in a higher mapping resolution (Ofner et al., 2016). These mapping populations have widely been used to study agronomical traits, like total yield and total soluble solids content but also leaf morphology and flowering time (Eshed & Zamir, 1995; Fulop et al., 2016; Gur et al., 2011; Gur & Zamir, 2015). Similarly, our own group has used these resources extensively for metabolomics experiments (Alseekh et al., 2015, 2017; Kuhalskaya et al., 2020; Schauer et al., 2006, 2008). Besides primary and secondary metabolites, an interesting group of compounds produced by tomato fruits are volatile compounds and increasing efforts are allocated to understanding their regulation and production as they are major components contributing to flavor (Rambla et al., 2014, 2017).

Another staple crop of high economic and nutritional importance is, the common bean *Phaseolus vulgaris*, which serves as a major protein source, especially in South America and Eastern and Southern Africa (Broughton et al., 2003). Common bean is an annual herbaceous plant, with a diploid genome, which makes it a suitable model plant for other legumes like soybean (*Glycine max*) (Nadeem et al., 2021). As a shift towards more plant-based diets has been identified as an important tool to mitigate global warming and to fight malnutrition, studying legumes, which are a suitable substitute for animal-based food, seems ever more important (Ferreira et al., 2021). Besides the protein, beans also contain, carbohydrates, secondary metabolites and a low amount of lipids (Martino et al., 2012). Common bean is believed to have undergone two separate domestication events in Mesoamerica and the Andes mountains, leading to two separate genepools, which have been hybridized during their introduction into Europe, leading to even greater genetic diversity (Raggi et al., 2019). Currently, around 10000 accessions have genotyping-by-sequencing (GBS) information and 220 accessions have whole genome sequencing information, while several other panels exist, genotyped by different approaches (Cortinovis et al., 2021). Although plenty of traits in common bean have been studied through GWAS approaches (Oladzad 2019) and protein and oil content has been assessed in soybean by GWAS (Hwang et al., 2014), the application of GWAS for metabolites of common bean has not yet been thoroughly exploited . Recently

however, beans are also starting to be used for metabolomics studies (Niron et al., 2020; Souza et al., 2019).

1.4. Phenotypic plasticity, canalization, variation

On purpose, the preceding paragraphs were focused on looking at traits in general as a single fixed or mean value. However in reality each trait will have a certain variation, depending from which angle I look on it. An early perspective to look at variation was offered by Waddington with the term canalization (Waddington, 1942). He described developmental processes as generally canalized, as they converge to a certain final state, irrespective of minor variations in the conditions during the development. A more general look on variation of any phenotype and their sources are provided by (Laitinen & Nikoloski, 2018). If I consider plants from different genotypes, grown under different environments, I can observe 3 types of variation for any specific quantitative trait (Figure 1).
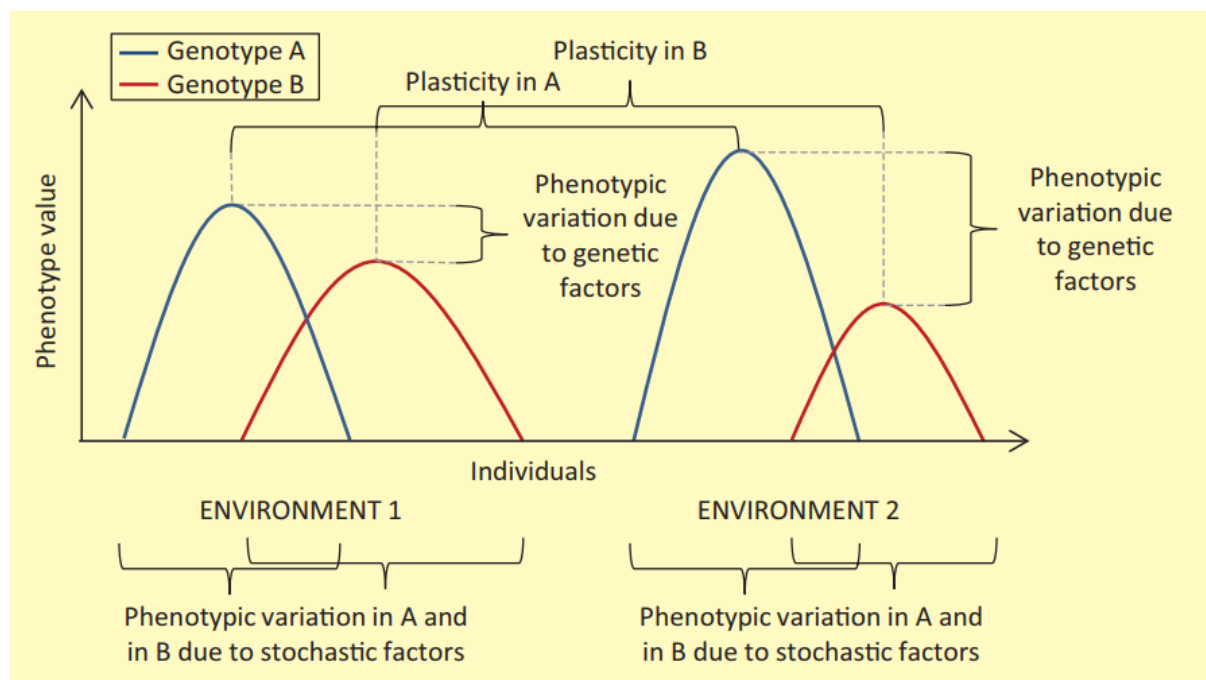
Figure 1: Types of variation between individuals, genotypes and environments (Laitinen & Nikoloski, 2018)

As an example, let us think about growing tomato plants from two cultivars (genotype) in two different locations (environment) and measuring a trait like fruit yield per plant (phenotype value). First of all, I will observe some variation of fruit yield between plants of the same genotype, within one environment. As the plants are genetically identical, this variation will likely be due to stochastic factors. The second type of variation, is the difference of the mean level of the trait between the two cultivars in either of the environments, which is due to genetic factors. The third type of variation is the variation of the trait level between the two

environments for each of the two cultivars. This type of variation is also referred to as phenotypic plasticity (Laitinen & Nikoloski, 2018). While there are many different ways to estimate variation, a relatively simple metric is the so-called coefficient of variation (CV), which is a dimensionless value consisting of the standard deviation divided by the mean of a certain trait (Laitinen & Nikoloski, 2018).

Variation and robustness has been studied from simple organisms like bacteria or yeasts to more complex organisms like crop plants. For example, in bacterial chemotaxis, the adaptation precision remains robust under varying concentrations of one key protein of the adaptation mechanism, while two other traits, adaptation time and steady-state tumbling frequency, change plastically (Alon et al., 1999). In yeast it has been shown by knocking out non-essential genes, that environmental, genetic and stochastic robustness are correlated (Lehner, 2010). Another study in yeast found out, by reducing gene expression of essential genes, that they are even more important, than non-essential genes for phenotypic robustness (Bauer et al., 2015). A study in *Arabidopsis thaliana* found that variation of hypocotyl length in dark grown seedlings is mediated by a gene from the BZR/BEH gene family, likely in connection to HSP90 (Lachowiec et al., 2018). In both archived tomato mapping data and a "crop garden" experiment, with additional plants like eggplant, watermelon, sunflower and maize, a bimodal distribution of stable and plastic traits was found (Fisher et al., 2017). In a large experiment of 976 maize hybrids, grown under 11 environments, the plasticity of 12 agronomic traits were mapped, yielding hundreds of QTL responsible for phenotypic plasticity (N. Liu et al., 2021).

This phenotypic plasticity, studied in crop plants, is especially interesting for the application in agriculture. Not only is it interesting to breed cultivars for different regions of the world or for different applications, as to maximize production efficiency, but it may also become more and more important to breed crops which have a certain robustness to changing climates, as increasing climate variability is connected to decreasing yield stability (Reckling et al., 2021). A large metastudy has shown that already, around a third of global yield variability of maize, wheat, rice and soybean, can be attributed to climate variation (Ray et al., 2015). Considering the same crops, which provide around 75% of globally consumed calories, a different study has projected already median yield losses of 3%-12% mid-century, under a rigorous warming scenario (Wing et al., 2021).

Our group has wondered, whether the variation of metabolites between different environments can also be traced back to genomic regions of plants. For that purpose, metabolomics data from

the tomato introgression line population grown in three seasons in the field, were used together with the genetic information to perform QTL mapping (Alseekh et al., 2017). Not using metabolite level, but its variation, several loci could be identified which had a significant correlation to the variation of a metabolite (Figure 2).
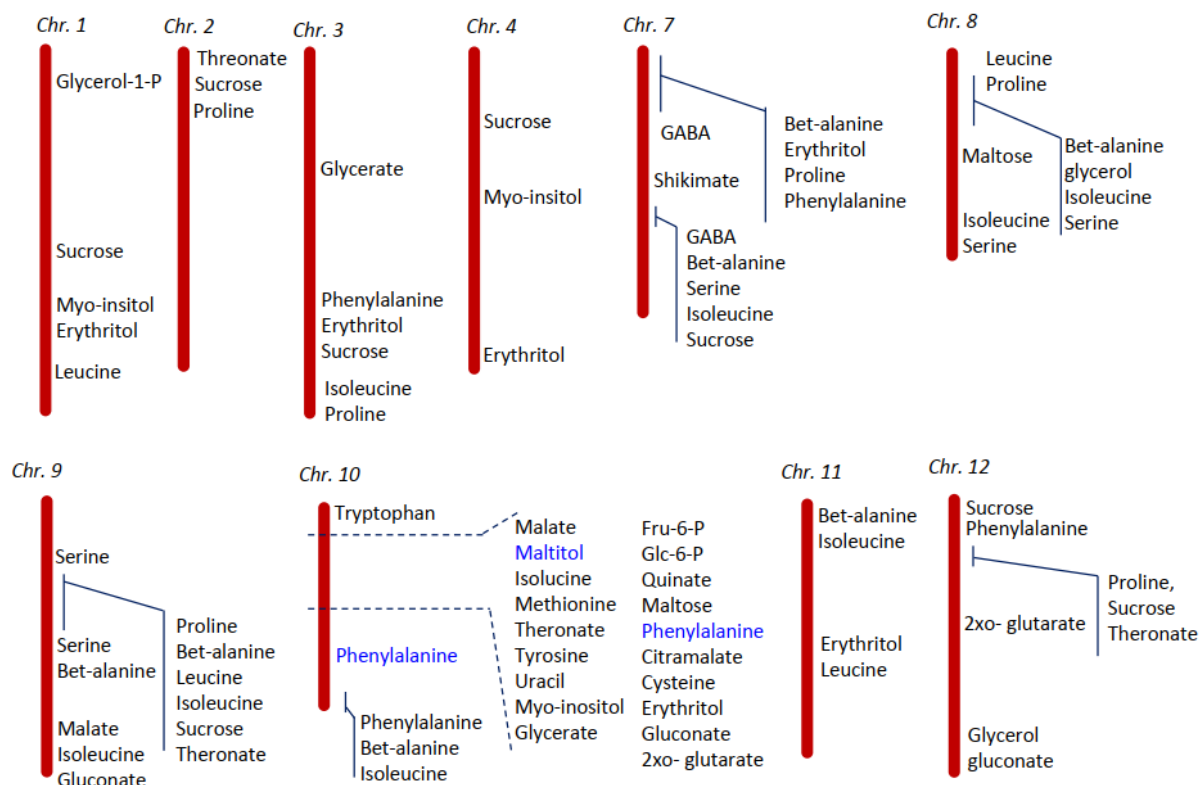
Figure 2: Identified cmQTL across chromosomes of tomato using data from multiple field trials of S. pennellii ILs. To account for multiple hypothesis testing bonferroni correction was applied. Metabolites which had significant cmQTL even after the correction at a significance threshold of 0.05 are marked in blue. (Alseekh et al., 2017)

Chromosome 10 seemed to have some hotspots, which had associations to several metabolites. A subset of the backcross inbred line population has been used to validate the findings and narrow down the relatively large genomic loci with a significant correlation to metabolite variation. This yielded loci, containing 20-40 genes, which are prime candidate genes for further validation (Alseekh et al., 2017). This experiment proved that the concept of canalization could also be applied to metabolism.

1.5. Aim of the thesis

Building on the knowledge gained from the previously mentioned experiments, I wanted to further understand canalization of metabolism, but also yield and on this basis three projects were developed.

Chapter 1: General Introduction

First of all I wanted to replicate the aforementioned procedure of cmQTL mapping in other species and populations with a GWA approach. Secondly I wanted to select candidate genes from the already found cmQTLs and try to validate their role in metabolic canalization. Finally I wanted to investigate a yield canalization mutant, discovered by our collaborator. In the following chapters I am addressing these projects one-by-one.

Chapter 2: cmGWAS

2.1. Introduction

Genome wide association studies (GWAS) and its traditional counterpart linkage analysis are the most commonly used strategies to identify causal genes of trait variation, not only in plants but also in animals and humans (Sukumaran & Yu, 2014). In principle both methods associate trait variance with different markers across the genome (Alseekh et al., 2021; Takuno et al., 2012). Different kinds of genetic markers exist. Early markers like restriction fragment length polymorphism (RFLP), were followed by PCR-based markers and finally SNP-based markers, which have widely become available with the ever-decreasing costs for sequencing through next-generation sequencing technologies (NGS) (Zargar et al., 2015).

In contrast to traditional QTL mapping, GWA approaches are not as limited in terms of allelic diversity and mapping resolution defined by the parental lines, but face additional limitations (Korte & Farlow, 2013). While GWAS is suitable to find the genetic basis for traits with a simple genetic architecture, traits with a more complex architecture (e.g. rare variants or small effect sizes) may be difficult to be elucidated by GWAS (Korte & Farlow, 2013). As groups of individuals may be more closely related, due to their geographical origin or local adaptation, this can lead to an unequal genetic relationship between those groups and if this population structure is left unaccounted for, this could lead to false positive results (Sukumaran & Yu, 2014). Several different single-locus or multi-locus models can be used to calculate associations between trait variation and SNPs (Kaler et al., 2020). Mixed linear models (MLM) are able to correct for bias in population stratification and inflation from many small genetic loci and are therefore the most popular method for GWAS (Fang & Luo, 2019).

GWAS and QTL mapping have both widely been used to investigate many traits in different sections of plant science (Fang & Luo, 2019; Koornneef et al., 2004). Recently, using quantitative genetic experiments to study the robustness of traits across varying environments or between different individuals is also starting to gain more attention (Jimenez-Gomez et al., 2011; Kikuchi et al., 2017; Tan et al., 2020). GWAS has also widely been used to elucidate metabolic pathways. Studies have shown that content of primary metabolites is usually determined by many small effect loci, while secondary metabolite content is determined by few large effect loci (Fang & Luo, 2019). Both enzymes and transcription factors have been found as causal genes for metabolite content, by GWAS (Fusari et al., 2017; S. Wu et al., 2016, 2018; Ye et al., 2019). However the potential of genomic mapping to study the robustness of

metabolism across different environments has so far remained underexploited (Alseekh et al., 2017).

In the following chapter I am using the coefficient of variation or fold-change of metabolic content of both primary and secondary metabolism in *Arabidopsis thaliana* and *Phaseolus vulgaris* across different environments as an estimator for metabolic canalization. I use this estimator as an input for a genome wide association study to discover novel loci, which may be responsible for the control of metabolite levels in regard to different environments.

2.2. Materials and Methods

2.2.1. Ecotypes

Different sets of ecotypes were used for GWAS. For the *Arabidopsis thaliana* GWAS, the HapMap Panel consisting of 350 natural accessions (Thoen et al., 2017) was used, which is a subset of the larger RegMap Panel consisting of 1307 natural accessions(Horton et al., 2012; Y. Li et al., 2010). From the RegMap Panel, which had previously been used for mGWAS (S. Wu et al., 2016, 2018), 315 ecotypes could be grown until maturity and used for analysis. For the common bean GWAS a set of 200 *Phaseolus vulgaris* landraces, containing 100 American and 100 European landraces, was selected for which whole genome sequencing data was available (Bellucci et al., in preparation).

2.2.2. Growth conditions

Growth conditions varied, depending on the specific experiment. The first experiment was planned and started by Fayezeh Arabi and partly analyzed by myself. In this experiment concerning acclimation to high light, 315 *Arabidopsis thaliana* ecotypes were sown onto soil. After germination, 2 plants per ecotype were transplanted into individual pots and cultivated for 6 weeks under a controlled environment of a York walk-in Phytochamber (York International/Johnson Controls; Cork, Ireland). During this time, plants received 150 µmol m^{-2} s^{-1} of light in a day/night cycle, with an 8h/16h photoperiod and a 22°C/16°C temperature regime. After 6 weeks, one plant of each ecotype was transferred to a phytochamber supplying 600 µmol m^{-2} s^{-1} and acclimated for 3 hours. After that time, for each ecotype, a single biological replicate from both light conditions was collected as a sample.

A separate experiment was performed by Feng Zhu and recently published (F. Zhu et al., 2021). In this experiment concerning metabolite catabolism under extended darkness, a similar set of 288 *Arabidopsis thaliana* ecotypes was used. In autumn 2018, plants were sown to soil in a controlled chamber of the MPIMP greenhouse (Potsdam-Golm, Germany) and transplanted to

individual pots after 2 weeks and cultivated for 3 more weeks. Plants were grown under short day conditions with an 8h light/16h dark short day photoperiod, at 250 µmol m^{-2} s^{-1} of light a 20°C/16°C day/night temperature and 60%/70% humidity. After 35 days one plant per genotype was harvested as a control and each one after 3 days and 6 days of darkness. This experiment was repeated independently in spring 2019. Samples of the same ecotype in both experiments were treated as two biological replicates.

Three separate experiments were conducted with 201 landraces of *Phaseolus vulgaris* grown in a Polytunnel with drought conditions or in a greenhouse with long day (16 h) or short day conditions (8h) (Bellucci et al., in preparation). Samples were collected from 1-3 biological replicates.

2.2.3. Sample collection

Plant tissue samples were generally, shock frosted after collection and stored at -80°C until further processing. For *Arabidopsis thaliana* experiments, the root was cut at soil level and the whole leaf rosette was collected and for *Phaseolus vulgaris* experiments a single leaflet per plant was collected.

2.2.4. Metabolite extraction

Frozen plant tissue was ground to powder using a Mixer Mill (Retsch; Haan, Germany) at 30 Hz for 1 min. Aliquots of 50 mg were used for metabolite extraction, as described before (Salem et al., 2016). Here, 1 ml of pre-cooled MTBE-MeOH (3:1;vol/vol) extraction buffer is added to each sample and samples are incubated 10 minutes on an orbital shaker at 4°C. Samples are sonicated for 10 minutes in an ice bath before 0.5 ml of MeOH-H$_2$O (3:1; vol/vol) is added and samples are centrifuged at 14000 rpm for 5 minutes at 4°C, leading to a phase separation. An aliquot of the upper (apolar) phase was taken for lipid analysis, the rest of the upper phase was aspirated with the BVC fluid aspiration system (Vacuubrand Inc; Essex, CT, U.S.A.) and two aliquots of the lower (polar) phase were taken for GC-MS and LC-MS. Extracts were dried in a Scan Speed 40 centrifugal vacuum concentrator (Labogene; Allerød, Denmark) coupled to a Scanvac CoolSafe cryo unit (Labogene; Allerød, Denmark) with 1000 g and 30°C for 3h or overnight. Dried extracts were kept at -80°C until further use.

2.2.5. Metabolite profiling

Primary metabolites were analyzed as described before (Lisec et al., 2006). For the analysis via GC-MS-TOF a dried aliquot of the polar phase was derivatized by adding 40 µL of (20 mg * mL^{-1}) methoxyamine hydrochloride and incubating at 37°C first for 2 h followed by addition

of 70 µL of MSTFA and continued incubation at 37°C for 30 min. The samples were injected by an autosampler Gerstel Multi-Purpose system (Gerstel GmbH & Co.KG, Mülheim an der Ruhr, Germany) into a gas chromatograph coupled to a time-of-flight mass spectrometer (Leco Pegasus HT TOF-MS) (LECO Corporation; St. Joseph, MI, U.S.A.). Analysis of secondary metabolites followed a previously described protocol (Giavalisco et al., 2009). Here a dried aliquot of the polar phase was resuspended in 150 µl H2O:MeOH (50:50) and samples injected to an Acquity UPLC system (Waters Corporation; Milford, MA, U.S.A.) coupled to an Exactive Orbitrap mass detector (ThermoFisher Scientific; Waltham, MA, U.S.A.) via a heated electrospray source (ThermoFisher Scientific; Waltham, MA, U.S.A.) . Mass spectra were obtained by running samples in negative ionization mode. For analysis of lipophilic compounds a previously established protocol was used (Hummel et al., 2011). An aliquot of the dried organic phase was resuspended in 100 µL of UPLC-grade acetonitrile:isopropanol (70:30) mix, of which 2 µl were injected onto an Acquity UPLC system (Waters Corporation; Milford, MA, U.S.A.) equipped with an RP C8 column (Hummel et al., 2011). Mass spectra were obtained by running samples in positive ionization mode on an Orbitrap high-resolution mass spectrometer: Fourier-transform mass spectrometer (FT-MS) coupled with a linear ion trap (LTQ) Orbitrap XL (ThermoFisher Scientific; Waltham, MA, U.S.A.).

2.2.6. Peak picking/Area calculation

For targeted analysis peaks were picked manually and peak area calculated with Xcalibur Version 4.2.47 (ThermoFisher Scientific; Waltham, MA, U.S.A.). Non-targeted analysis was performed with Genedata Expressionist ® 14.0.5 (Genedata; Basel, Switzerland).

2.2.7. Metabolite data normalization

Missing values were imputed by the half-minimum of a certain metabolic feature in the respective run or by using a QRILC approach as suggested by literatures (Wei et al., 2018). Raw peak area of each metabolic feature was then normalized to the area of an internal standard (ribitol for GC-MS and isovitexin for LC-MS) if applicable. To account for batch and drift effects a simple linear model was fitted to pooled quality control samples in each batch, which were regularly injected. For slopes with a p-value ≤ 0.05 and an adjusted $R^2 \geq 0.75$ or 0.8 the predicted values were used for a normalization, otherwise the median value of the pooled quality control samples was used. Further normalization, was done by resuspension volume and sample weight. Values of each metabolite were subjected to Box-Cox transformation (Box & Cox, 1964) to bring them closer to a normal distribution.

2.2.8. SNP data source

For GWAS in *Arabidopsis thaliana* publicly available imputed SNPs with an imputation accuracy of ≥ 95% (Arouisse et al., 2020) was used and filtererd for a minor allele frequency of 0.05, leaving around 1.2 million SNPs. For GWAS in *Phaseolus vulgaris* an as of yet unpublished set of whole genome sequencing markers was used, containing around 3.6 million polymorphic SNPs when filtering for a minor allele of 0.05.

2.2.9. GWAS

A mixed-linear model was used to calculate phenotype-to-genotype associations. Variance components were estimated by the efficient mixed model association ("EMMA") method, which corrects for population structure and genetic relatedness of the sample population (Kang et al., 2008).

2.2.10. Computational analysis and used packages

All analysis was performed using R statistical software v4.0.2 in the RStudio environment or on a unix-based high performance computing cluster. Code was written in base R or with the help of the packages tidyverse, broom, data.table, modelr (Dowle et al., 2021; Robinson et al., 2021; Wickham et al., 2019; Wickham & RStudio, 2020). GWAS was performed using rMVP v1.0.6 (L. Yin et al., 2021). Imputation by QRILC was done with the package imputeLCMD (Lazar, 2015). Heatmaps were created with the pheatmap package (Kolde, 2019).

2.3. Results

From the available datasets, I generated different new datasets, by calculating the metabolite fold-change or CV of each ecotype across the selected conditions. Table 1 gives an overview of the datasets from the different experiments, which were conducted.

Table 1: Available datasets, which were combined as input for GWAS

Species	Experiment	Condition	# genotypes
A. thaliana	Light acclimation	Normal light	310
		High light	
	Darkness	0 days (control)	264
		3 days	
		6 days	
P. vulgaris	Day length	Short day	201
		Long day	

	Drought	Drought	

I used these datasets to carry out individual GWAS. I always considered associations at a strict bonferroni cut-off and a more relaxed LOD threshold corresponding to $-\log_{10}(1/\text{number of markers})$, as it has been suggested that the bonferroni cut-off may be too strong and this marker-defined threshold has been shown to allow detecting true positive results (S. Wu et al., 2016).

2.3.1. _Arabidopsis thaliana_ primary metabolites CV across darkness conditions

For the first dataset I calculated the ecotype-wise CV across the conditions of 0, 3 and 6 days. Average values of the CV range from 0.14 to 1.64, with only 8 compounds with a CV higher than 1 (Figure 3).
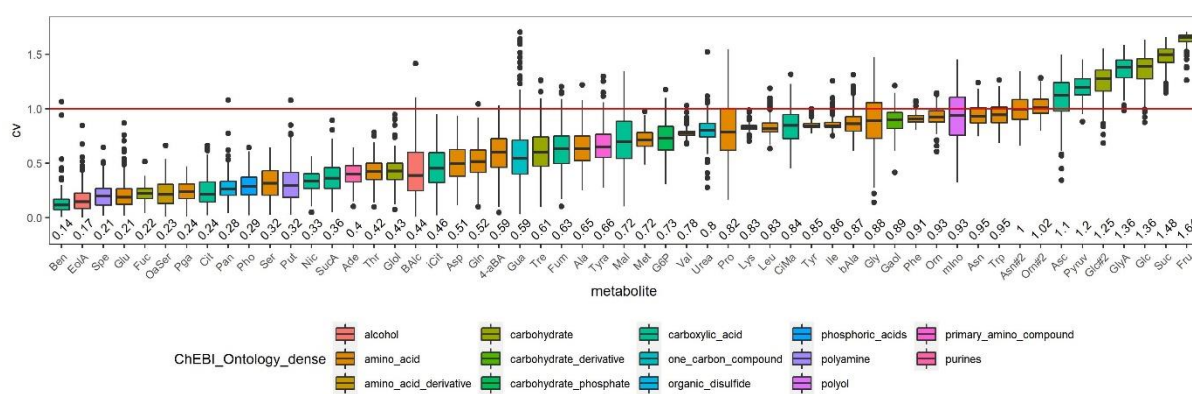
Figure 3: Boxplot of raw CV values of metabolites across 0, 3 and 6 days of darkness of all used ecotypes. The red horizontal line denotes a CV of 1. Numbers along the x-axis show mean CV.

Individual values range however from close to 0 to around 1.7 (Figure 4). Different clusters of metabolites with different CV profiles can be seen. The highest CV values can be found in the lowest cluster containing only carbohydrates and carboxylic acids. The cluster above this mainly contains amino acids with moderately low to moderately high CV values. The largest cluster with a diverse set of metabolites has majorly low to very low CVs across all ecotypes. The last cluster also contains diverse compounds with a higher variation between low and moderate CV values in addition to a few extreme outliers (Figure 4).
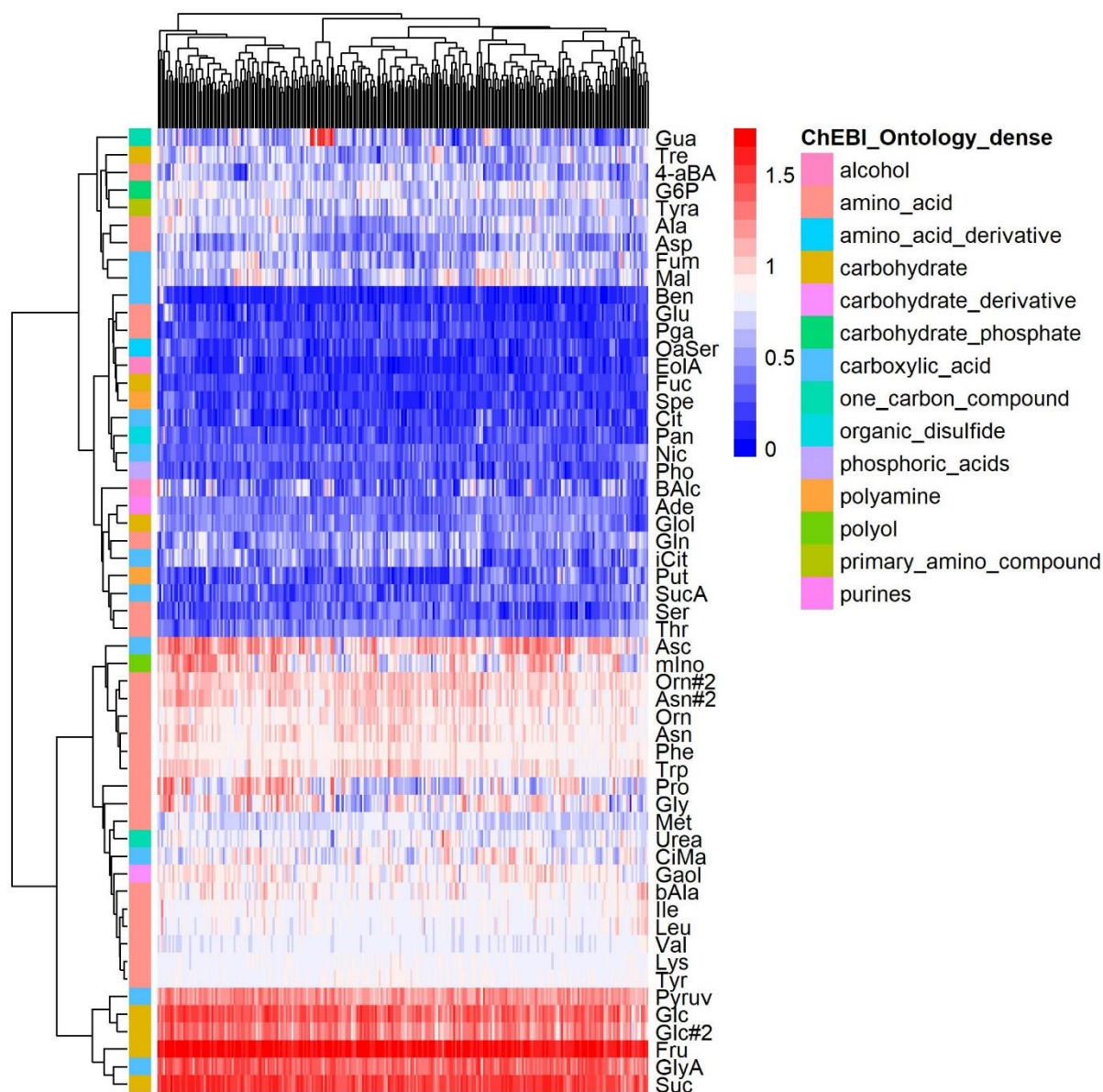
Figure 4: Heatmap of raw CV values of metabolites across 0, 3 and 6 days of darkness of all used ecotypes

To see if there is any connection between CV values of different metabolites, I correlated metabolite CV values to each other (Figure 5). As can be seen, many amino acids cluster together with moderate to high positive correlation values. The amino acids valine, leucine, isoleucine and β-alanine have moderate to high positive correlation values to each other but low positive to moderately negative correlation to most other amino acids (Figure 5). Also fructose, glucose and glucose-6-phosphate show moderate to high positive correlations to each other.
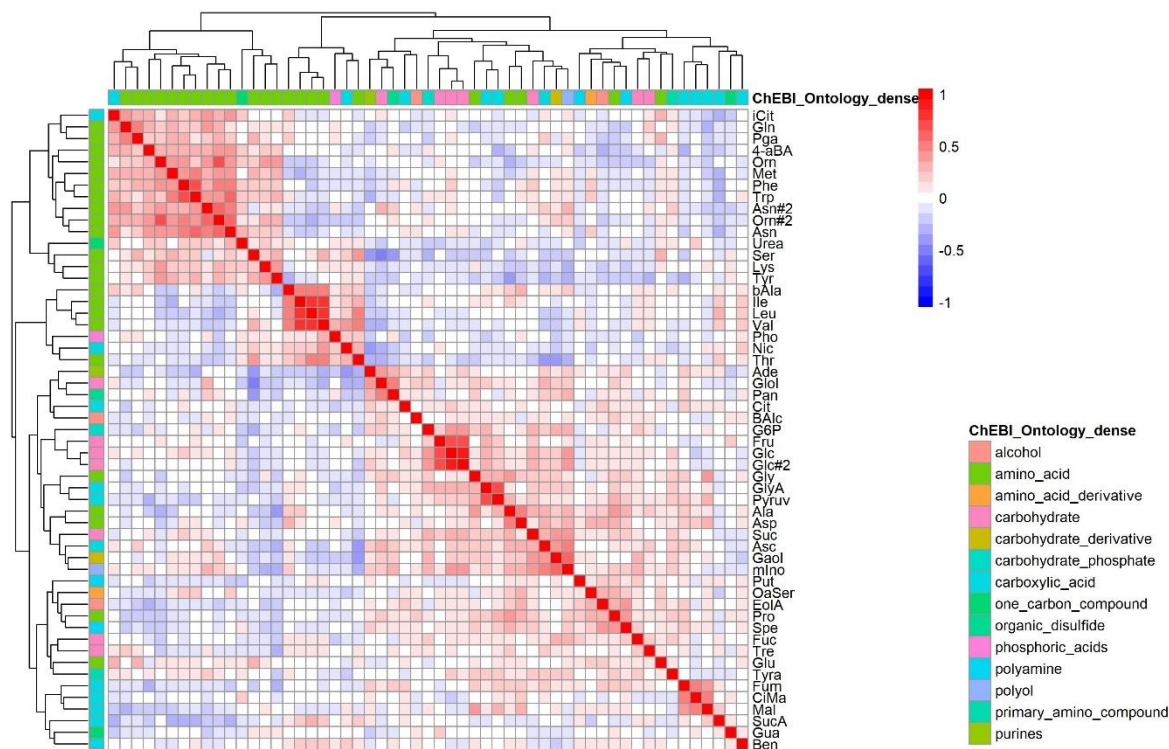
Figure 5: Heatmap of pearson-correlation coefficient of metabolites across 0, 3 and 6 days of darkness

After transformation, the ecotype-wise CV values were also used as input for the GWAS model. Although no association passed the strict bonferroni cut-off I could see some associations at the lower threshold.

For fucose I detected a significant association on chromosome 1 (Figure 6).

Figure 6: Manhattan plot of association of fucose CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The closest gene to the lead SNP is AT1G52330, which is a late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family. Closely next to this gene is ABA2 (AT1G52340), which is a gene related to aba production and glucose signalling (Rook et al., 2001).

I found an even stronger association between aspartic acid levels and a locus on chromosome 2 (Figure 7).
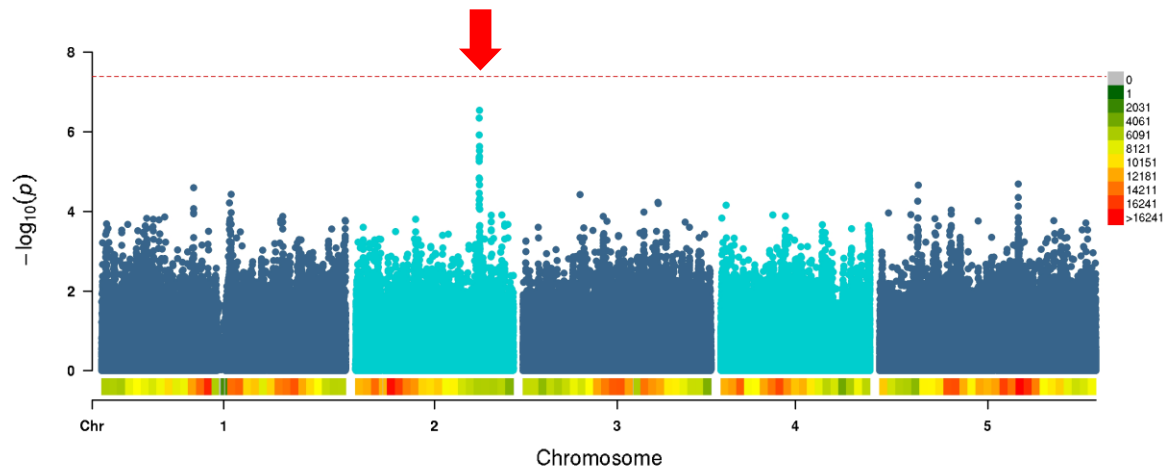
Figure 7: Manhattan plot of association of aspartic acid CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The closest gene is AT2G36860, which codes for a pre-tRNA. In the same locus, nearby is AT2G36870, which is a xyloglucan endotransglucosylase/hydrolase 32. In a locus on chromosome 3 with significant association to CV of glutamine levels (Figure 8), the gene closest to the leading SNP is a gene of unknown function AT3G27390. However, in the same locus I also find a gene from the pectin lyase-like superfamily (AT3G27370) and succinate dehydrogenase 2-1 (AT3G27380).

Figure 8: Manhattan plot of association of glutamine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

I also found a weak association between, lysine CV and another locus on chromosome 3 (Figure 9).

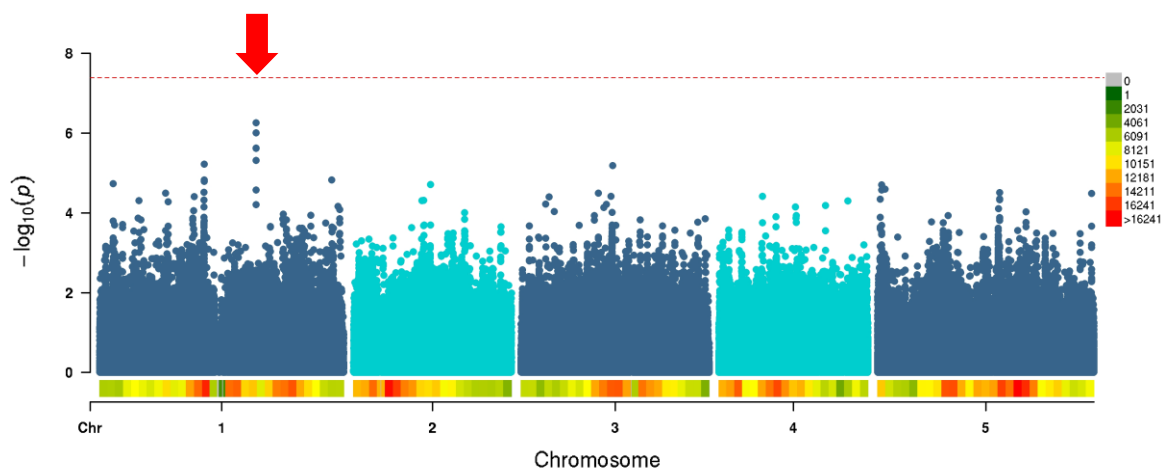*Figure 9: Manhattan plot of association of lysine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.*

The gene closest to the leading SNP is a lysophosphatidyl acyl transferase (AT3G57640). In vicinity are a gene coding for U2.2; snRNA (AT3G57650) and nuclear RNA polymerase (AT3G57660)

Finally, I also found a close association between malic acid and a region on chromosome 4 (Figure 10).

Figure 10: Manhattan plot of association of malic acid CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The leading SNP resides in AT4G13920 as well as AT4G13918, which are a receptor-like protein, and a potential natural antisense locus, respectively. The locus also contains a serine hydroxymethyltransferase4 (AT4G13930).

In total, for this dataset I found here 5 cmQTL for 5 different metabolites on 4 chromosomes.

2.3.2. *Arabidopsis thaliana* primary metabolites CV across darkness and light conditions

To explore an even more complex scenario, I included data from another experiment with normal and high light conditions and again calculated the ecotype-wise CV across all environments.

On average the CV values range from 0.27 for pyroglutamic acid to 1.5 for ornithine (Figure 11) while individual values range from close to 0 to over 2 (Figure 12).
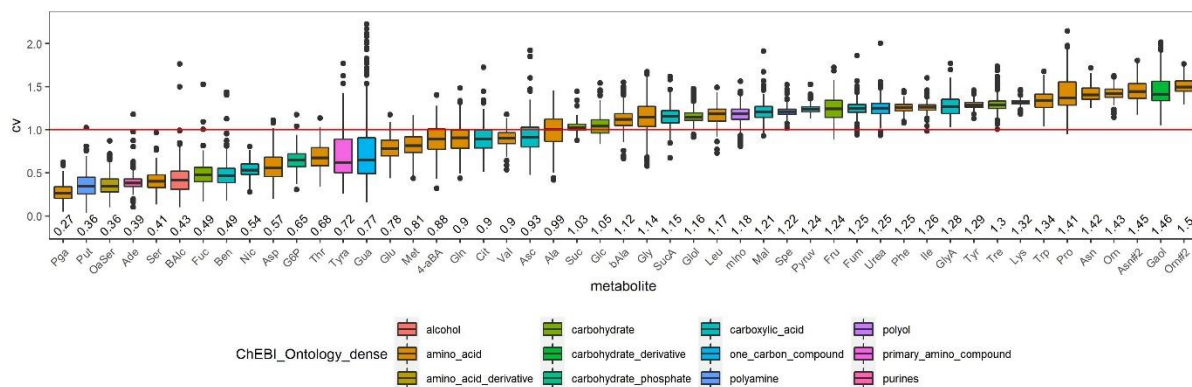
Figure 11: Boxplot of raw CV values of metabolites across 0, 3 and 6 days of darkness, normal light and high light conditions of all used ecotypes. The red horizontal line denotes a CV of 1. Numbers along the x-axis show mean CV.

Again I can see a metabolite-dependent CV value profile (Figure 12).
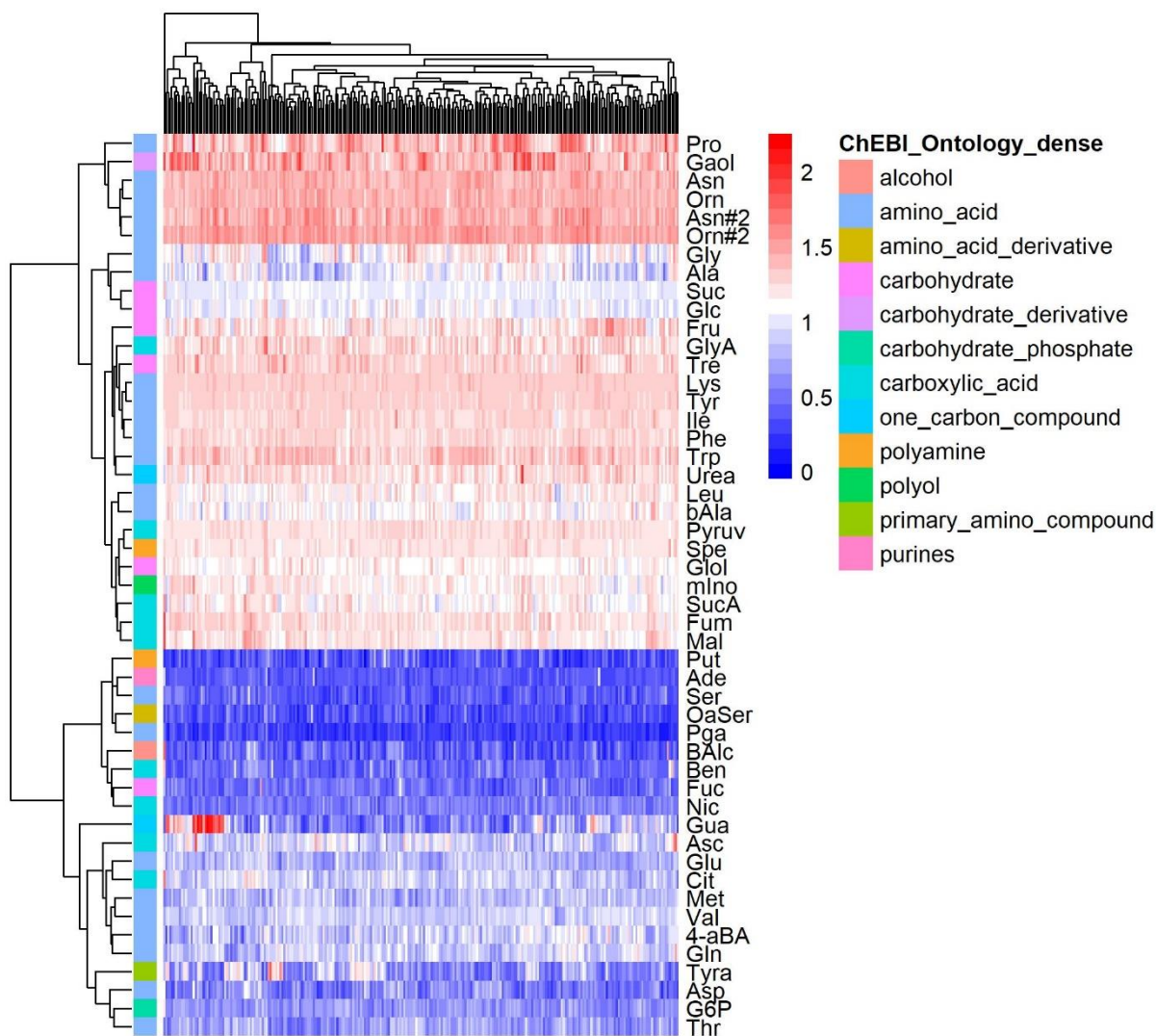
Figure 12: Heatmap of raw CV values of metabolites across 0, 3 and 6 days of darkness, normal light and high light conditions of all used ecotypes

With more than half of the compounds with a CV higher than 1, CV values are in this combination much higher as the ones in the dataset containing the 3 darkness treatments (Figure 3, Figure 11). Additionally, the cluster with the highest CV values here contains galactinol and the amino acids, asparagine, ornithine and proline (Figure 12). Glycine, alanine, sucrose and glucose form a small subcluster in a larger cluster with metabolites that all have a CV around 1 or higher. All other metabolites, which all have a CV lower than 1, form another large cluster together (Figure 12).

Also here I used the raw CV values to calculate Pearson's correlation coefficients between metabolites (Figure 13).
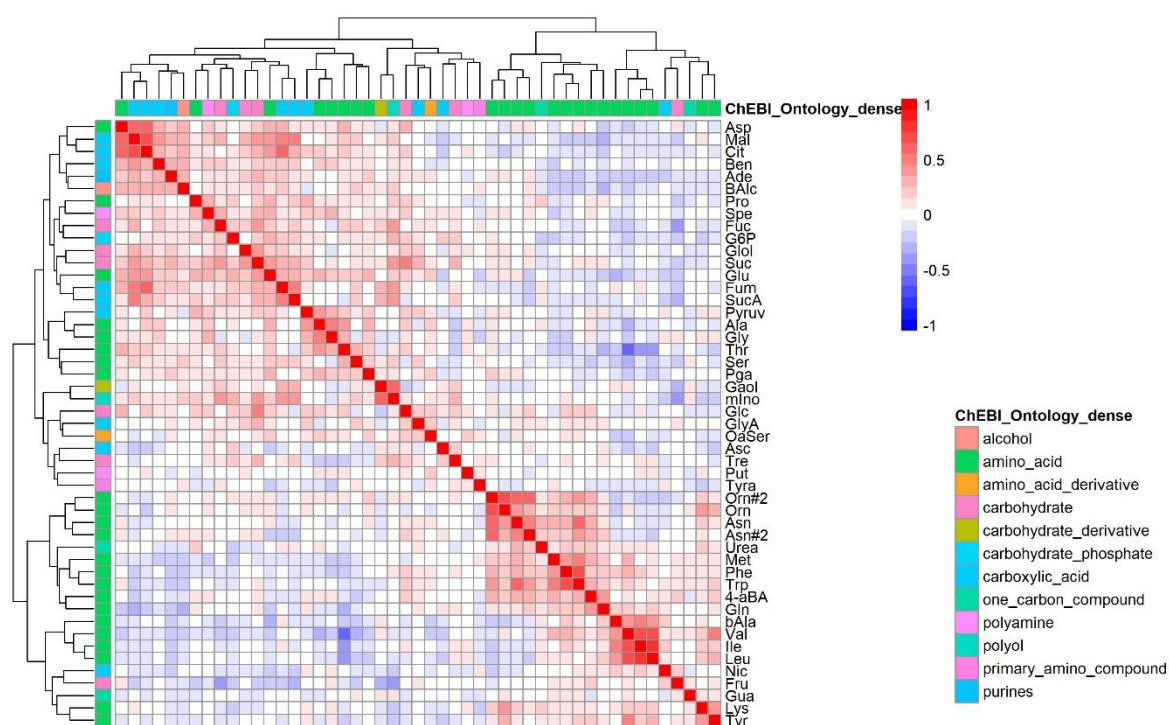
Figure 13: Heatmap of pearson-correlation coefficient of metabolites across 0, 3 and 6 days of darkness, normal light and high light conditions

Again I can see a high number of amino acids which show a moderate to high positive correlation to each other. The cluster between the amino acids valine, leucine, isoleucine and β-alanine, with moderate to high positive correlations to each, which I observed earlier in the dataset containing only the darkness treatments also persists here (Figure 5, Figure 13). Further on, aspartate, malate and citrate show a moderately high positive correlation pattern (Figure 13).

After transformation of the raw CV values, I used this dataset as input for the GWAS model. Again no associations breach the strict bonferroni cut-off, but I find associations at the marker-

defined threshold. The strongest association is one I found already in both individual conditions of the normal and high light experiment, which are not discussed in this thesis. This locus resides on chromosome 4 and associates to the CV of tyramine (Figure 14).
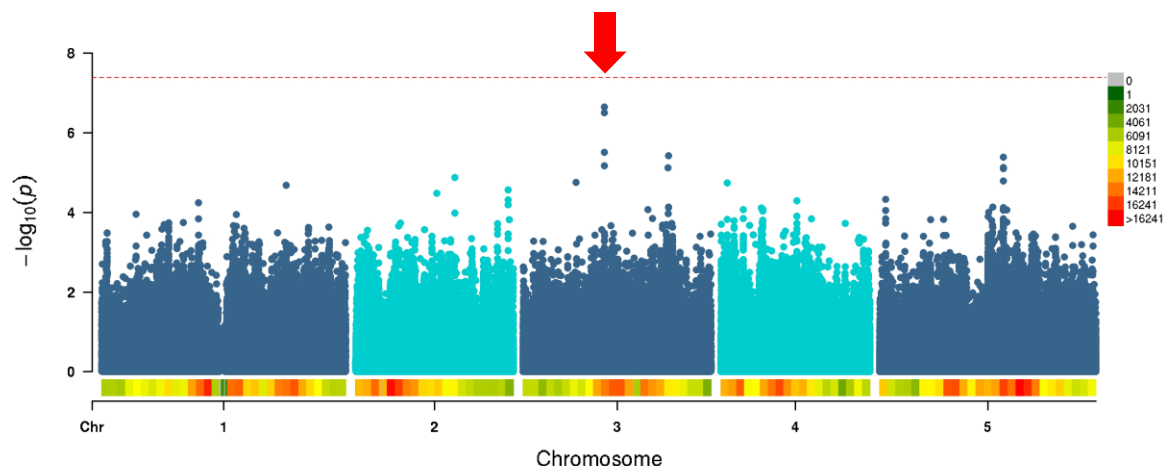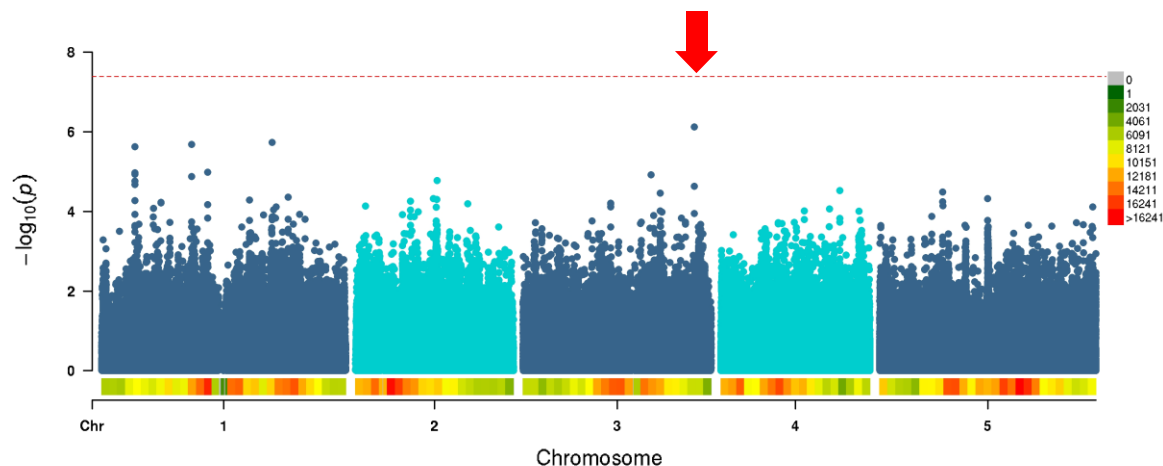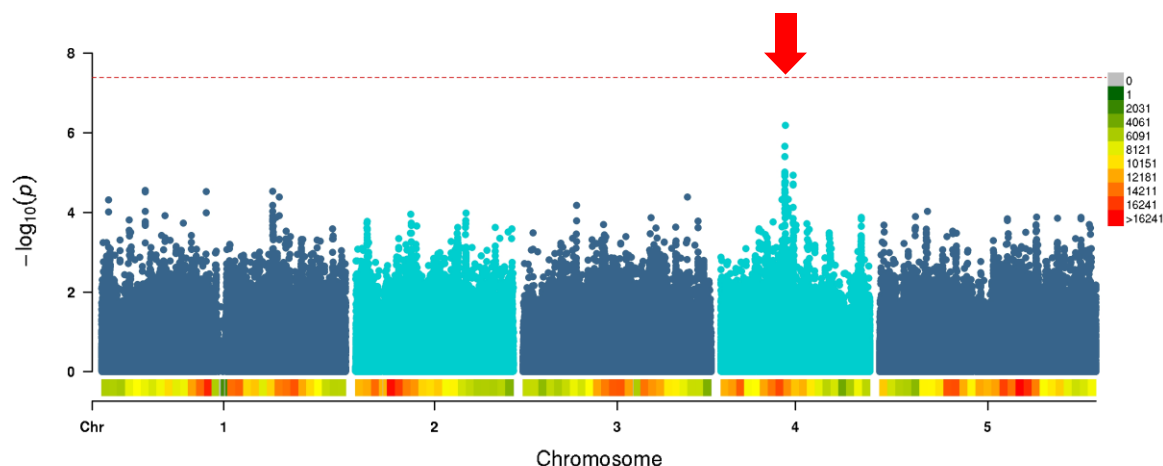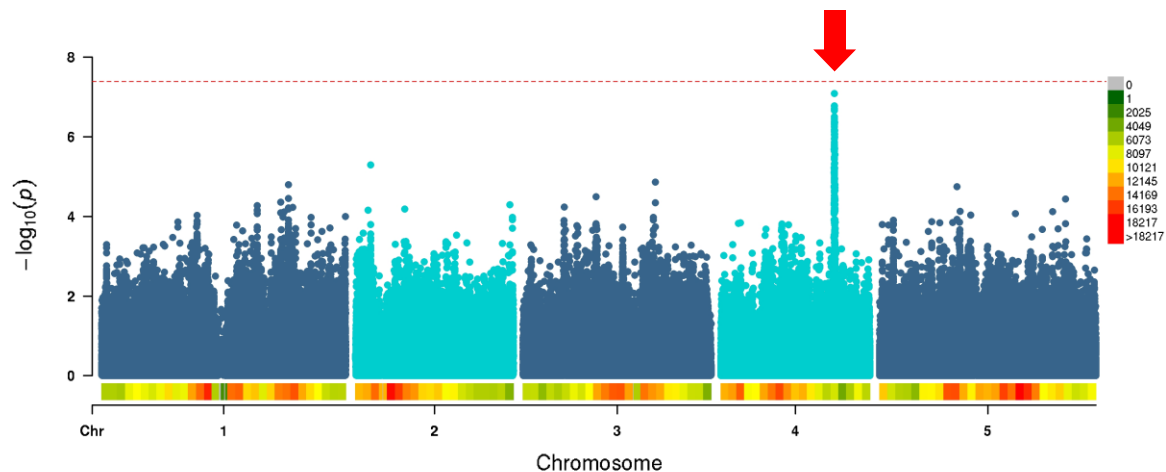

Figure 14: Manhattan plot of association of tyramine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

However in this case the gene closest to the leading SNP is not the known causal gene tyrosine decarboxylase (AT4G28680), but a gene coding for an RmlC-like cupins superfamily protein (AT4G28703). For the CV of methionine levels, I found an association to two genomic loci; one on chromosome 1 and one on chromosome 5 (Figure 15).
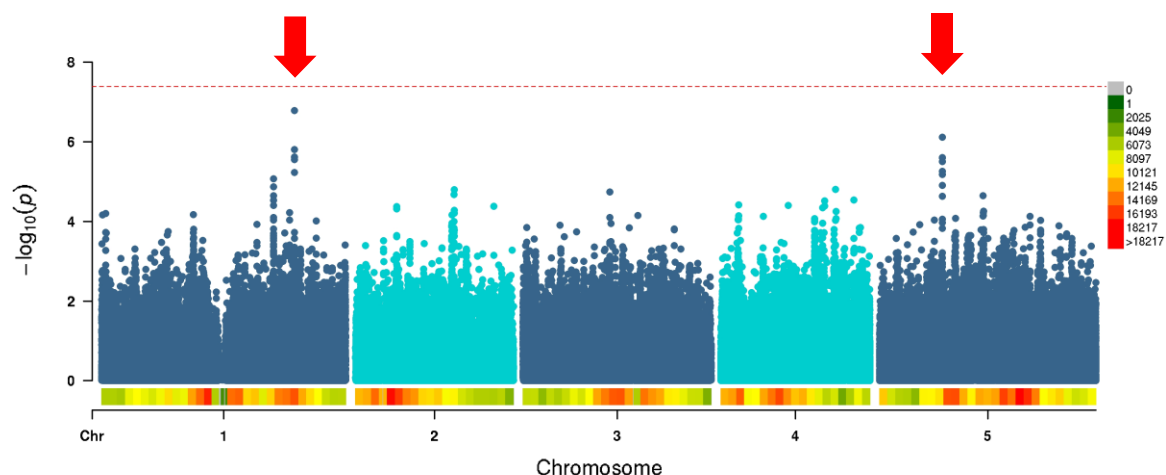
Figure 15: Manhattan plot of association of methionine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow point to the discussed loci.

The locus on chromosome 1 has a methionine gamma-lyase (AT1G64660) as the gene closest to the leading SNP and the locus on chromosome 5 has a Defensin-like (DEFL) family protein gene (AT5G23212) closest to the leading SNP. Other genes in the locus on chromosome 1 are a gene coding for a major facilitator superfamily protein (AT1G64650) and a gene coding for an alpha/beta-Hydrolases superfamily protein (AT1G64670). In the QTL on chromosome 5 I find a gene annotated as serine carboxypeptidase-like 34 (AT5G23210) and two nicotinamidases (AT5G23220 + AT5G23230).

For the CV of glycine I found an association to a SNP on chromosome 1 (Figure 16).
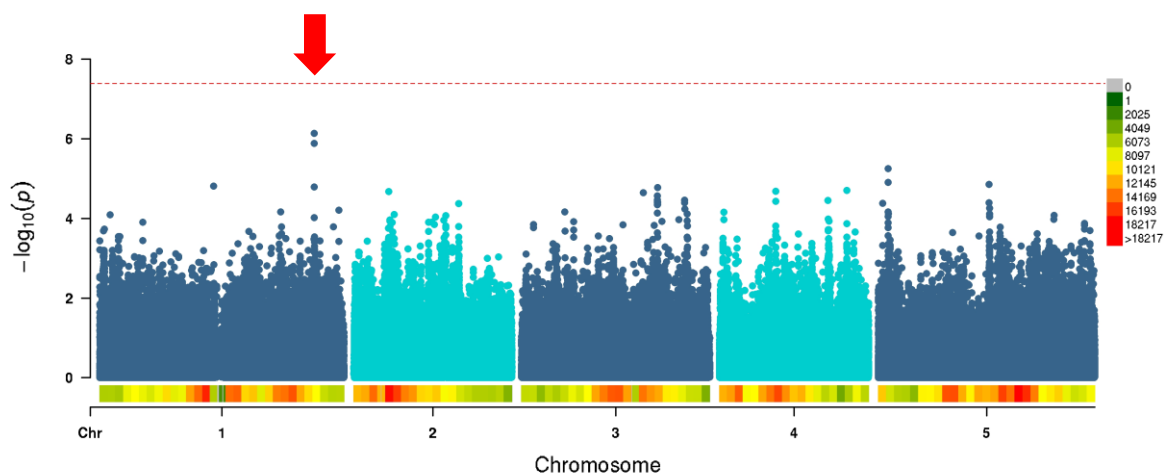
Figure 16: Manhattan plot of association of tyramine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

Closest to the SNP is a gene coding for a protein of unknown function (AT1G70750). Also close by is a the gene CRR23 (AT1G70760), which codes for a subunit of the chloroplast NAD(P)H dehydrogenase complex (Shimizu et al., 2008).

The CV of asparagine showed an association to a locus on chromosome 2 (Figure 17). The gene closest to the leading SNP is a transposable element gene. Interestingly, within around 20 kb up- and downstream of the leading SNP I find 7 more transposable element genes.

Figure 17: Manhattan plot of association of asparagine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

On a locus on chromosome 3, I found an association, to the CV of malic acid (Figure 18).

Figure 18: Manhattan plot of association of malic acid CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The closest gene is a Major facilitator superfamily gene and upstream, within ~13 kb of the leading SNP, there are 2 more genes of this superfamily. Downstream, close to the SNP I also find a Phosphoglycerate mutase family gene (AT3G05170).

Finally, I find an association between a locus on chromosome 5 and the CV of urea (Figure 19).
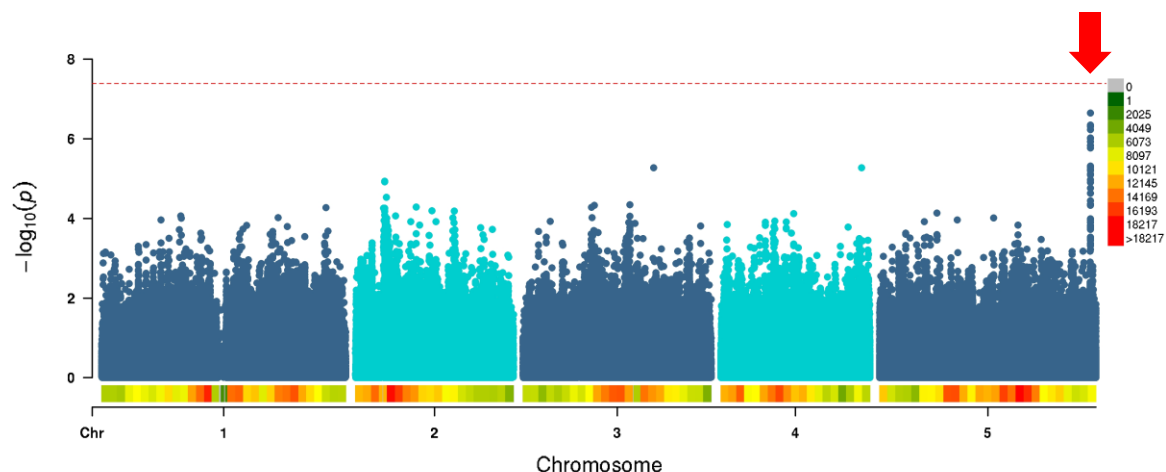
Figure 19: Manhattan plot of association of urea CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the $-log_{10}(p\text{-}value)$ of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

Also here I found a major facilitator superfamily gene (AT5G65687) as the gene closest to the leading SNP. The next closest gene is phosphoenolpyruvate carboxykinase 2 (AT5G65690).

In total in this dataset I found 7 cmQTLs of 6 metabolites, across all chromosomes.

2.3.3. *Arabidopsis thaliana* primary metabolites CV across light and control conditions

Since the starvation condition under the dark treatment, may have influenced our results I tried to combine the two datasets from the normal and high light condition, with the control set of the darkness experiment, which had not been subjected to dark.

In this combination the average CV values range from 0.18 for adenine to 1.57 for ornithine (Figure 20). Here only 2 metabolites have a CV higher than 1, asparagine and ornithine (I picked two peaks for ornithine, which both have a high CV). The two ornithine peaks build a distinct small cluster with very high CV values (Figure 21). A large cluster contains many metabolites with mostly very low CVs. Another large cluster contains metabolites with CVs, which show greater variation between ecotypes. Within this cluster tyramine shows a very high CV for some ecotypes and on the other hand very low CV for most other ecotypes (Figure 21).
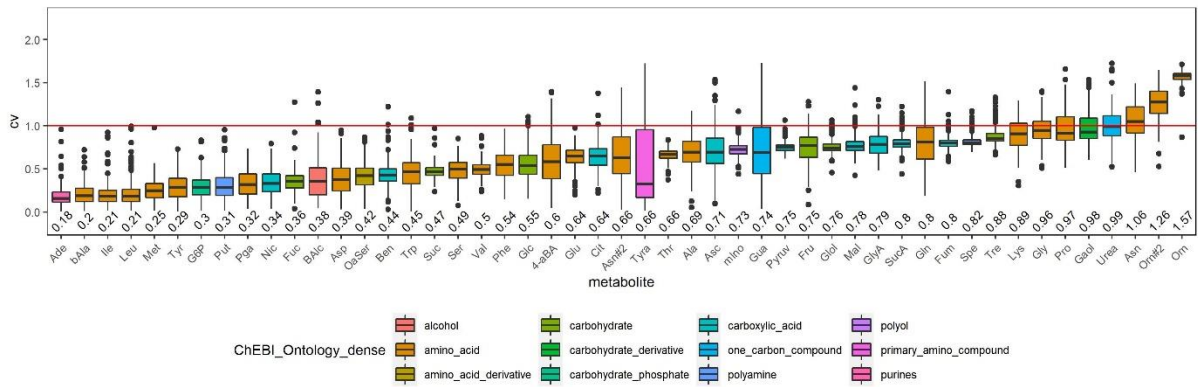
Figure 20: Boxplot of raw CV values of metabolites across normal light and high light and control conditions of the darkness experiment of all used ecotypes. The red horizontal line denotes a CV of 1. Numbers along the x-axis show mean CV.
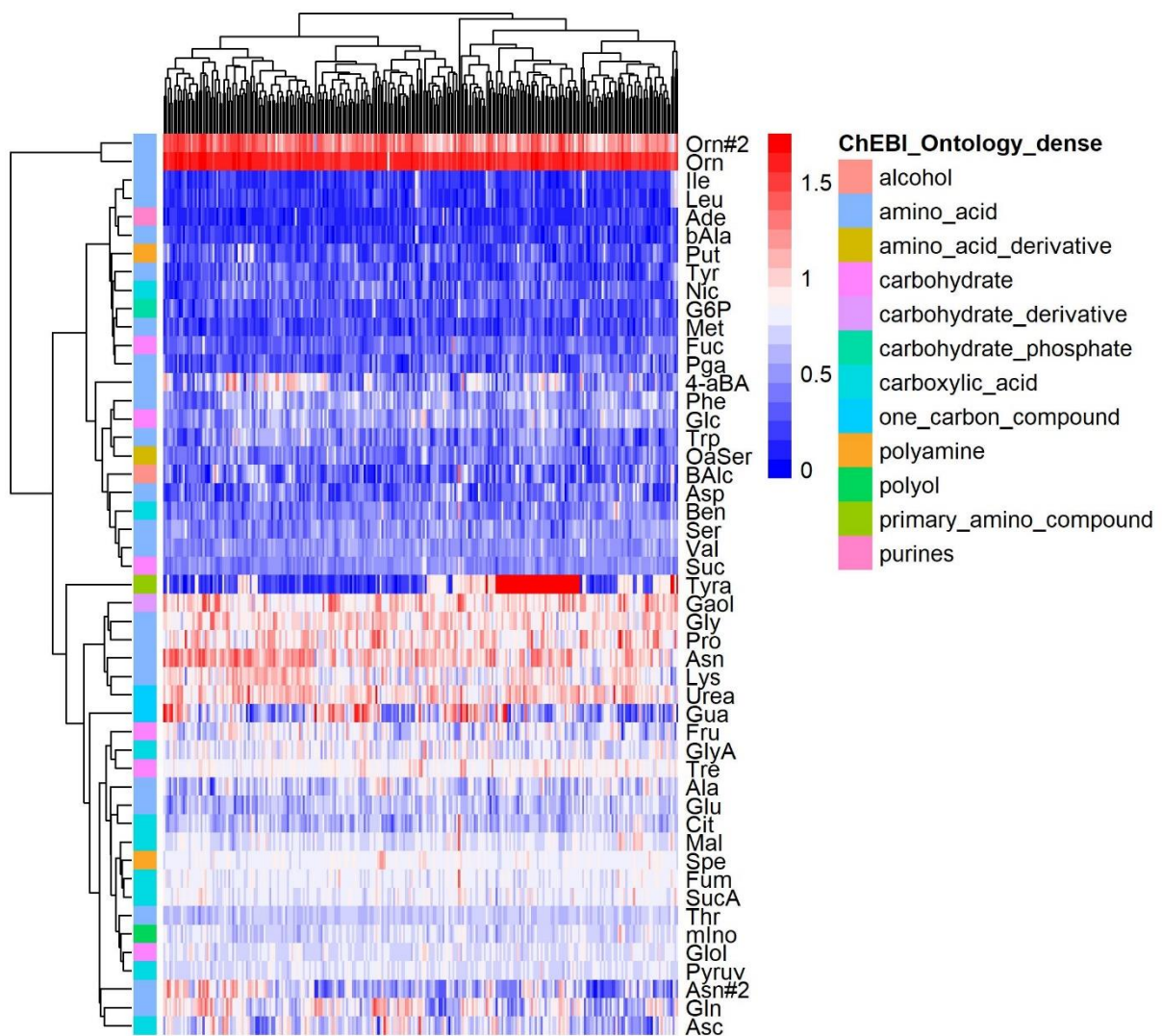
Figure 21: Heatmap of raw CV values of metabolites across normal light and high light and control conditions of the darkness experiment of all used ecotypes

When correlating CVs between metabolites, I can see a similar pattern as before (Figure 22).
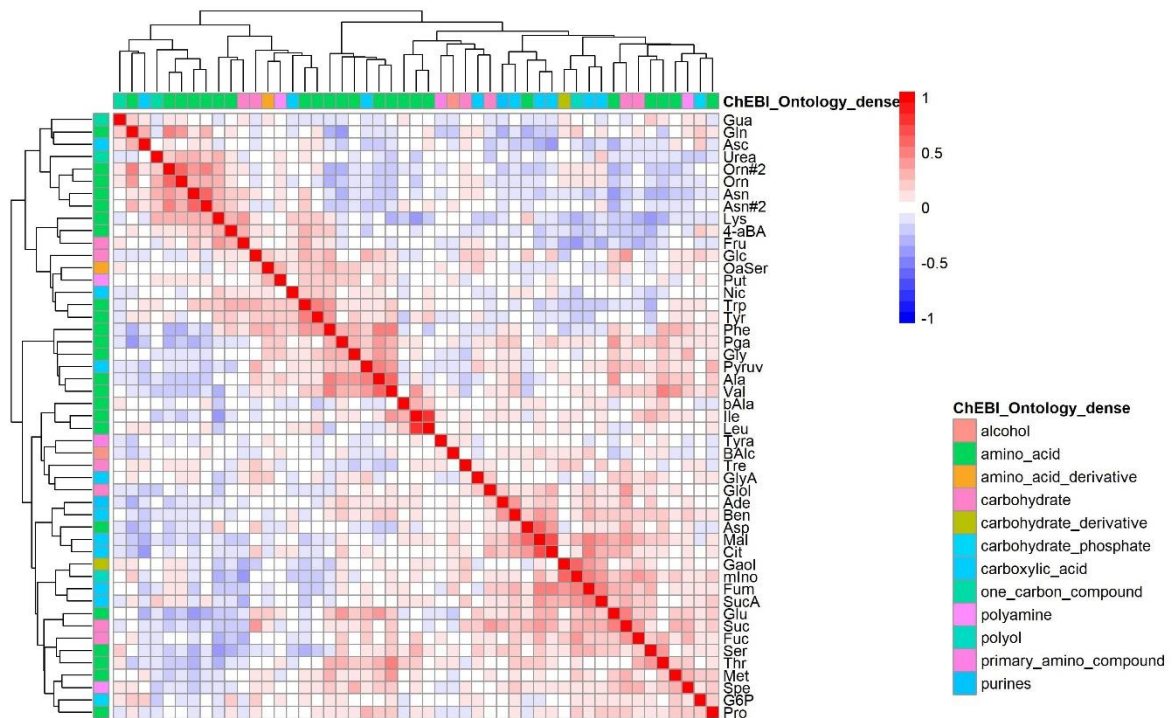
Figure 22: Heatmap of pearson-correlation coefficient of metabolites across normal light and high light and control conditions of the darkness experiment

Many amino acids show a positive correlation although only some with higher and most with relatively low coefficients. Leucine and isoleucine show a strong positive correlation (0.77). Also aspartate, malate and citrate show moderately strong positive correlations (Figure 22).

When using the transformed CV values as input for the GWAS model, I again detected the association between the tyrosine decarboxylase locus on chromosome 4 and the CV of tyramine, this time even surpassing the bonferrroni cut-off.

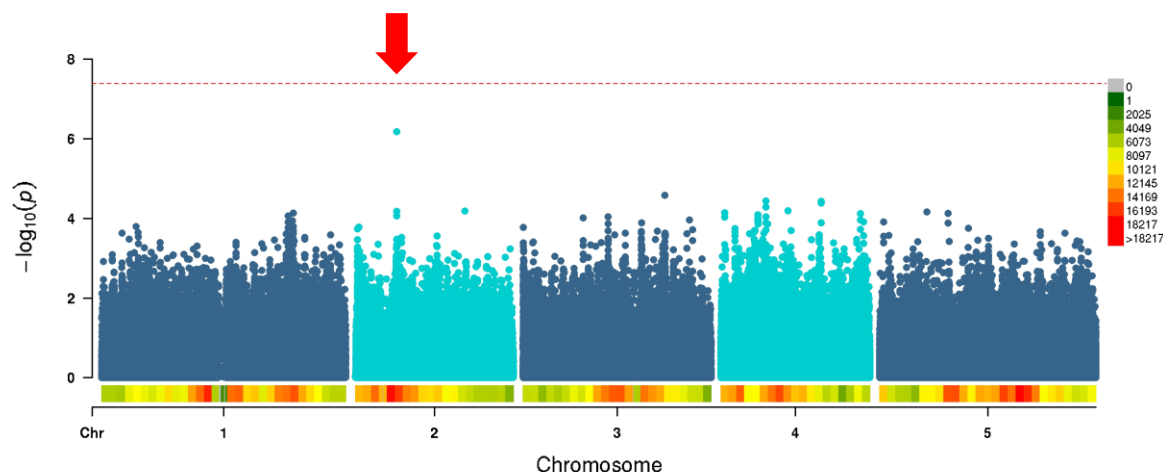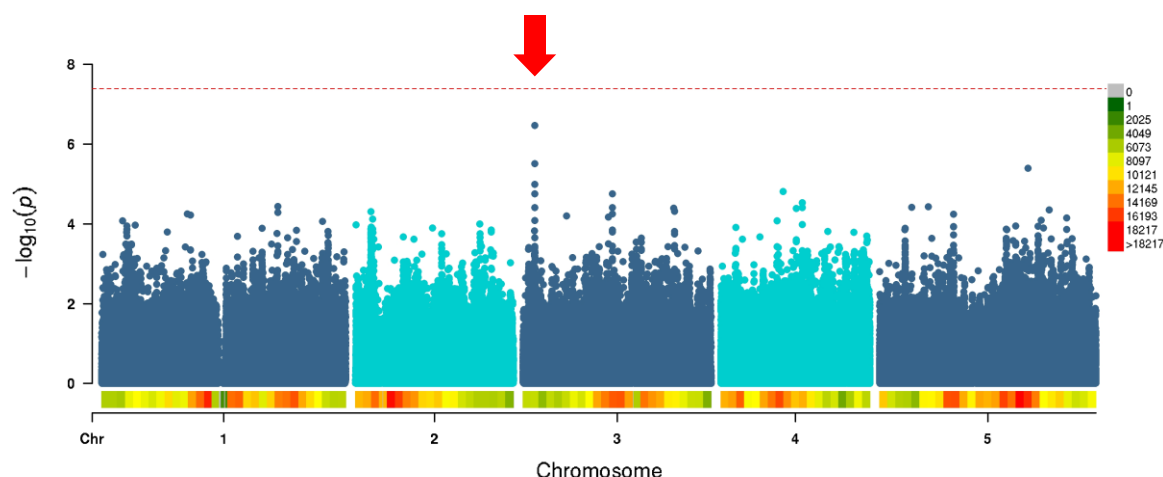For the CV of leucine I found a cmQTL on chromosome 2 (Figure 23).
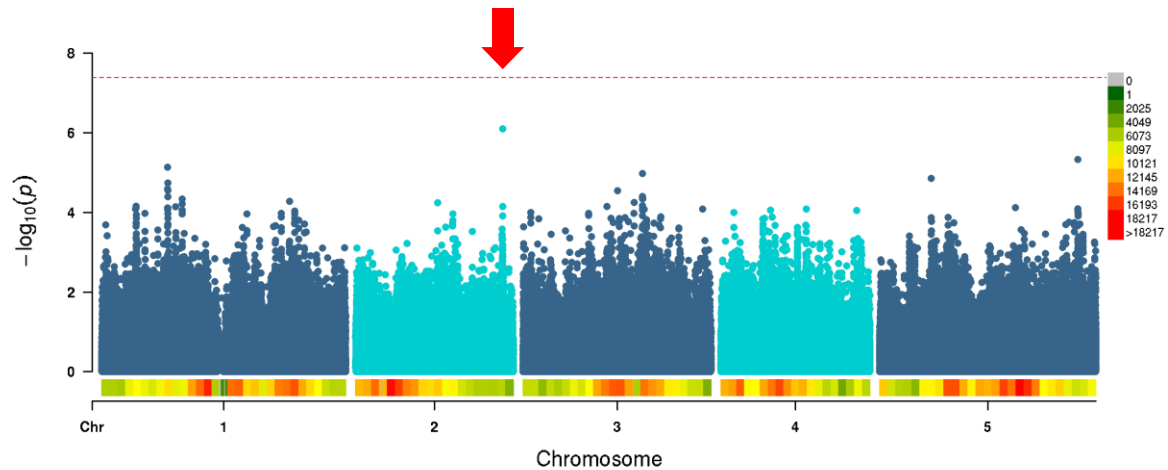
Figure 23: Manhattan plot of association of leucine CV to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the $-log_{10}(p\text{-}value)$ of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The closest gene to the lead SNP is beta glucosidase 17 (AT2G44480). Upstream are 3 more beta glucosidases and downstream a glycosyl hydrolase (AT2G44490).

Additionally I found cmQTLs for the CV of isoleucine, benzyl alcohol and trehalose as summarized in Table 2.

Table 2: Selection of putative candidate genes controlling the CV of primary metabolites across different light conditions

#	Distance to best SNP	LOD	Gene ID	Gene annotation	Metabolite
1	-5050	6.25	AT2G01630	O-Glycosyl hydrolases family 17 protein	tyrosine
	-680	6.25	AT2G01650	plant UBX domain-containing protein 2	
	3430	6.23	AT2G01660	plasmodesmata-located protein 6	
2	-17021	6.10	AT2G44450	beta glucosidase 15	leucine
	-10904	6.10	AT2G44460	beta glucosidase 28	
	-2191	6.10	AT2G44470	beta glucosidase 29	
	0	6.10	AT2G44480	beta glucosidase 17	
	3886	6.10	AT2G44490	Glycosyl hydrolase superfamily protein	
	13422	6.10	AT2G44500	O-fucosyltransferase family protein	
3	-262	6.36	AT3G01140	myb domain protein 106	isoleucine
	-262	6.36	AT3G01142	Potential natural antisense gene	
4	-655	6.52	AT3G45210	Protein of unknown function, DUF584	benzyl alcohol
5	0	9.65	AT4G28680	L-tyrosine decarboxylase	tyramine
6	-20208	6.38	AT5G33424	transposable element gene	trehalose
	-17602	6.38	AT5G33427	transposable element gene	
	-6897	6.38	AT5G33428	transposable element gene	
	8381	6.76	AT5G33431	transposable element gene	
	-4849	6.38	AT5G33432	transposable element gene	
	4831	6.76	AT5G33433	transposable element gene	
	11268	6.76	AT5G33434	transposable element gene	
	0	6.76	AT5G33436	pseudogene	
	18037	6.76	AT5G33438	transposable element gene	
	4415	6.76	AT5G33439	other RNA	
	2596	6.76	AT5G33441	pseudogene	
	21158	6.76	AT5G33442	transposable element gene	

In these loci I find a myb transcription protein, along with a potential natural antisense gene, a gene of unknown function, pseudogenes and several transposable element genes. In total in this dataset I was able to find 6 cmQTL from 6 metabolites across 4 chromosomes.

2.3.4. *Arabidopsis thaliana* secondary metabolites fold change between low and high light conditions

For the Arabidopsis normal and high light experiment, I also looked at secondary metabolites. Since only data from two conditions were available, I used the metabolite fold-change as input for the GWAS.

I found a significant association of one unannotated metabolite to a locus on chromosome 2 (Figure 24).
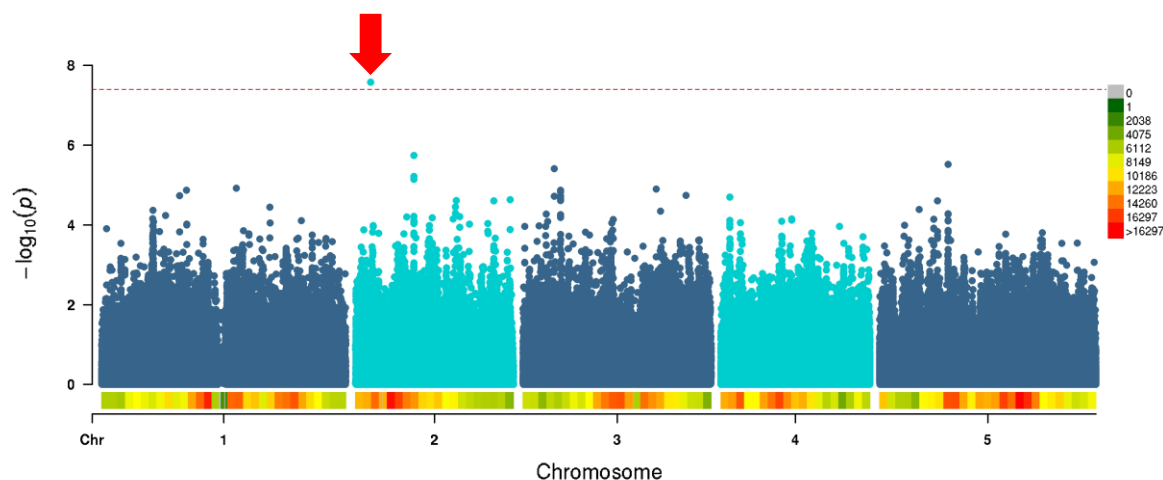
Figure 24: Manhattan plot of association of the fold-change of an unknown compound to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The lead SNP is close to a gene coding for a DNAJ heat shock N-terminal domain-containing protein (AT2G05230) and another gene with the same annotation is around 14 kb apart (AT2G05250).

A locus on chromosome 3 showed a significant association to 4 glucosinolates (4-pentenylglucosinolate, 3-butenylglucosinolate, 6-methylsulfinylhexyl glucosinolate, 1-methoxy-3-indolylmethyl-glucosinolate) as well as one unannotated compound (Figure 25). In the case of 4-pentenylglucosinolate the LOD score is even above the bonferroni threshold.
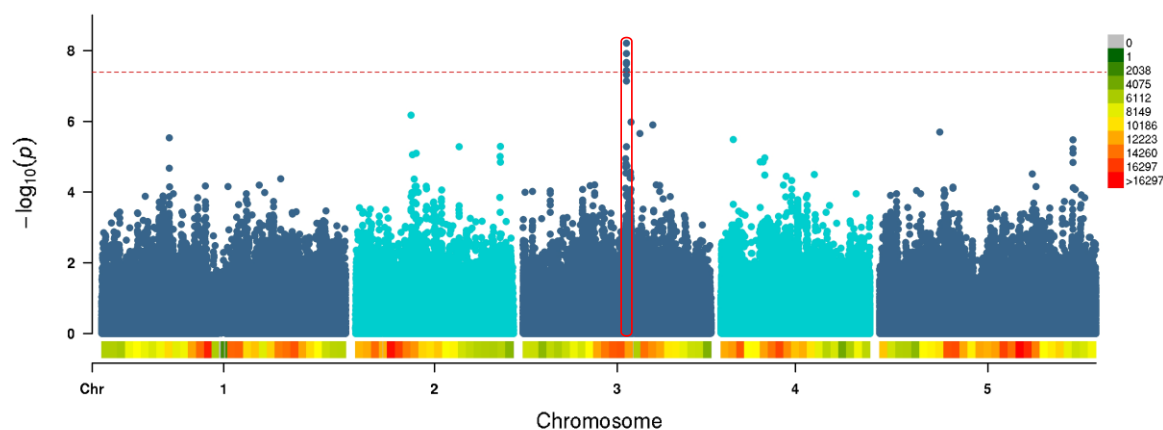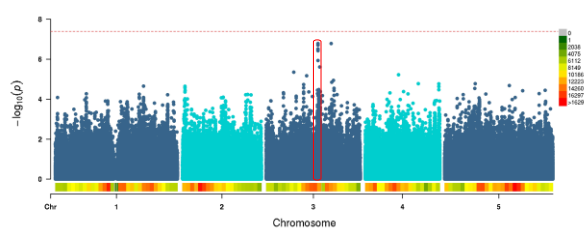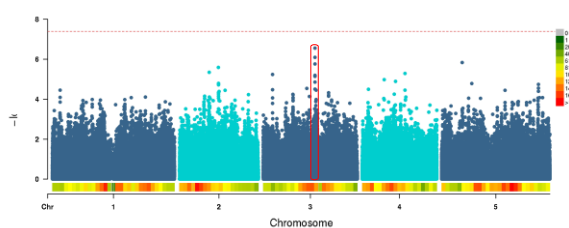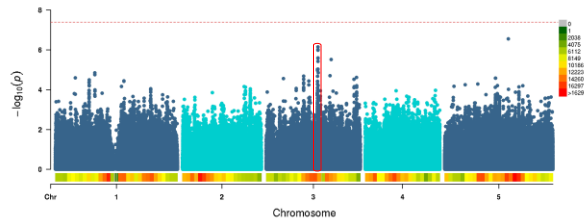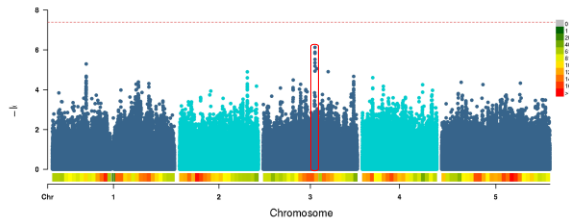
Figure 25: Manhattan plot of association of the CV of 4 glucosinolates (A: 4-pentenylglucosinolate; B: 3-butenylglucosinolate; C: 6-methylsulfinylhexyl glucosinolate; D: 1-methoxy-3-indolylmethyl-glucosinolate; E: unknown compound) to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the – $\log_{10}$(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red rounded rectangle encloses the discussed locus.

The closest gene to the lead SNP is a transposable element gene (AT3G31904), which may be a pseudogene for a helicase. In that region (~10 kb up- and downstream) I find several other transposable element genes and pseudogenes. The nearest protein coding gene, with a predicted function is an ATP-dependent helicase family gene (AT3G31900).

I found another locus on chromosome 3 showing an association to a glucosinolate (4-methylsulfinylbutyl glucosinolate) (Figure 26).
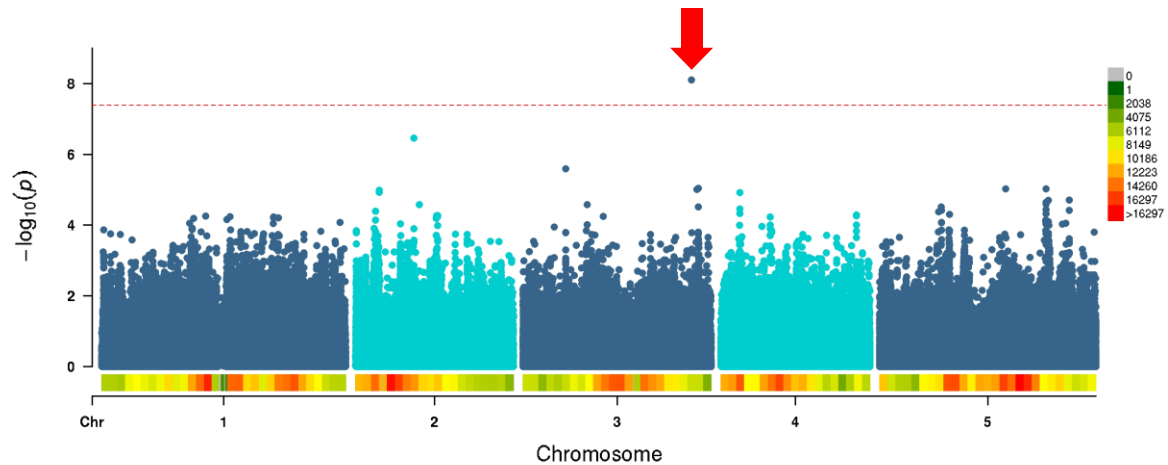
Figure 26: Manhattan plot of association of the CV of 4-methylsulfinylbutyl glucosinolate. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the $-log_{10}$(p-value) of a calculated association. The red line shows the bonferroni threshold. Colors in the bottom panel show the marker density binned into 1 Mb windows, according to the color legend at the top right. The red arrow points to the discussed locus.

The lead SNP was closest to a gene coding for Cam interacting protein 111 (AT3G56690). Also very close was the gene for fatty acid reductase 6 (AT3G56700).

All other associations do not surpass the bonferroni threshold, but are significant at a marker-defined threshold. A list of selected putative candidate genes is shown in Table 3.

Chapter 2: cmGWAS

Table 3: Selection of putative candidate genes controlling the fold change of secondary metabolites across normal light and high light conditions

#	Distance (best SNP)	LOD (nearest SNP)	Gene ID	gene_syn	Putative Metabolites
1	0	6.58	AT1G18500	methylthioalkylmalate synthase-like 4	unannotated#74
2	-1196	6.37	AT1G30260	NA	glucoiberin
	1933	6.37	AT1G30270	CBL-interacting protein kinase 23	
3	-620	6.64	AT1G34490	MBOAT (membrane bound O-acyl transferase) family protein	guanosine
	523	6.64	AT1G34500	MBOAT (membrane bound O-acyl transferase) family protein	
4	-6568	6.30	AT1G64760	O-Glycosyl hydrolases family 17 protein	quercitrin
	-352	6.30	AT1G64780	ammonium transporter 1;2	
5	0	6.57	AT1G65440	global transcription factor group B1	7,8-Dihydroxycoumarin
6	2423	6.15	AT2G01190	Octicosapeptide/Phox/Bem1p family protein	Indole-3-acetyl-L-aspartic acid
	5465	6.15	AT2G01200	indole-3-acetic acid inducible 32	
7	0	6.15	AT2G03340	WRKY DNA-binding protein 3	2,3-DHBAG peak2
	6465	6.15	AT2G03360	Glycosyltransferase family 61 protein	
8	297	7.58	AT2G05230	DNAJ heat shock N-terminal domain-containing protein	2,5-DHBAX peak1
	14128	7.58	AT2G05250	DNAJ heat shock N-terminal domain-containing protein	
9	123	6.27	AT2G07788	transposable element gene	tryptophan
	4977	6.27	AT2G07789	transposable element gene	
10	-564	6.66	AT2G13280	transposable element gene	unannotated#52
	0	6.66	AT2G13290	beta-1,4-N-acetylglucosaminyltransferase family protein	
11	0	6.18	AT2G15910	CSL zinc finger domain-containing protein	4-pentenylglucosinolate
	1294	6.18	AT2G15920	transposable element gene	
12	-808	6.46	AT2G16790	P-loop containing nucleoside triphosphate hydrolases superfamily protein	glucoraphanin
	0	6.46	AT2G16800	high-affinity nickel-transport family protein	
13	-291	6.22	AT2G22830	squalene epoxidase 2	unannotated#8
14	-1091	6.82	AT3G24180	Beta-glucosidase, GBA2 type family protein	unannotated#25
	437	6.82	AT3G24190	Protein kinase superfamily protein	
15	0	6.40	AT3G25805		glucoalyssin
16	-13023	8.20	AT3G31900	ATP-dependent helicase family protein	4-pentenylglucosinolate, gluconapin, 6-methylsulfinylhexyl glucosinolate, 6-methylthiohexylglucosinolate, unannotated#60
	0	8.20	AT3G31904	transposable element gene	
17	151	6.79	AT3G44610	Protein kinase superfamily protein	unannotated#60
18	-247	8.10	AT3G56690	Cam interacting protein 111	glucoraphanin
	789	8.10	AT3G56700	fatty acid reductase 6	
19	208	6.47	AT3G57070	Glutaredoxin family protein	unannotated#92
20	-3414	6.81	AT3G59850	Pectin lyase-like superfamily protein	2,5-DHBAG peak 1+2
	-115	6.81	AT3G59860	transposable element gene	
21	-3005	6.48	AT4G06550	transposable element gene	

	1565	6.48	AT4G06551	transposable element gene	Kaempferol 3-O-glucosyl-glucoside 7-O-rhamnoside
22	0	6.38	AT4G22770	AT hook motif DNA-binding family protein	2,5-DHBAG peak 1+2
23	315	6.62	AT5G39580	Peroxidase superfamily protein	6-methylsulfinylhexyl glucosinolate, unannotated#45
24	-483	6.15	AT5G43020	Leucine-rich repeat protein kinase family protein	unannotated#70
25	-2806	6.11	AT5G50915	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	unannotated#70
26	2326	6.59	AT5G54520	Transducin/WD40 repeat-like superfamily protein	unannotated#10
	8537	6.59	AT5G54530	Protein of unknown function, DUF538	

2.3.5. *Phaseolus vulgaris* secondary metabolites CV across short and long day and drought conditions

Switching from Arabidopsis to common bean, I also investigated secondary metabolism. Data from 3 previous experiments was available, where beans were grown under short day, long day and drought conditions. As a first step, I matched the metabolic features, by their closest match in retention time and m/z-value to a list of manually checked chromatogram peaks and filtered out all features, where more than 50% of values were missing. The remaining 169 features are listed in Table 4.

For these features, I calculated the CV across conditions. The average CV values range from 0.34 to 1.73, with most features having a mean CV higher than 1 (Figure 27).

Chapter 2: cmGWAS

Table 4: Matched metabolic features detected in polar LC-MS of samples of common bean from short day, long day and drought conditions; RT: retention time; m/z: mass-charge ratio

Feature	RT	m/z	Feature	RT	m/z	Feature	RT	m/z	Feature	RT	m/z
m_2	2.48	295.1	m_51	5.89	295.05	m_96	7.85	491.08	m_143	11.51	591.34
m_3	2.48	337.08	m_52	6	295.05	m_97	7.92	565.23	m_144	11.51	839.37
m_4	2.48	369.1	m_53	6.08	433.21	m_98	7.92	609.22	m_145	11.71	1097.52
m_5	2.92	309.12	m_54	6.08	625.14	m_100	8.04	533.09	m_146	11.71	1143.52
m_6	2.92	381.13	m_55	6.08	755.2	m_101	8.04	839.34	m_147	11.71	837.36
m_7	2.99	309.12	m_56	6.15	741.19	m_102	8.13	677.28	m_148	11.76	1229.6
m_9	3.2	331.07	m_57	6.17	609.15	m_104	8.2	429.18	m_149	11.95	635.3
m_10	3.28	371.06	m_58	6.23	755.21	m_105	8.2	497.16	m_150	11.95	1213.6
m_11	3.38	371.06	m_59	6.28	433.21	m_106	8.33	561.2	m_151	11.99	1081.52
m_12	3.57	315.07	m_60	6.34	431.19	m_107	8.33	605.19	m_152	11.99	1127.53
m_13	3.62	371.06	m_61	6.39	613.18	m_108	8.36	457.21	m_153	12.09	627.3
m_14	3.69	305.09	m_62	6.43	595.13	m_109	8.36	411.2	m_154	12.09	1081.52
m_15	3.74	329.09	m_63	6.41	535.19	m_110	8.44	463.26	m_156	12.18	441.25
m_16	3.87	371.06	m_64	6.58	427.2	m_111	8.44	679.28	m_157	12.32	953.48
m_17	3.96	311.04	m_65	6.58	523.22	m_112	8.52	677.28	m_158	12.32	999.48
m_18	4.06	311.04	m_66	6.6	547.16	m_113	8.62	467.21	m_160	12.37	343.26
m_20	4.15	311.04	m_67	6.64	725.2	m_114	8.69	579.21	m_163	12.44	987.52
m_22	4.37	323.13	m_68	6.75	279.05	m_115	8.71	675.27	m_164	12.91	1067.54
m_24	4.25	371.06	m_69	6.75	581.09	m_116	8.96	675.27	m_165	12.91	1113.55
m_25	4.25	203.08	m_70	6.75	889.2	m_117	8.75	579.21	m_167	13.06	1097.55
m_27	4.51	355.07	m_71	6.81	609.15	m_118	8.97	371.13	m_168	13.06	1119.54
m_28	4.63	385.08	m_72	6.89	431.1	m_119	8.83	467.21	m_171	13.34	789.35
m_29	4.63	447.19	m_73	6.93	279.05	m_120	9.19	771.4	m_172	13.39	328.21
m_30	4.68	599.16	m_74	6.93	559.11	m_121	9.39	577.21	m_173	13.43	562.32
m_31	4.68	299.08	m_75	6.93	581.09	m_122	9.39	492.23	m_174	13.47	721.37
m_32	4.68	621.14	m_76	6.93	859.19	m_123	9.39	781.29	m_175	13.54	562.32
m_33	4.83	295.05	m_77	6.99	477.07	m_124	9.44	805.33	m_176	13.58	593.27
m_34	4.96	385.08	m_78	6.99	955.14	m_125	9.51	569.35	m_177	13.58	319.14
m_35	5.05	325.09	m_79	6.99	1433.22	m_126	9.72	629.28	m_178	13.91	559.31
m_36	5.05	651.19	m_80	7.14	193.05	m_127	9.72	652.28	m_179	14.12	416.37
m_37	5.12	433.21	m_81	7.14	309.06	m_128	9.88	1259.57	m_180	14.17	540.33
m_38	5.12	771.2	m_82	7.14	619.13	m_130	10.2	377.18	m_181	14.27	221.15
m_39	5.32	325.06	m_83	7.17	593.15	m_131	10.23	855.37	m_182	14.42	505.26
m_40	5.36	371.23	m_84	7.28	309.06	m_132	10.6	753.39	m_183	14.42	265.15
m_41	5.45	355.1	m_85	7.38	593.15	m_133	10.67	855.37	m_184	14.81	311.2
m_42	5.57	399.09	m_86	7.38	661.14	m_134	10.67	753.39	m_185	15.28	356.19
m_43	5.57	447.15	m_87	7.38	839.34	m_135	10.8	403.2			
m_44	5.65	387.17	m_88	7.42	429.18	m_136	10.8	855.37			
m_45	5.65	583.2	m_89	7.52	623.16	m_137	11.05	643.28			
m_46	5.73	433.21	m_90	7.56	429.18	m_138	11.08	837.36			
m_47	5.73	493.06	m_91	7.62	461.07	m_139	11.11	242.18			
m_48	5.73	583.2	m_92	7.62	692.11	m_140	11.37	643.28			
m_49	5.8	431.19	m_94	7.62	1154.69	m_141	11.37	593.35			
m_50	5.89	591.1	m_95	7.82	429.18	m_142	11.37	839.37			
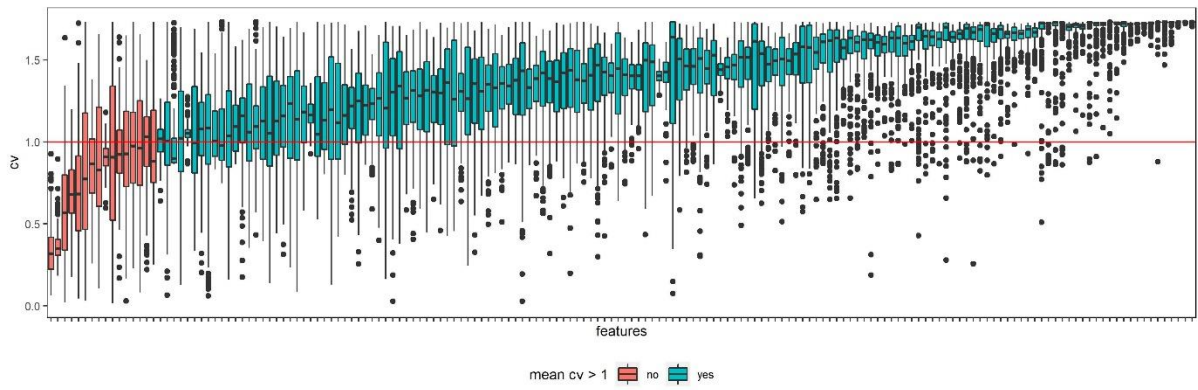
Figure 27: Boxplot of raw CV values of features across short day long day and drought conditions of all used ecotypes. The red horizontal line denotes a CV of 1.
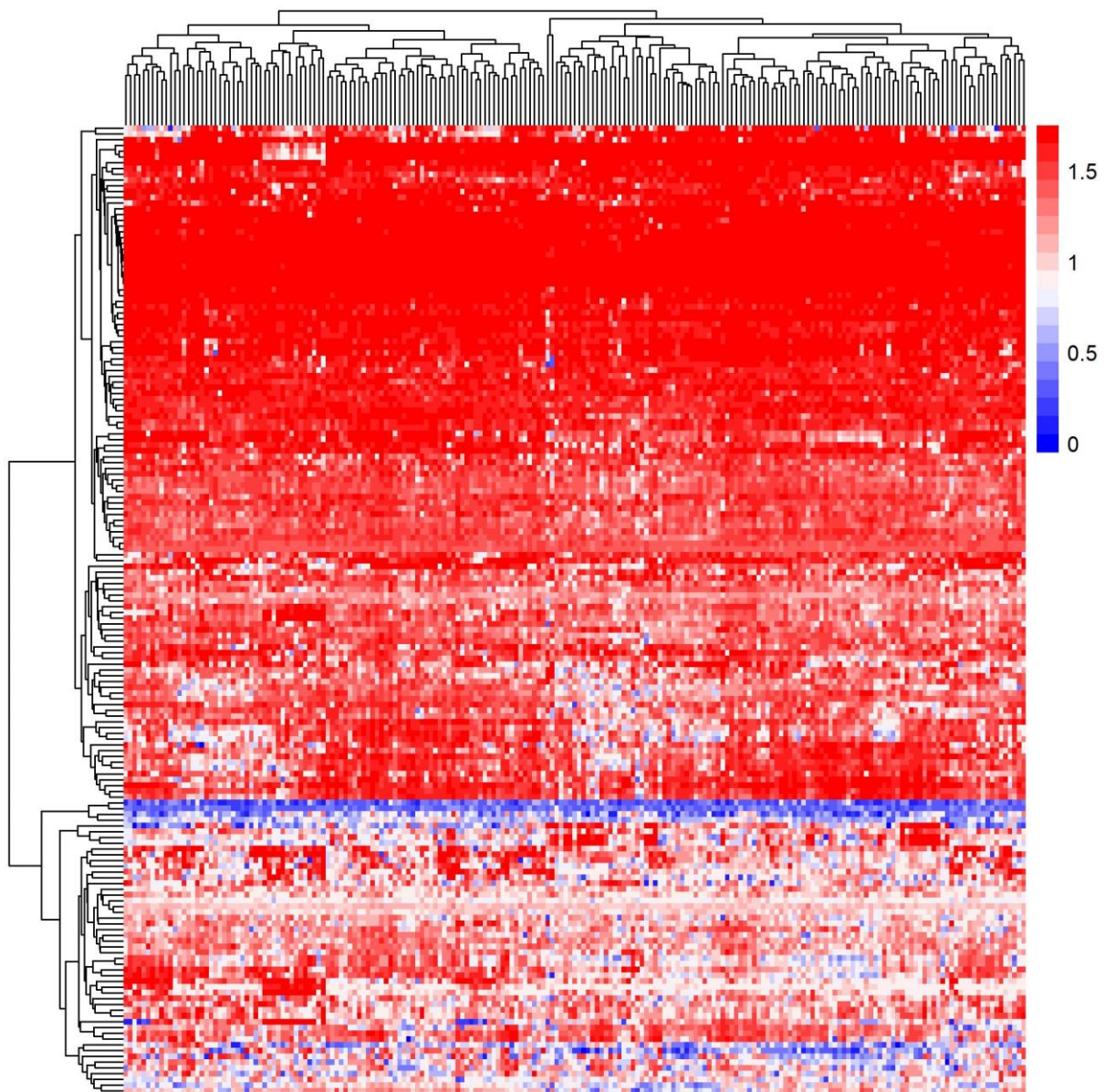
Figure 28: Heatmap of raw CV values of metabolites from Phaseolus vulgaris leaf samples across short day, long day and drought conditions of all used ecotypes. Rows show metabolites and columns show ecotypes

In addition to generally having a higher CV than primary metabolites, features of secondary metabolism also seem to have a higher variation of the CV between the ecotypes. I can subdivide the features into two clusters (Figure 28). In the upper cluster, where features generally have a high CV, it seems that the cause for the variation are few ecotypes, which have a relatively low CV. In the lower cluster, which has a few features with a much lower CV, the variation seems to be more equally distributed, between several ecotypes.

When calculating the Pearson's correlation coefficient I can also see some correlation patterns (Figure 29).
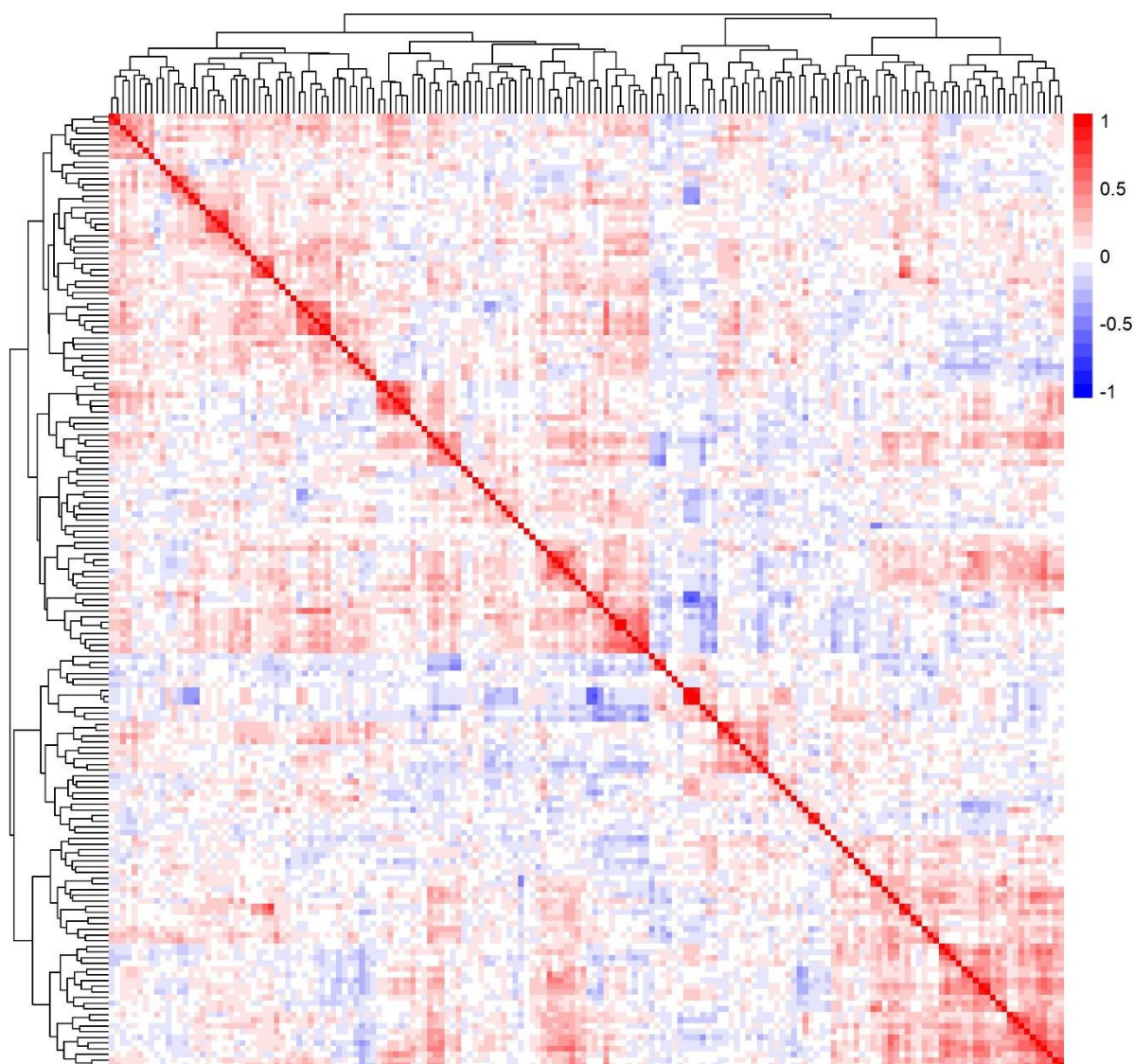
Figure 29: Heatmap of pearson-correlation coefficient of metabolites across normal light and high light and control conditions of the darkness experiment

Generally I can see several cluster with positive correlations along the diagonal. A cluster to the lower right contains mostly positive or no correlation values but only few negative ones.

38

Chapter 2: cmGWAS

I also used the transformed CV values as input for the GWAS model. I found one significant association between a locus on chromosome 5 and the CV of 2 metabolites (Figure 30).
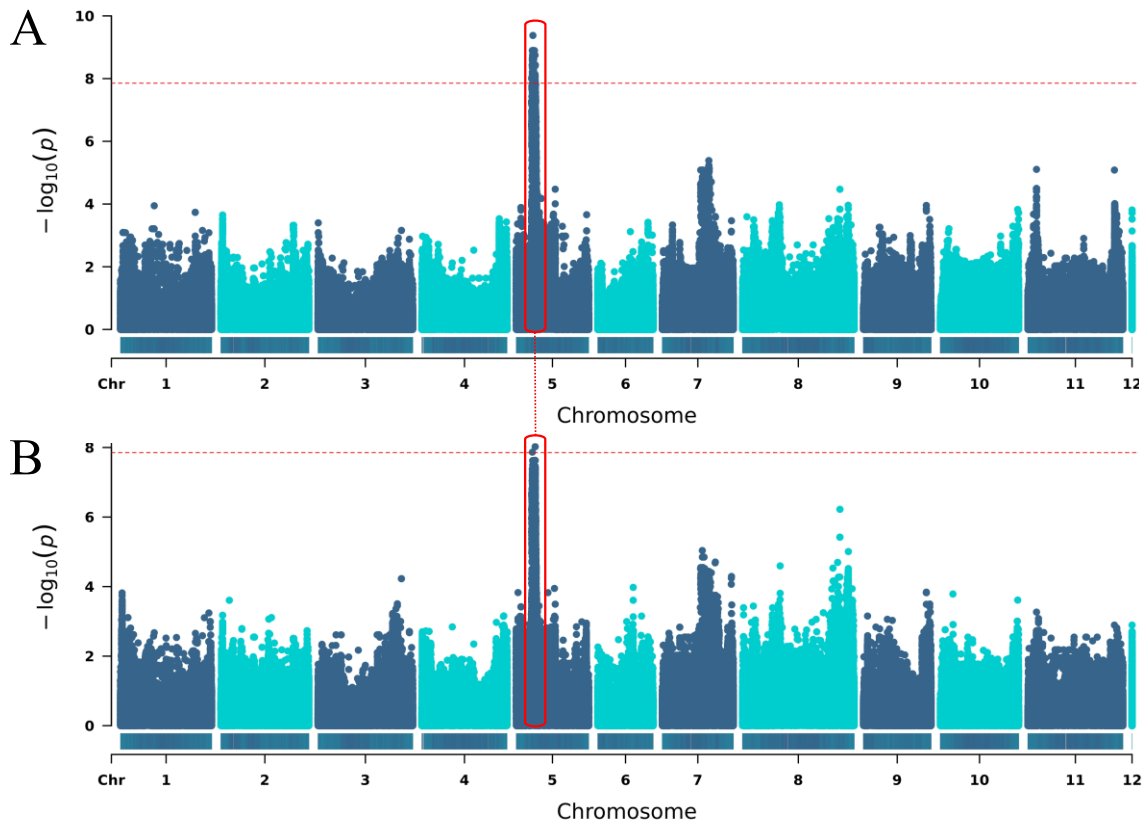
Figure 30: Manhattan plot of association of the CV of two features (A: m_126; B: m_127) to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the –log₁₀(p-value) of a calculated association. The red line shows the bonferroni threshold. The red rounded rectangle points to the discussed locus.

The gene closest to the leading SNP is annotated as an 11-oxo-beta-amyrin 30-oxidase (Phvul.005G063200). In the vicinity of this gene I found 2 more genes with the same putative annotation.

One cmQTL on chromosome 6 was particularly interesting, as it mapped 8 different features.
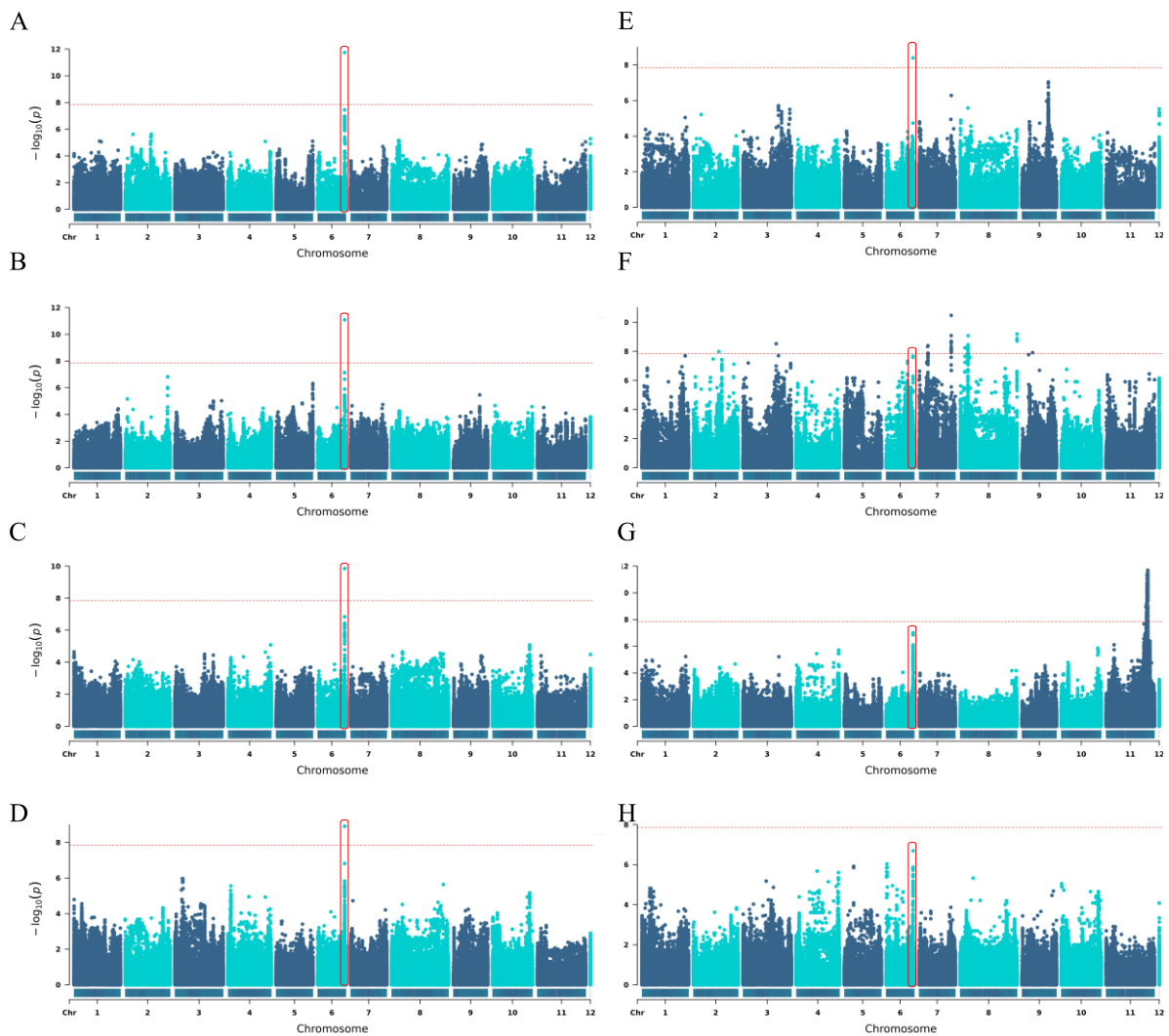
Figure 31: Manhattan plot of association of the CV of two features (A: m_47, B: m_91, C: m_79, D: m_77, E: m_92, F: m_55, G: m_56, H: m_78) to genomic regions. The x-axis shows the genomic position of the SNPs organized into chromosomes, while the y-axis displays the $-log_{10}(p\text{-}value)$ of a calculated association. The red line shows the bonferroni threshold. The red rounded rectangle points to the discussed locus.

The closest gene to the lead SNP is Phvul.006G200700, which codes for a U3 small nucleolar RNA-associated protein 18 (UTP18). In proximity is also Phvul.006G201100, which is a gene coding for a Flavonol 3-O-glucosyltransferase. I found several other cmQTL, many with more than one feature mapping to them and also many above the bonferrroni cut-off. A selection of the highest associations are shown in Table 5.

Chapter 2: cmGWAS

Table 5: Selection of putative candidate genes controlling the CV of secondary metabolites in Phaseolus vulgaris across short day, long day and drought conditions

#	LOD (best SNP)	Distance (best SNP)	Gene ID	Gene annotation	Metabolites mapping to this locus
6	8.34	0	Phvul.001G182100	UDP-glucosyl transferase 73C	m_86, m_55, m_85
11	8.74	-2770	Phvul.002G025800	MAJOR POLLEN ALLERGEN-LIKE PROTEIN	m_52, m_80, m_50, m_72, m_84
13	9.05	4362	Phvul.002G134300	ESCRT-II complex subunit VPS25	m_97, m_98
14	9.33	0	Phvul.002G135000	Domain of unknown function (DUF4219)	m_97, m_55, m_98
18	8.67	0	Phvul.002G139200	Phosphoenolpyruvate carboxykinase	m_97
27	8.53	0	Phvul.003G041700	CELL DIVISION CONTROL PROTEIN 48 HOMOLOG B	m_61
32	9.42	-6684	Phvul.003G142400	Core-2/I-Branching enzyme (Branch)	m_59
40	8.52	19083	Phvul.003G151700	EamA-like transporter family (EamA)	m_55
59	8.53	0	Phvul.005G005200	CHLORIDE CHANNEL PROTEIN CLC-F	m_61
60	8.89	2379	Phvul.005G031150		m_53, m_49, m_61
63	8.43	-13436	Phvul.005G036200		m_53
93	8.00	-13482	Phvul.005G061132	GENOMIC DNA, CHROMOSOME 3, P1 CLONE:MJM20	m_126, m_127
94	8.04	0	Phvul.005G061200	GIBBERELLIN 2-BETA-DIOXYGENASE 1	m_126, m_135, m_127
95	9.38	1734	Phvul.005G063200	11-oxo-beta-amyrin 30-oxidase / CYP72A154	m_126, m_127, m_135
95	9.38	37105	Phvul.005G063300	11-oxo-beta-amyrin 30-oxidase / CYP72A154	m_126, m_127, m_135
96	8.89	9482	Phvul.005G063650	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	m_126, m_127
98	8.44	249	Phvul.005G064000		m_126, m_127
99	8.66	13236	Phvul.005G064100	11-oxo-beta-amyrin 30-oxidase / CYP72A154	m_126, m_127
100	8.89	-10899	Phvul.005G064300		m_126, m_127
101	8.89	18425	Phvul.005G057920	phosphatidylglycerophosphatase GEP4 (GEP4)	m_126, m_127
103	8.75	12318	Phvul.005G064400	CARBOHYDRATE-BINDING X8 DOMAIN-CONTAINING PROTEIN-RELATED	m_126, m_127, m_135
364	11.74	0	Phvul.006G200700	U3 small nucleolar RNA-associated protein 18 (UTP18)	m_47, m_91, m_79, m_77, m_92, m_55, m_56, m_78
370	12.84	0	Phvul.007G016500		m_18, m_85
371	11.19	-5952	Phvul.007G018300	PROTEIN PHOSPHATASE 2C 64-RELATED	m_18
378	8.33	-7943	Phvul.007G090901	RNA SPLICING PROTEIN MRS2, MITOCHONDRIAL	m_55, m_119
380	8.39	-9388	Phvul.007G093900	ATH SUBFAMILY PROTEIN ATH8	m_55, m_119
402	8.30	-1430	Phvul.007G228800		m_61
403	8.30	-5996	Phvul.007G228900	AN1-TYPE ZINC FINGER PROTEIN	m_61
404	10.47	-211	Phvul.007G230700	B3 DOMAIN-CONTAINING TRANSCRIPTION FACTOR VAL3	m_55
406	20.77	-5354	Phvul.008G004000	SNARE associated Golgi protein (SNARE_assoc)	m_61
407	9.01	-601	Phvul.008G028800		m_116
408	12.15	1272	Phvul.008G032501	Isoflavone-7-O-beta-glucoside 6"-O-malonyltransferase	m_116
409	8.94	514	Phvul.008G034000	Cyanohydrin beta-glucosyltransferase	m_116
411	8.24	0	Phvul.008G060100	FERRIC REDUCTION OXIDASE 2-RELATED	m_55
413	8.41	-2777	Phvul.008G087901	ENTH/VHS/GAT FAMILY PROTEIN	m_55, m_58
414	9.07	1035	Phvul.008G089400		m_55

434	9.18	0	Phvul.008G290800		m_55
444	8.71	-3443	Phvul.009G093500	PHOSPHOLIPASE D ALPHA 1-RELATED	m_61
451	8.13	0	Phvul.009G127900		m_86, m_55
456	14.78	-2328	Phvul.009G258600	Anthranilate N-methyltransferase	m_114, m_117
463	8.41	0	Phvul.011G016400	RNA polymerase primary sigma factor (SIG1, rpoD)	m_46
464	16.06	0	Phvul.011G021800	Assimilatory sulfite reductase (ferredoxin)	m_46, m_119
471	8.12	0	Phvul.011G088200	PROTEIN ARGONAUTE 5	m_62
472	8.12	4458	Phvul.011G088900		m_62
515	9.17	-13706	Phvul.011G157000	IQ-DOMAIN 9 PROTEIN	m_56
519	11.36	-3574	Phvul.011G158700	PROTEIN SUPPRESSOR OF PHYA-105 1	m_56, m_62
520	9.65	3543	Phvul.011G160100		m_56, m_62
522	8.96	14831	Phvul.011G161100	11-oxo-beta-amyrin 30-oxidase / CYP72A154	m_56, m_62
523	11.70	-2088	Phvul.011G162200	RAB3-GAP REGULATORY DOMAIN	m_56, m_62
526	8.12	0	Phvul.011G180700	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 37-RELATED	m_58, m_86, m_85
527	8.93	0	Phvul.L001785	SF7 - GB	m_135, m_86

2.4. Discussion

To estimate variation of metabolic content across different environments I reutilized metabolic data from different experiments. I used the coefficient of variation or fold-change between conditions as a measure of metabolic canalization or lack thereof.

It should be mentioned at this point that combining the datasets was in reality connected with rather large challenges. In order to be able to calculate a sensible coefficient of variation, it is necessary to arrive at a sensible normalized value first. Such normalization is necessary, to overcome technical variation, which can appear as batch variation or machine drift and several methods have been developed to correct these confounding factors (Alseekh et al., 2018; Dunn et al., 2011; Fan et al., 2019; Rusilowicz et al., 2016). Unfortunately, there is always the risk of removing the wanted biological variation within one experiment along with the unwanted variation (Livera et al., 2015), so one can easily imagine it to be more difficult to additionally remove unwanted technical variation between experiments and retain true differences. I tried out several methods, which I did not show here and finally settled with a linear model based on the quality control (QC) values as described in the methods section. More promising methods certainly exist, but a higher number of QC samples is either required or at least suggested (Dunn et al., 2011; Fan et al., 2019), therefore using those methods with a small number of QCs may have added additional unwanted variation. The QC samples from the *Arabidopsis thaliana* datasets were all based on pooled extracts from Col-0 ecotype, which gives me a fairly high confidence that I have normalized experiments to a sensible baseline.

However for the *Phaseolus vulgaris* datasets I only have experiment specific pooled QCs, which meant that data had to be rescaled after normalization. I used the median of metabolic abundance as a rescaling factor for the median-centered normalized values. Even though this is a plausible approach and similar approaches have been suggested before in literature (Rusilowicz et al., 2016), this limitation should be kept in mind. Due to this difficulty of relative values for investigation of variation, obtaining absolute values could be very valuable. Absolute quantification via GC-MS is possible, but requires the availability of standard compounds for the creation of standard curves (Rosado-Souza et al., 2019). Alternatively, several metabolites such as starch, sucrose, malate, fumarate as well other organic acids and sugars can be absolutely quantified with enzymatic assays, although satisfactory time-efficiency and accuracy may only be achievable with a robotized system (Cross et al., 2006; Gibon, Blaesing, et al., 2004; Nunes-Nesi et al., 2007).

In the datasets containing primary metabolites, the CV seemed to be both metabolite and context dependent. For example metabolites like adenine or pyroglutamic acid had relatively low mean CVs in all combinations. On the other end of the spectrum, metabolites with high CV were represented by carbohydrates in the darkness treatment dataset, but contained amino acids like ornithine and asparagine, when data from the normal and high light conditions were included. This effect can easily be explained. It is well known that carbohydrates are tightly controlled in a diurnal manner, with starch almost but not completely used up at the end of the night and sucrose and glucose maintained at a more or less regular level, independent of the photoperiod (Sulpice et al., 2014). However if the night is just extended by 6 hours, starch is not detectable anymore and sucrose decreases 3-fold (Gibon, Bläsing, et al., 2004). An even longer darkness will likely have more detrimental effects to carbohydrate levels, which would explain the high CV levels. Similarly it could be possible that the levels of amino acids change with different conditions. For example, it has been shown that urea cycle compounds like ornithine accumulate under low CO_2 conditions likely through increase of photorespiration (Blume et al., 2019). As photorespiration is in general light-dependent (Peterhansel et al., 2010), it may be possible that differences in light conditions between our experiments, could explain the high CV of urea cycle metabolites.

I also investigated the correlation of metabolite CVs to each other. A high degree of correlation between different compound groups like amino acids, has already been shown in *Arabidopsis thaliana*, tomato and maize (Schauer et al., 2006; Sulpice et al., 2010; Toubiana et al., 2016).

Here, I also found the CV especially of amino acids to show some degree of correlation in different combinations of conditions.

In the GWAS for primary metabolites I only found a relatively small number of cmQTLs, most of which also do not surpass the bonferroni cut-off. As the basis of metabolic canalization is not well-understood, it is difficult to determine, which gene within the cmQTL is the causal regulator of the respective metabolite. However, in light of previous experiments I anticipated to find potential genes with clear regulatory functions (Alseekh et al., 2015). Although I do find several genes, that have known regulatory functions, I also find many enzymes in the cmQTL. Then again, only some of these enzymes are known to be directly involved in the production or degradation of the compounds they are associated to. For example in the dataset for the 3 darkness conditions I found an association of the CV of glutamine and a succinate dehydrogenase, which converts succinate to fumarate in the TCA cycle. Even though glutamate and from this glutamine can be synthesized from the TCA cycle metabolite 2-oxoglutarate, it is at least 4 conversion steps apart. However, there are some candidate enzymes, which seem even further apart from the metabolite for which I found an association. For the 3 amino acids, leucine, tyrosine and aspartic acid, I found different glucosidases and glycosylhydrolases, which are related to carbohydrate metabolic processes. Nevertheless I also found some examples of enzymes, which are directly related to the associated metabolites. The most pervasive association is the one of tyramine to tyrosine decarboxylase. The association of tyramine level to this gene has been discovered before (S. Wu et al., 2016) and I also found it in both individual conditions of the GWAS with normal and high light conditions as well as the CV datasets including these conditions. This could mean that this enzyme is not only important for the mean level of tyramine, but also for the robustness or plasticity across different environments. Tyrosine decarboxylase has been shown to be upregulated upon wounding and drought stress (Lehmann & Pollmann, 2009). Similarly, in apple plants, overexpression of tyrosine decarboxylase mediates tolerance to drought and alkalinity stress (Gao et al., 2021; X. Liu et al., 2021). How the enzyme itself is regulated, is not known but the results found here, could mean that it regulates itself for example by feedback inhibition or by its own mRNA levels. Another enzyme as potential candidate gene for the CV of its metabolite is methionine gamma-lyase (AT1G64660). This enzyme has previously been shown to be involved in the catabolism of methionine and is suggested to be responsible for methionine homeostasis (Rébeillé et al., 2006). The notion that methionine levels are highly controlled, has also been confirmed in soybean, where methionine gamma-lyase (MGL) activity seems to

be responsible for control of methionine levels and if insufficient, leads to the hyperaccumulation of S-methylmethionine, through a separate enzyme (Teshima et al., 2020). In *Arabidopsis thaliana* mutation of MGL leads to accumulation of methionine and other amino acids in flowers and may even affect transcription of other methionine biosynthesis genes (Joshi & Jander, 2009). MGL has also been shown to be upregulated by dual nematode and water stress and overexpression of the gene seems to confer resistance against nematodes (Atkinson et al., 2013).

Another group of genes I found in several cmQTLs of primary metabolites are major facilitator superfamily proteins. This protein family is together with the ATP-binding cassette (ABC) superfamily one of the major transporter families (Pao et al., 1998). These transporters use electrochemical gradients to transport small molecules across membranes and in *Arabidopsis thaliana* major facilitator superfamily proteins are suggested to be important for plant tolerance of environmental stresses (Niño-González et al., 2019).

Other putative candidate genes, could also be found, which are related to stress responses but not directly involved in primary metabolism, such as the LEA family protein, ABA2 and the myb family protein. LEA proteins are a diverse protein family, which is suggested to play an important role for stress tolerance, especially in regard to desiccation (Hundertmark & Hincha, 2008). ABA2 is involved in the production of abscisic acid (ABA), which is a plant hormone, that regulates different physiological processes but also stress responses (González-Guzmán et al., 2002). MYB proteins are a diverse family of transcription factors, which are involved in transcriptional regulation of many processes including metabolism and stress responses (Dubos et al., 2010).

Regarding the fold-change of secondary metabolites in *Arabidopsis thaliana* I found even more significant associations. Among the putative candidate genes are again different enzymes and transcription factors. Interestingly, among the enzymes, I could not find any enzyme, which have a direct known role in the metabolism of the associated compound. Although AT1G64760, a putative candidate gene for an association to quercitrin, has a beta-1-3-glucanase activity, it seems rather related to tissue morphogenesis than metabolism (Vaddepalli et al., 2017). The most striking association I found was one between a locus on chromosome 3 and 4 different glucosinolates and one unannotated compound. The closest protein coding gene is an ATP-dependent helicase family protein (AT3G31900). Unfortunately not much is known about the gene but besides the helicase activity it also has a GO annotation for "gene silencing

by RNA-directed DNA methylation"(*At3g31900 - ATP-Dependent Helicase Family Protein - Arabidopsis Thaliana (Mouse-Ear Cress) - At3g31900 Gene & Protein*, n.d.). Since helicases generally unwind RNA or DNA, they can be involved in many processes like DNA recombination and repair or transcription and RNA splicing (Abdelhaleem, 2010), it is conceivable that they may have a regulatory effect. Interesting in regard to this locus and also some of the other loci is that I found many transposable element genes. As mobile and repetitive DNA sequences, transposable elements, are believed to be major drivers of genome evolution and contributors to genetic diversification (Domínguez et al., 2020; Quesneville, 2020). In *Arabidopsis thaliana* two transposable elements have been found, which seem to have been targets of positive selection and may have been involved in the adaptation of the plant (Z.-W. Li et al., 2018).

Among the putative candidate genes, I also found some genes belonging to two known transcription factor groups, namely WRKY3 and a basic helix-loop-helix (bHLH) DNA-binding superfamily protein (Qu & Zhu, 2006). WRKY proteins are a large family of transcription factors, involved in response to various abiotic stresses and WRKY3 specifically has been shown to play a role in response to salt and MeJA stress(P. Li et al., 2021). The gene annotated as bHLH protein I found, has recently been described as a cytokinin-responsive growth regulator, which regulates cell expansion and cell cycle progression (Park et al., 2021).

The CV of metabolic features in the bean dataset is generally much higher than the CV I obtained for primary metabolites in *Arabidopsis thaliana*. Distributions are generally considered low variance or high variance if their CV is lower or higher than 1, respectively (Ospina & Marmolejo-Ramos, 2019). Given that a large fraction of metabolic features have a CV higher than one I can conclude that they have a rather large variance. Although there may of course also be a species effect, the difference is likely caused by less stringent control of secondary metabolism in contrast to primary metabolism, which is more tightly controlled (Pott et al., 2019). The observation that the CV approaches a limit of ~1.73 may seem like a technical artifact, but is actually delimited by the number of conditions across which the CV is calculated. Due to the relationship of standard deviation and mean, the CV always has a ceiling at the square root of the number of individual values (Supplementary Formula 1).

In the candidate gene list from our bean dataset I can find again many genes coding for enzymes. A good example are the genes coding for the CYP72A154 enzymes, which I found in multiple cmQTL. This enzyme has been characterized to be involved in the production of

glycyrhizzin, which is a triterpenoid saponin, present in underground parts of *Glycyrrhiza* plants (Seki et al., 2011). Saponins are well known to occur in beans and other legume plants (Fenwick & Oakenfull, 1983). An interesting locus is also the one which mapped to 8 different features. As I did not perform any proper annotation, I do not know what metabolites they may be. However as a gene in this locus close to the leading SNP was Phvul.006G201100, which is a gene coding for a Flavonol 3-O-glucosyltransferase, it is possible that some of these features are flavonoids. For example the feature m_47 for example could be the formic acid adduct of astragalin, which has previously been identified in red and pinto bean (Hu et al., 2006) and the feature m_77 could be isorhamnetin 3-O glucoside as identified in elderflower (Lin & Harnly, 2007). Phvul.006G201100 has been found to be part of a Flavonoid 3-*O*-glycosyltransferases (F3GT) cluster, enzymes of which in *Arabidopsis thaliana* transfers glucosyl, rhamnosyl and arabinosyl units to flavonoids (Souza et al., 2019). Another interesting candidate gene may be the gene which codes for an Isoflavone-7-O-beta-glucoside 6"-O-malonyltransferase. This enzyme has already been characterized in 1984 in roots of chickpeas, to transfer malonyl residues onto the 7-O-glycosides of biochanin A and formononetin (Koester et al., 1984). The candidate gene has recently also been identified in a GWAS for bean pod color (García-Fernández et al., 2021). Another candidate gene in a locus with a strong association was an Anthranilate N-methyltransferase (Phvul.009G258600). So far this enzyme has only been characterized as part of the biosynthesis of acridone alcaloids, specific to the Rutaceae family (Rohde et al., 2008), but maybe the enzyme has a different function in common bean. The strongest association I found was of one feature and a locus on chromosome 8, producing a LOD score over 20. Although the closest gene codes for a SNARE associated golgi protein, a gene annotated as Heparan-alpha-glucosaminide N-acetyltransferase (Phvul.008G004100) is also in this locus. Although not well characterized, this gene seems to be orthologous to genes, that have been lost in plant lineages that have abandoned arbuscular mycorrhizal symbiosis (Radhakrishnan et al., 2020), so maybe it is responsible for producing a compound essential for this association.

Besides enzymatic genes, I also found some genes, which could have a regulatory role. For example in one locus I found a gene coding for Protein Argonaute 5 (Phvul.011G088200). Argonaute proteins are well known for transcriptional and posttranscriptional gene silencing and Argonaute 5 seems to be involved in legume nodulation (Reyero-Saavedra et al., 2017), but it may well have other regulatory functions. Another candidate gene with potential role in transcriptional regulation is RNA polymerase primary sigma factor (SIG1, rpoD). In

Chapter 2: cmGWAS

Arabidopsis thaliana SIG1 is important for transcription of a small set of plastidic genes (Macadlo et al., 2020).

In summary I can say that I found many associations between the CV or fold change of metabolic features and genomic regions. Even with a threshold less conservative then the bonferroni cut-off I do not find a large number of associations for the primary metabolites of *Arabidopsis thaliana*. For the fold-change of secondary metabolites and the CV of secondary metabolites of common bean I find however a higher number of associations as I do for the primary metabolite datasets. This is in agreement with the observation from regular mGWAS, where primary metabolism has been shown to be controlled by many small effect loci and secondary metabolism by few large effect loci (Fang & Luo, 2019). This is likely due to the fact that central metabolism is generally more tightly regulated and interconnected, then specialized metabolism (Pott et al., 2019). This may even be more relevant when considering the CV of metabolites, as primary metabolites will likely be more constant under different non-stress environments, while secondary metabolites could allow for a fine-tuning adaptation to a specific environment.

Considering putative candidate genes I find both enzymatic and regulatory genes, that could plausibly be causal for the canalization of metabolite content across different environments which is in agreement with findings from a cmQTL mapping approach in tomato introgression lines (Alseekh et al., 2017). In contrast to that study however, I found much fewer associations in total and I did not find a hotspot for regulatory function of several primary metabolites. The reason for this may be related to the different mapping approaches. Firstly, the tomato introgression line mapping population may have synthetically decanalized metabolites in some lines, by replacing genome segments of the domesticated tomato with those of the wild species where not all genes are homologous. In fact, a closely investigated cmQTL for phenylalanine was characterized, by several presence-absence variants in the *S. pennellii* genome (Alseekh et al., 2015). Such replacements would lead to a much larger genetic variability. On the other hand with a GWAS approach, I am limited to genetic variation and connected to that phenotypic variation that is already present in the population. Considering that primary metabolites are needed for growth under all conditions, the natural variation and therefore the CV of the respective metabolite might be relatively low in the whole population. In secondary metabolism however, where I generally find much more variation, I am also able to find stronger associations. Here also, both in *Arabidopsis thaliana* and *Phaseolus vulgaris* I find some loci with multiple metabolites associating. Not always do I find apparent enzymatic genes

48

in those loci, which could mean, that there are key regulatory factors, which are responsible for the robustness of several metabolites simultaneously. Given the relatively large amount of enzymatic genes I find, albeit not always directly related to the associated metabolite, one may consider that enzymes or metabolism itself confers a certain amount of canalization to metabolites. The observation that the CV of metabolites like amino acids show moderate to high correlations to each other may support this hypothesis.

In conclusion, based on our results, I hypothesize that metabolic robustness may be mediated, both by the inherent robustness of metabolic networks and regulatory genes. However, there may be even more mechanisms responsible for canalization of metabolism, which has yet to be uncovered.

Chapter 3: cmQTL validation

3.1. Introduction

The previous chapter dealt with finding new candidate genes, responsible for mediating canalization of metabolism. In this chapter I am describing experiments to validate some candidate genes that were found in a previous cmQTL mapping experiment (Alseekh et al., 2017). This experiment found several cmQTL for the variation of tomato pericarp metabolites across different conditions and is one of the first to show that the concept of canalization can be applied to metabolism. Genomic loci, could be strongly reduced to only several genes, which builds a great basis for further validation of potential candidate genes.

Before candidate genes can be validated, the relevant candidate genes first have to be selected, which can be quite challenging, especially if *a priori* knowledge is limited. Once suitable candidate genes are determined, they can be validated by different reverse genetic tools like gene silencing, insertional mutagenesis or transgene overexpression (Ben-Amar et al., 2016). A newer technique, which is now commonly used for gene validation is CRISPR/Cas9 (Curtin et al., 2017). The clustered regularly interspaced short palindromic repeat (CRISPR) DNA sequences and CRISPR associated (Cas) genes are part of a bacterial immune system against invading bacteriophages(Garneau et al., 2010). The endogenous immune system works by first integrating short viral sequences, so-called protospacers next to repeats of the host genome, which are then used together to produce CRISPR RNA (crRNA), which hybridize with complementary protospacers and are silenced by Cas proteins (Jinek et al., 2012). Although canonically, a transactivating crRNA (tracrRNA), that hybridizes with crRNA is needed for Cas9 cleavage, it was soon discovered that this duplex can be replaced by a chimeric single self-pairing RNA, that contains both the crRNA and tracrRNA parts (Jinek et al., 2012). As this allowed to target essentially any desired DNA sequence, this discovery widely opened the door for CRISPR/Cas9 as a biotechnological tool (Adli, 2018).

Besides its original function of introducing double-strand breaks, the CRISPR/Cas9 system has further been developed to be used for gene activation or repression, epigenome editing, DNA imaging and base editing and has since become a versatile tool for molecular biology as well as a promising strategy for precision plant breeding, promising to substantially cut down on development time of elite plant varieties (Chen et al., 2019). In order to ensure food security for a growing population under additional challenges posed by climate change, a tool allowing the quick development may prove quite helpful (Massel et al., 2021).

In this chapter I am selecting candidate genes and are using a CRISPR/Cas9 system, with a double target approach, to try to introduce larger deletions into these candidate genes in tomato plants, thereby rendering the genes unfunctional. By measuring the metabolite profile of tomato pericarp of different knock-out lines, grown under different water conditionsI am able to test whether their function was relevant for robustness of metabolites.

3.2. Materials and Methods

3.2.1. Orthology search

Orthology search between *Arabidopsis thaliana* (TAIR10), *Solanum lycopersicum* (ITAG3.2), *Solanum tuberosum* (PGSC v3.4) and *Solanum pennellii* (Spenn v2) was performed by OrthoFinder v2.0.0 (Emms & Kelly, 2019). Data was obtained from TAIR (https://www.arabidopsis.org/) and Solgenomics (https://www.solgenomics.net/).

3.2.2. Candidate gene selection

Candidate genes for cmQTL validation were selected from previously characterized cmQTLs (Alseekh et al., 2017), by a combination of a correlation approach and a manual curation. For the correlation approach a permutation correlation test was performed between the CV of metabolite values and gene expression values available in the tomato functional genomics database (Fei et al., 2011) at http://ted.bti.cornell.edu/.

3.2.3. guideRNA design

The webtool CRISPR-P 2.0 (http://crispr.hzau.edu.cn/CRISPR2/) (H. Liu et al., 2017) was used to design sequence-specific guide RNAs. Two guide RNAs per target were selected, which were ~40-200 bp apart, while considering low off-target and high on-target scores.

3.2.4. Cloning

Cloning was performed as described previously (Reem & Van Eck, 2019). The cloning procedure, uses a two-step Golden Gate Cloning reaction, to assemble two guideRNA-expressing cassettes, a kanamycin resistance gene, and the Cas9 nuclease in the final vector, which is transferred to *Agrobacterium tumefaciens*.

3.2.5. Plant transformation

Plant transformation was performed as a service by the greenteam of the MPIMP, similar to previously published literature (Van Eck et al., 2019). For the transformation, cotyledon segments of *Solanum lycopersicum cv. Money Maker*, are co-cultivated with *Agrobacterium tumefaciens* GV 2260, carrying the respective CRISPR/Cas9 vector and then regenerated to

calli on sterile medium. Calli are transferred several times onto different sterile selection media until shoots can be generated, and transferred to rooting medium.

3.2.6. Genotyping/Sequencing

After plant transformation, shoots from well-rooted putatively transformed tomato plants are cut and transferred to fresh sterile media containing containing agar (6.8% m/v), MS, sucrose (2% m/v) and 125 µg/ml Ticarcillin, to eliminate residual *Agrobacterium tumefaciens*. Plants are cultivated for 4-6 weeks under sterile conditions, until they have rooted again and grown back to the original size. This process is repeated at least 2 more times. At the last transfer step, leaf samples are collected and DNA is extracted, similarly as previously reported (Lu, 2011). A first PCR amplifying the chromosomal rpoB gene of *Agrobacterium tumefaciens* GV2260, to check for residual agrobacterial contamination. A second PCR is performed with primers specific to the Cas9 gene. The third PCR amplifies a region ranging from ~100 bp upstream of the first gRNA to ~100 bp downstream of the second gRNA. PCR fragments are separated on an agarose gel (4% m/v in TAE), to detect large or small indels. If bands from putative genome-edited plants could not be distinguished from bands of wild type plants, PCR fragments were subjected to a heteroduplex mobility assay, as described previously (Bhattacharya & Meir, 2019). If bands were still indiscernible, PCR fragments were send for Sanger sequencing to LGC Genomics (LGC Genomics GmbH, Berlin, Germany). Agrobacterium-free, Cas9-positive, genome-edited plants were transferred to the greenhouse for propagation and investigation.

3.2.7. Growth conditions

Putative transformants were kept under sterile conditions in a tissue culture chamber (York International/Johnson Controls; Cork, Ireland) with ~35 µmol m^{-2} s^{-1} of light in a (16h/8h)-(22°C/22°C)-(day/night) cycle. Selected transgenic plants were propagated under standard greenhouse conditions for seed production.

Seeds were generally first sown on soil in a young plant phytotron (York International/Johnson Controls; Cork, Ireland) with 150 µmol m^{-2} s^{-1} of light in a (16h/8h)-(22°C/18°C)-(70% RH/70%RH)-(day/night) cycle and transplanted into individual pots after cotyledon were fully expanded. After 4 weeks, seedlings were transplanted into bigger pots and transferred to greenhouse chambers. Plants were fertilized once after transplanting to big pots and once at flowering. Axillary meristems were continuously removed and the primary shoots were trimmed off after development of the 4th truss. During growth, pots are watered multiple times

daily by an automated drip irrigation system. Standard greenhouse settings are a (16h/8h)-(22°C/20°C)-(50% RH/50%RH)-(day/night) cycle.

In an experiment to test the mutant's effect on metabolic canalization, different watering regimes were applied. Plants were watered with 100%, 80%, 60% and 40% of the standard watering amount.

3.2.8. Phenotyping

Different phenotypic traits were assessed. Plants were checked every 2 days and the time of the first red ripe fruit was documented. At fruit harvest, all ripe and unripe fruits per plant were counted and, weighed on a scale (Sartorius; Göttingen, Germany). Average fruit weight per plant was calculated by dividing, total fruit weight per plant by the number of fruits. After fruit harvest, the remaining above-ground biomass was weighed (Sartorius; Göttingen, Germany) to estimate shoot fresh weight. Total biomass and harvest index were calculated as the sum and the ratio, respectively, of total fruit weight and shoot fresh weight. Shoot dry weight was estimated after drying the fresh shoots in an oven at 70°C for 3 days.

3.2.9. Sample collection

Leaf samples were collected from mature non-senescing plants and pericarp samples were collected from red ripe tomato fruits, when 80%-100% of fruits were ripe.

3.2.10. Metabolite extraction

Frozen plant tissue was ground to powder using a Mixer Mill (Retsch; Haan, Germany) at 30 Hz for 1 min. Aliquots of 50 mg were used for metabolite extraction, as described before (Salem et al., 2016). Here, 1 ml of pre-cooled MTBE-MeOH (3:1; vol/vol) extraction buffer is added to each sample and samples are incubated 10 minutes on an orbital shaker at 4°C. Samples are sonicated for 10 minutes in an ice bath before 0.5 ml of MeOH-H_2O (3:1; vol/vol) is added and samples are centrifuged at 14000 rpm for 5 minutes at 4°C, leading to a phase separation. An aliquot of the upper (apolar) phase was taken for lipid analysis, the rest of the upper phase was aspirated with the BVC fluid aspiration system (Vacuubrand Inc; Essex, CT, U.S.A.) and two aliquots of the lower (polar) phase were taken for GC-MS and LC-MS. Extracts were dried in a Scan Speed 40 centrifugal vacuum concentrator (Labogene; Allerød, Denmark) coupled to a Scanvac CoolSafe cryo unit (Labogene; Allerød, Denmark) with 1000 g and 30°C for 3h or overnight. Dried extracts were kept at -80°C until further use.

3.2.11. Metabolite profiling

Primary metabolites were analyzed as described before (Lisec et al., 2006). For the analysis via GC-MS-TOF a dried aliquot of the polar phase was derivatized by adding 40 µL of (20 mg $* \text{mL}^{-1}$) methoxyamine hydrochloride and incubating at 37°C first for 2 h followed by addition of 70 µL of MSTFA and continued incubation at 37°C for 30 min. The samples were injected by an autosampler Gerstel Multi-Purpose system (Gerstel GmbH & Co.KG, Mülheim an der Ruhr, Germany) into a gas chromatograph coupled to a time-of-flight mass spectrometer (Leco Pegasus HT TOF-MS) (LECO Corporation; St. Joseph, MI, U.S.A.). Analysis of secondary metabolites followed a previously described protocol (Giavalisco et al., 2009). Here a dried aliquot of the polar phase was resuspended in 150 µl H2O:MeOH (50:50) and samples injected to an Acquity UPLC system (Waters Corporation; Milford, MA, U.S.A.) coupled to an Exactive Orbitrap mass detector (ThermoFisher Scientific; Waltham, MA, U.S.A.) via a heated electrospray source (ThermoFisher Scientific; Waltham, MA, U.S.A.) . Mass spectra were obtained by running samples in negative ionization mode. For analysis of lipophilic compounds a previously established protocol was used (Hummel et al., 2011). An aliquot of the dried organic phase was resuspended in 100 µL of UPLC-grade acetonitrile:isopropanol (70:30) mix, of which 2 µl were injected onto an Acquity UPLC system (Waters Corporation; Milford, MA, U.S.A.) equipped with an RP C8 column (Hummel et al., 2011). Mass spectra were obtained by running samples in positive ionization mode on an Orbitrap high-resolution mass spectrometer: Fourier-transform mass spectrometer (FT-MS) coupled with a linear ion trap (LTQ) Orbitrap XL (ThermoFisher Scientific; Waltham, MA, U.S.A.).

3.2.12. Peak picking/Area calculation

For targeted analysis peaks were picked manually and peak area calculated with Xcalibur Version 4.2.47 (ThermoFisher Scientific; Waltham, MA, U.S.A.). Non-targeted analysis was performed with Genedata Expressionist ® 14.0.5 (Genedata; Basel, Switzerland).

3.2.13. Metabolite data normalization

Missing values were imputed by the half-minimum of a certain metabolic feature in the respective run or by using a QRILC approach as suggested by literatures (Wei et al., 2018). Raw peak area of each metabolic feature was then normalized to the area of an internal standard (ribitol for GC-MS and isovitexin for LC-MS) if applicable. To account for batch and drift effects pooled quality control samples were regularly injected in between samples as a basis for a QC-based locally estimated scatterplot smoothing. Further normalization, was done by sample weight.

3.2.14. Computational analysis, data visualization and used packages

All analysis was performed using R statistical software v4.0.2 in the RStudio environment or on a unix-based high performance computing cluster. Code was written in base R or with the help of the packages tidyverse, broom, data.table, modelr, ggpubr, ggtext, glue and cowplot (Dowle et al., 2021; Hester, 2020; Kassambara, 2020; Robinson et al., 2021; Wickham et al., 2019; Wickham & RStudio, 2020; Wilke, 2020a, 2020b). Imputation by QRILC was done with the package imputeLCMD (Lazar, 2015). Heatmaps were created with the pheatmap package (Kolde, 2019). I am using the bootstrap::jackknife function to generate jack-values, which I am using for statistical analysis (original et al., 2019).

3.3. Results

I selected putative candidate genes, responsible for canalization of metabolites out of previously identified cmQTLs, resulting from a QTL mapping in tomato ILs, which were further narrowed down by the BIL population (Alseekh et al., 2017). An example for phenylalanine is shown below (Figure 32). I chose three separate cmQTL from chromosome ten for experimental validation. The first cmQTL showed an association to phenylalanine, the second one to malate and the third one for fructose-6-phosphate, glucose-6-phosphate and maltose (Table 6).
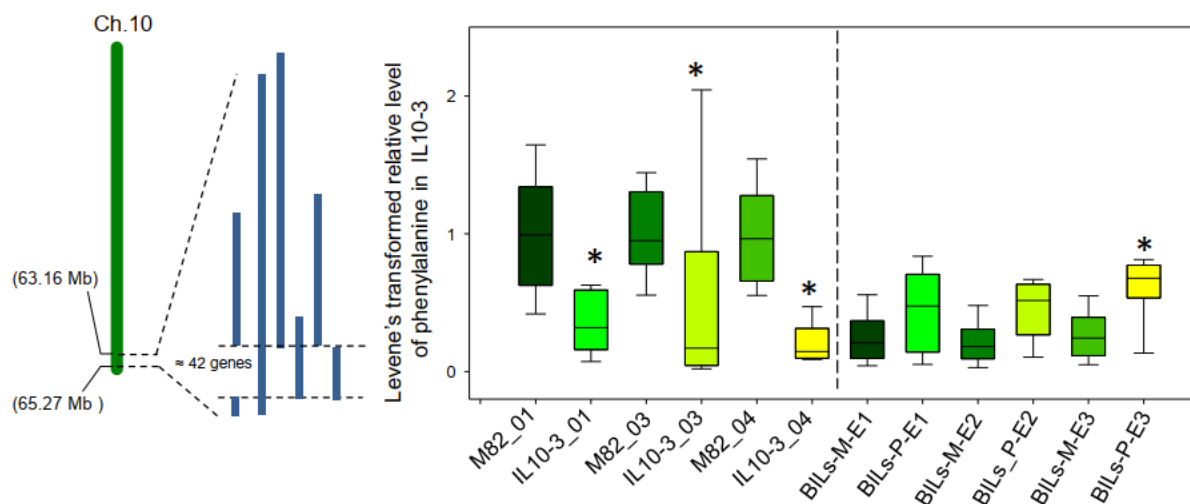
Figure 32: cmQTL for phenylalanine validated by BIL population (Figure 6 from (Alseekh et al., 2017))

3.3.1. Candidate gene selection

I correlated metabolite CV with gene expression values from pericarp of the ILs with a permutation correlation test. I additionally selected two genes based on their expression, amino acid identity, promoter conservation and orthology comparing the cultivated tomato *S. lycopersicum* and its wild relative *S. pennellii*. The results are shown in Table 6.

Table 6: Candidate genes selected by permutation correlation test and manual selection

IL segment	Metabolite	Gene ID	Annotation	Name	p-value	Correlation value
10-1	Fru6p, Glu6p, Maltose	Solyc10g007190	Transposon protein	TRANSP1		
10-1	Fru6p, Glu6p, Maltose	Solyc10g007340	Unknown protein	UP1		Manual selection
10-1	Fru6p, Glu6p, Maltose	Solyc10g007350	Multiprotein bridging factor 1	MBF1	0.014/0.017	-0.29/-0.28
10-1	Malate	Solyc10g074590	Pantothenate kinase 4	PANK4	0.05185	0.23
10-1	Malate	Solyc10g074630	ADP-ribosylation factor	ARLB	0.00835	-0.31
10-1	Malate	Solyc10g074720	Cell division protein kinase 2	CDKB1;2	0.0275	-0.26
10-3	Phenylalanine	Solyc10g086080	RING finger protein B	LOG2	0.05	-0.23
10-3	Phenylalanine	Solyc10g086190	Adenosine kinase	ADK1	0.03855	0.24

I used results from an orthology search between tomato, potato, *Solanum pennellii* and *Arabidopsis thaliana* to check if anything is already known about orthologues of the selected genes. Solyc10g007190 is annotated to code for a transposon protein, but it has no orthologues in *Arabidopsis thaliana*. Solyc10g007340 codes for an unknown protein, with no orthologues in *Arabidopsis thaliana*. Solyc10g007350 is orthologous to AT2G42680 (MBF1A) and AT3G58680 (MBF1B), which are transcriptional coactivators, induced by several stresses (Sugikawa et al., 2005). Solyc10g074590 is orthologous to the genes AT2G17320, AT2G17340 and AT4G35360 annotated as pantothenate kinases. The protein structure of the product of AT2G17320 has been determined and the C-terminal domain has been identified as homologous to PANK2 and it is therefore possibly also a pantothenate kinase, which phosphorylates pantothenate to 4`-phosphopantothenate (Bitto et al., 2005; Tilton et al., 2006). Solyc10g074630 is homologous to the ARF-like GTPase AT5G52210 (ARLB) (Paul et al., 2014). Solyc10g074720 is a cell division protein kinase, orthologous to AT2G38620. In an as of yet unpublished study it occurred as a differentially expressed gene in a mutant plant with chlorotic leaves (Gu et al., 2020). The gene Solyc10g086080 is orthologous to AT3G09770 (LOG2) and AT3G53410 (LUL2). LOG2 (LOSS OF GDU2) is a RING-type E3 ubiquitin ligase and together with GDU1 (GLUTAMINE DUMPER1) associates to membranes and is suggested to be involved in the regulation of amino acid export from plant cells (Pratelli et al., 2012). Solyc10g086190 is annotated as adenosine kinase and is orthologous to AT3G09820 (ADK1) and AT5G03300 (ADK2). Both these genes have been shown to be relevant for cytokinin interconversion (Schoor et al., 2011).

After selecting the candidate genes I designed two guide RNAs per construct, for the CRISPR/Cas9-mediated knock-out of the genes and cloned the constructs after which they were transformed into *Solanum lycopersicum* cv. MoneyMaker via Agrobacterium-mediated transformation. After regeneration I tested putative transformants for the presence of Cas9 and for potential indels at the target loci.

3.3.2. Transformation and gene-editing efficiency

Although I was always able to identify several transformed plants per line, sometimes only one of them actually had a gene-edit (Figure 33, Figure 34).

Figure 33: Combined boxplot and dot plot of transformation and gene-editing efficiency. Cyan: Total number of plants; Magenta: Efficiency in %; TF: transformation, GE: gene-editing

As indicated in the figure, the transformation efficiency was lower than the gene-editing efficiency (Figure 33). Either way I obtained several lines, which had a gene-edit confirmed by DNA sequencing of the target region (Figure 34).
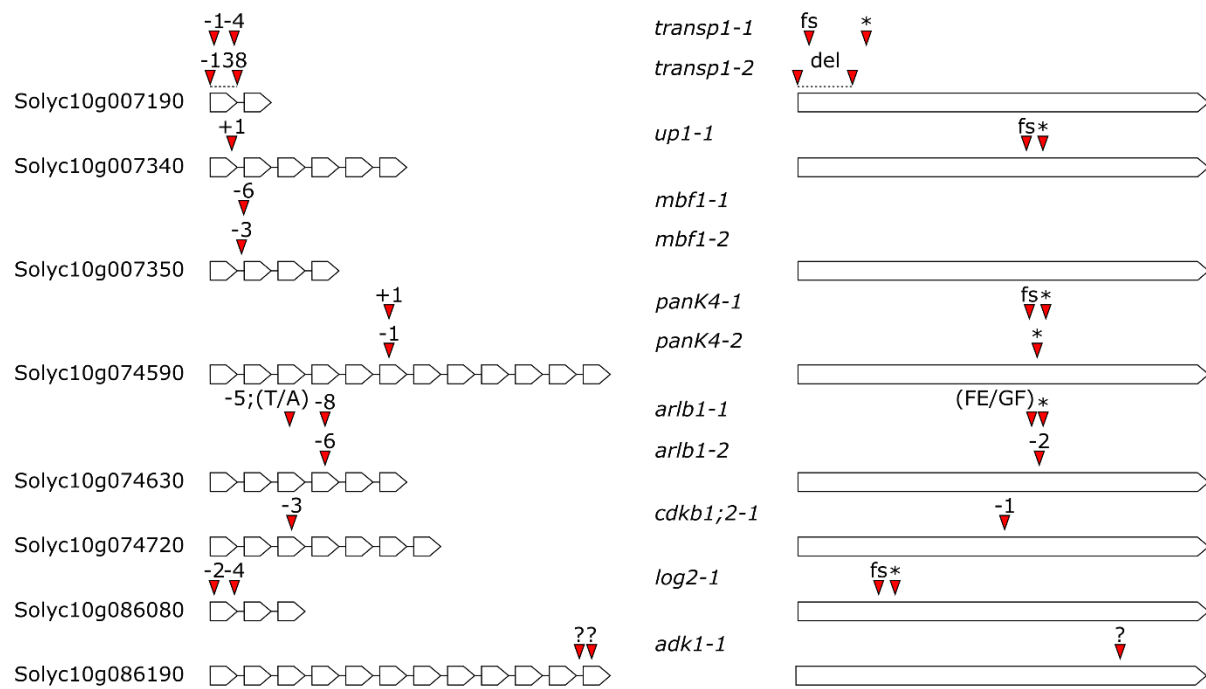
*Figure 34: Overview of genes and DNA edits (left) and proteins and resulting amino acid changes (right) in the .corresponding mutants. Boxes represent exons and whole proteins respectively. Insertions and deletions are marked with red arrowheads. Exons and introns as well as proteins are not drawn to scale and positions for mutations and amino acid changes are approximate to support readability. fs: frame shift; *: stop codon*

Not all of these lines were included in further investigations for differing reasons. Not always could homozygous plants be found in the initial screening, which is why their analysis has been put on hold. In the case of Solyc10g007340 one line did not produce any seeds, which is why I could not analyze it. In the case of Solyc10g007350 I could unfortunately only find gene-edits in the first intron, which resulted in no alteration of the amino acid sequence. The chromatograms of the *adk1-1* mutant still showed some unreadable sections, which makes it impossible to determine the exact change to the DNA sequence.

3.3.3. Screening of T1 generation

After obtaining seeds from the initial T0 transformants, I strived to screen for homozygous gene-edited plants. Optimally plants also did not have any Cas9 gene anymore, to prevent any future off-target editing. For screening T1 seedlings, literature recommends using at least 50 plants per line (Reem & Van Eck, 2019). As this is a relatively large number of plants, I implemented a heteroduplex mobility assay, originally developed for screening of gene-edited mice, that reduces the need for sequencing (Bhattacharya & Meir, 2019). With this method I was able to detect indels with a combined minimum length of only 5 bp (Figure 35).
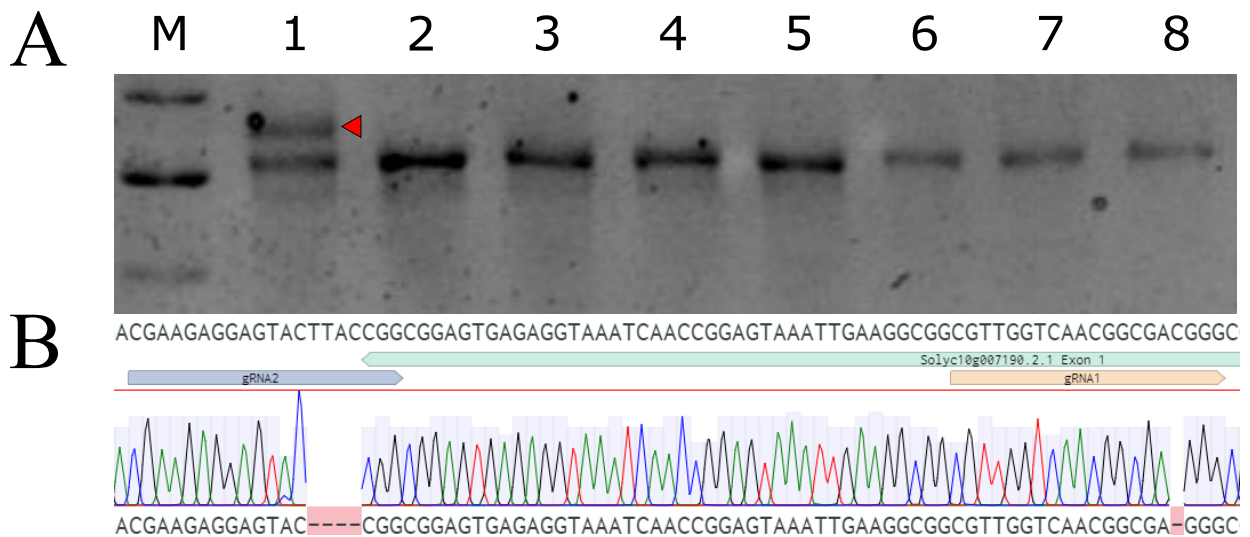
Figure 35: Example of a successful heteroduplex mobility assay, confirmed by Sanger sequencing. A: Image of gel-electrophoresis after additional heteroduplex inducing cycle. The red arrow points to the additional band caused by the heteroduplex. B: Sequencing chromatogram of sample 1 in A; M: Marker for DNA

In the example shown above I can see that the gene-editing introduced a 4 bp and 1 bp deletion into the Solyc10g007190 gene (Figure 35).

3.3.4. Morphological phenotypes

The lines I generated can be divided into two groups, with some showing a clearly visible phenotype and some, which look wild type-like. The lines *arlb1-1* and *arlb1-2* for example showed a slightly dwarfish phenotype and some outgrowth at nodes of leaflets (Figure 36).

Figure 36: Phenotype of arlb1 mutant lines compared to wildt ype. A: Whole plant B: Close-up on leafs at leaflet branching points.

It can be seen, that the heterozygous plants show an intermediate phenotype, concerning the outgrowth. Further on, the fruits from these plants also showed scarring of the cuticle (Figure 37).
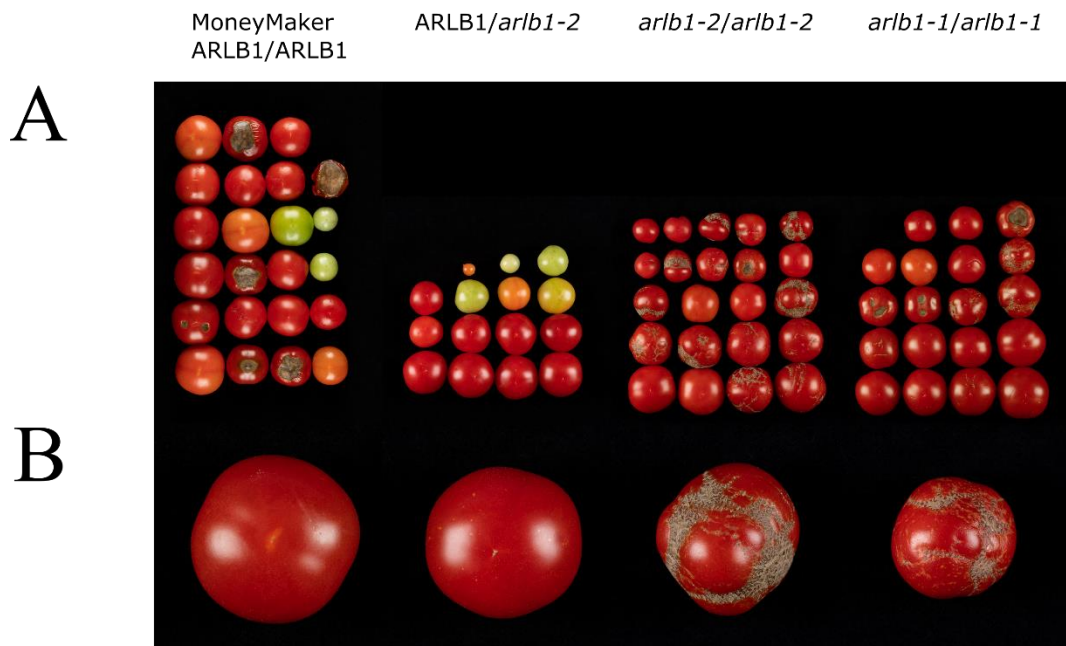

Figure 37: Phenotype of arlb1 mutant lines compared to wild type. A: All fruits of one plant B: Close-up of individual fruits with the stylar end facing upwards

In this case the fruits from heterozygous plants did not show an intermediate phenotype but were rather wild type-like (Figure 37). The fruits from the *adk1-1* line also showed a clear phenotype, as they were relatively small and most of them were completely seedless (Figure 38). The *cdkb1;2-1* line on the other hand had pale leaves and a few, small fruits (Figure 38).. Although fruits produced some seeds, attempts at propagating the next generation have so far failed, as seedlings do not survive the seedling stage, even under tissue culture conditions and therefore vegetative propagations have been performed.

Figure 38: Phenotypes of gene-edited mutant lines. A: fruits of adk1-1 plants; B: Cross-section of seedless adk1-1 fruits; C: Pale green leaves from cdkb1;2-1; D: Small fruits of cdkb1;2-1 on the plant

I grew plants in two experiments. The first experiment was performed with the lines showing no clear morphological or developmental phenotype. I used different watering conditions, with 40%, 60%, 80% and 100% of the standard optimal watering amount per day.

3.3.5. Drought experiment

Before investigating the metabolite profile, I measured whole plant phenotypic traits, like average fruit weight and days until ripening.

3.3.5.1. Whole plant phenotypic traits

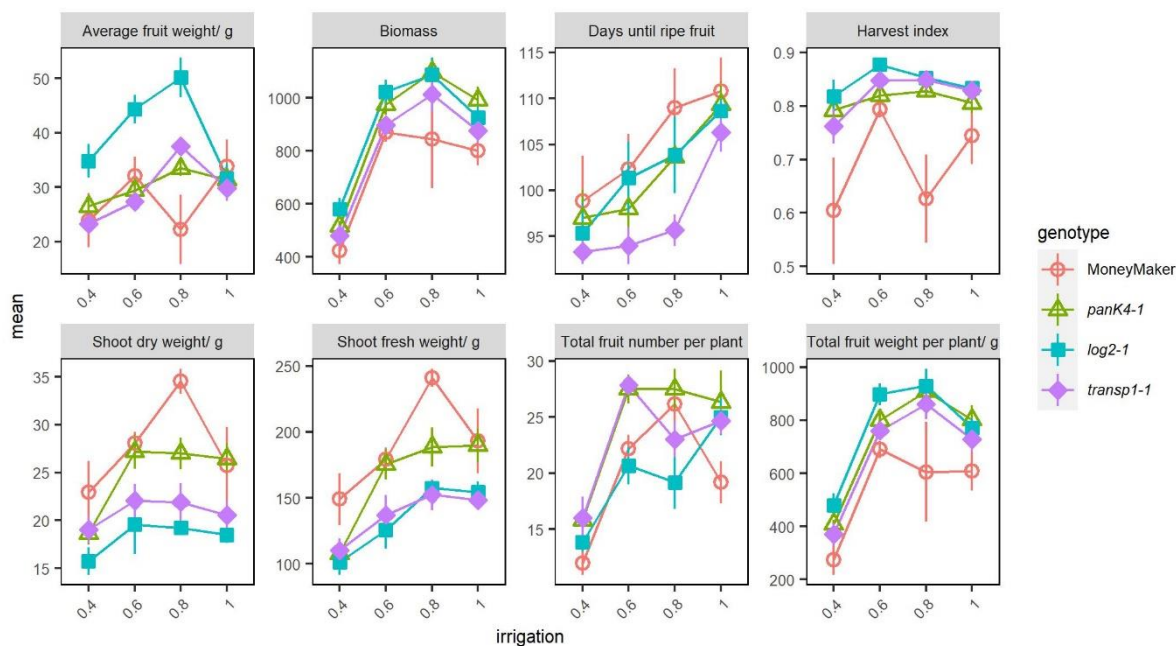During the growth period and at fruit harvest I measured several phenotypic traits (Figure 39).

Figure 39: Mean level of several phenotypic traits under 0.4, 0.6, 0.8 and 1x the optimal watering condition. Shown are mean values ± the standard error. Shapes: empty circle: Money Maker; empty triangle panK4-1; filled square: log2-1; filled diamond: transp1-1

As expected, for all genotypes, I see an increasing trend of traits like total fruit weight, when increasing watering amount. Interestingly, most of the traits peak at the 80% condition, which may indicate that the optimal watering amount may have been overestimated and plants actually suffered some mild water stress when receiving the full amount.

The ANOVA shows that genotype and environment have a significant effect on all traits, but only for the total fruit number per plant and the average fruit weight do I find a genotype-by-environment interaction (Table 7).

Table 7: P-values resulting on ANOVA of measured phenotypic traits. P-values lower than 0.05 are shaded in red. G: genotype, E: environment (here:irrigation); GxE: genotype-by-environment interaction.

Trait	G	E	GxE
Total fruit number per plant	1.96E-04	3.85E-14	1.94E-02
Total fruit weight per plant/ g	2.69E-05	1.32E-15	8.83E-01
Average fruit weight/ g	8.92E-07	1.91E-03	2.03E-03
Days until ripe fruit	7.68E-03	1.38E-06	9.53E-01
Shoot fresh weight/ g	5.33E-09	1.50E-10	2.59E-01
Harvest index	4.75E-07	1.13E-02	2.92E-01
Biomass	8.50E-04	1.37E-18	9.54E-01
Shoot dry weight/ g	5.07E-09	9.43E-05	2.94E-01

Chapter 3: cmQTL validation

When taking a closer look on the individual genotypes I can see that all of them have statistically significant higher total fruit weight per plant and lower shoot fresh weight and as a consequence a higher harvest index (Table 8).

Table 8: Adjusted p-values from TukeyHSD on significant effects of individual genotypes compared to MoneyMaker. P-values lower than 0.05 are shaded in red.

Trait	panK4-1	log2-1	transp1-1
Total fruit number per plant	1.99E-03	9.98E-01	6.25E-02
Total fruit weight per plant/ g	6.63E-04	2.67E-05	2.07E-02
Average fruit weight/ g	7.59E-01	2.86E-06	9.10E-01
Days until ripe fruit	4.73E-01	5.54E-01	3.80E-03
Shoot fresh weight/ g	2.74E-02	7.82E-08	2.36E-07
Harvest index	1.57E-04	1.04E-06	3.41E-05
Biomass	3.61E-03	1.88E-03	2.75E-01
Shoot dry weight/ g	1.53E-01	1.04E-08	2.65E-05

Some traits are only significant for one or two lines. For example the time until the first ripe fruit was significantly reduced in the *transp1-1* line and average fruit weight was significantly increased in the *panK4-1* line. However, as already indicated by the genotype-by-environment effect, only few traits showed significant differences at individual conditions, between the gene-edited lines and wild type plants. Two traits that are significantly increased in the *log2-1* mutant, compared to wild type plants, are total and average fruit weight at 80% water (Figure 40)
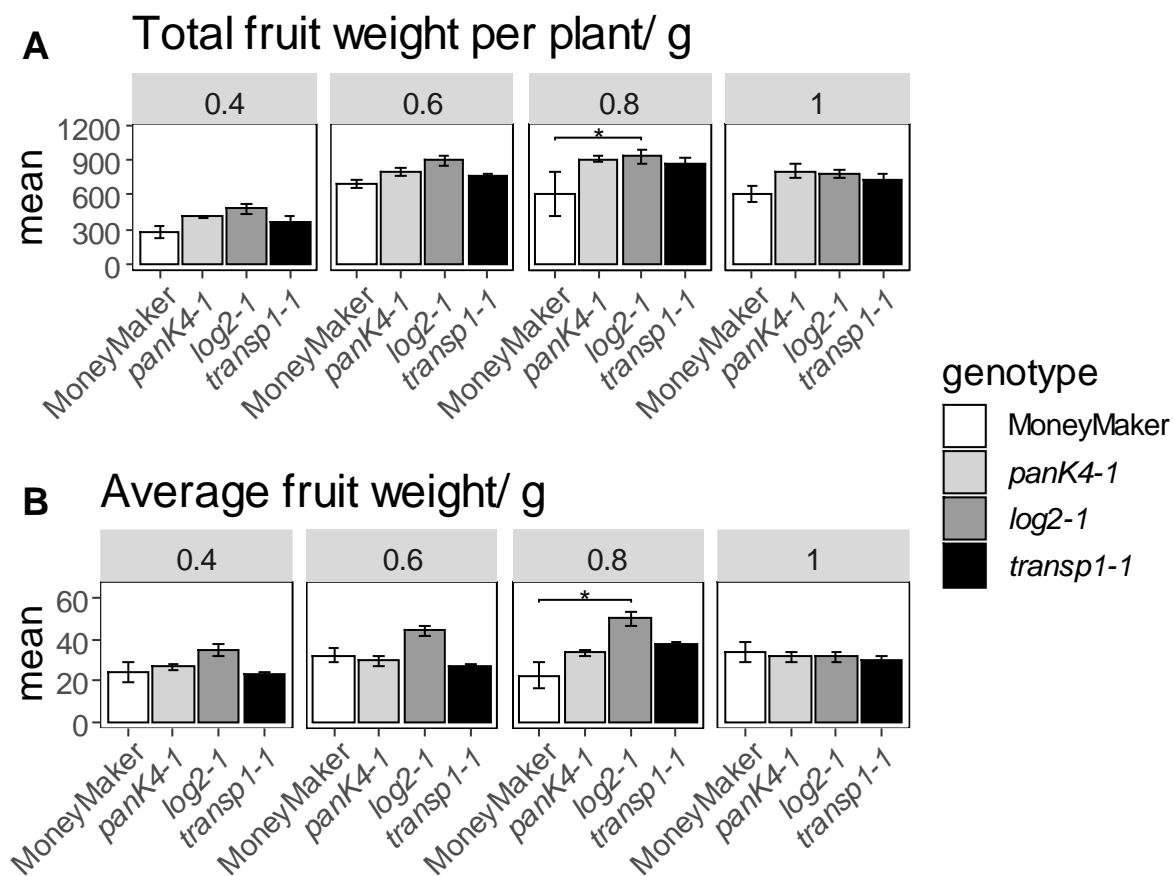
Figure 40: Whole plant traits under different irrigation conditions of MoneyMaker and mutant plants. A: Total fruit weight per plant; B: Average fruit weight; Panels show irrigation condtions from 0.4x to 1x of the standard optimal irrigation amount. Shown is the mean ± standard error; *: p-value ≤ 0.05 in TukeyHSD

By jackknifing conditions, I was also able to get values to perform statistical tests for the CV of different traits. One can see that the CV of average fruit weight is statistically reduced in *panK4-1* compared to MoneyMaker (Figure 41). Further on the CV of the harvest index is statistically reduced in all mutant lines in comparison to the wild type (Figure 41).
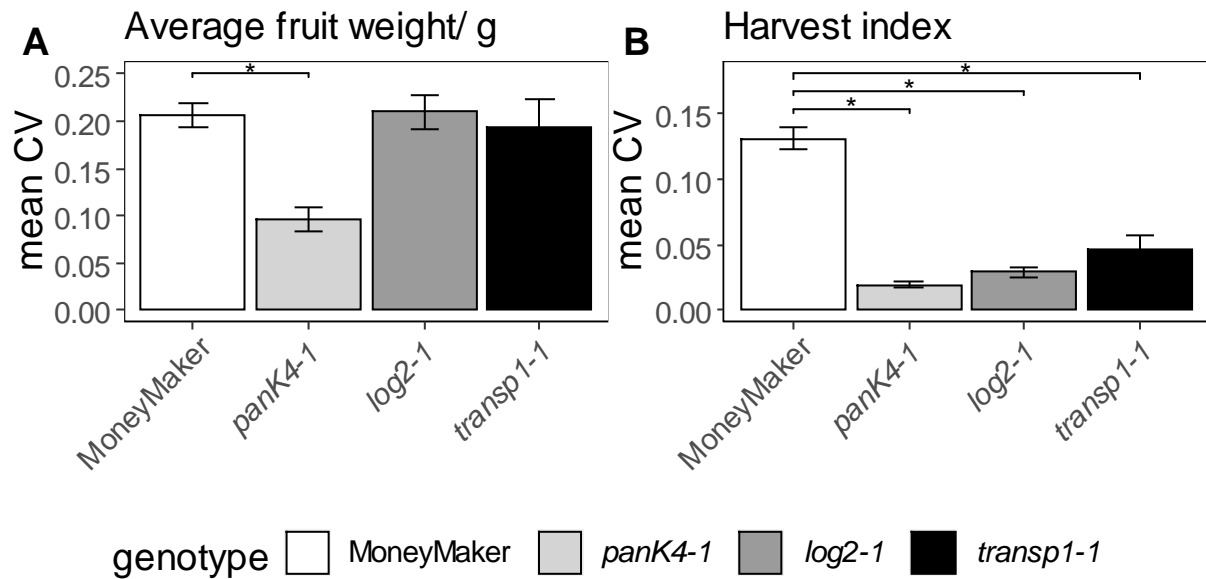
*Figure 41: Mean CV of whole plant traits calculated across different irrigation conditions of MoneyMaker and mutant plants. A: Average fruit weight; B: Harvest weight; Shown is the mean ± standard error; *: p-value ≤0.05 in TukeyHSD*

In summary I can say that the mutations seemed to have marked effects on several phenotypic traits.

3.3.5.2. Metabolomics

During the growth period I took leaf samples and after harvesting tomato fruits, I also collected samples from the pericarp. I extracted metabolites from both tissues and subjected them to metabolomics analysis.

3.3.5.2.1. Primary metabolism

I found a few metabolites, which show changes at individual conditions, both in leaves and fruits (Supplementary Figure 1). However, compared to the wild type, most changes are relatively mild (Figure 42).
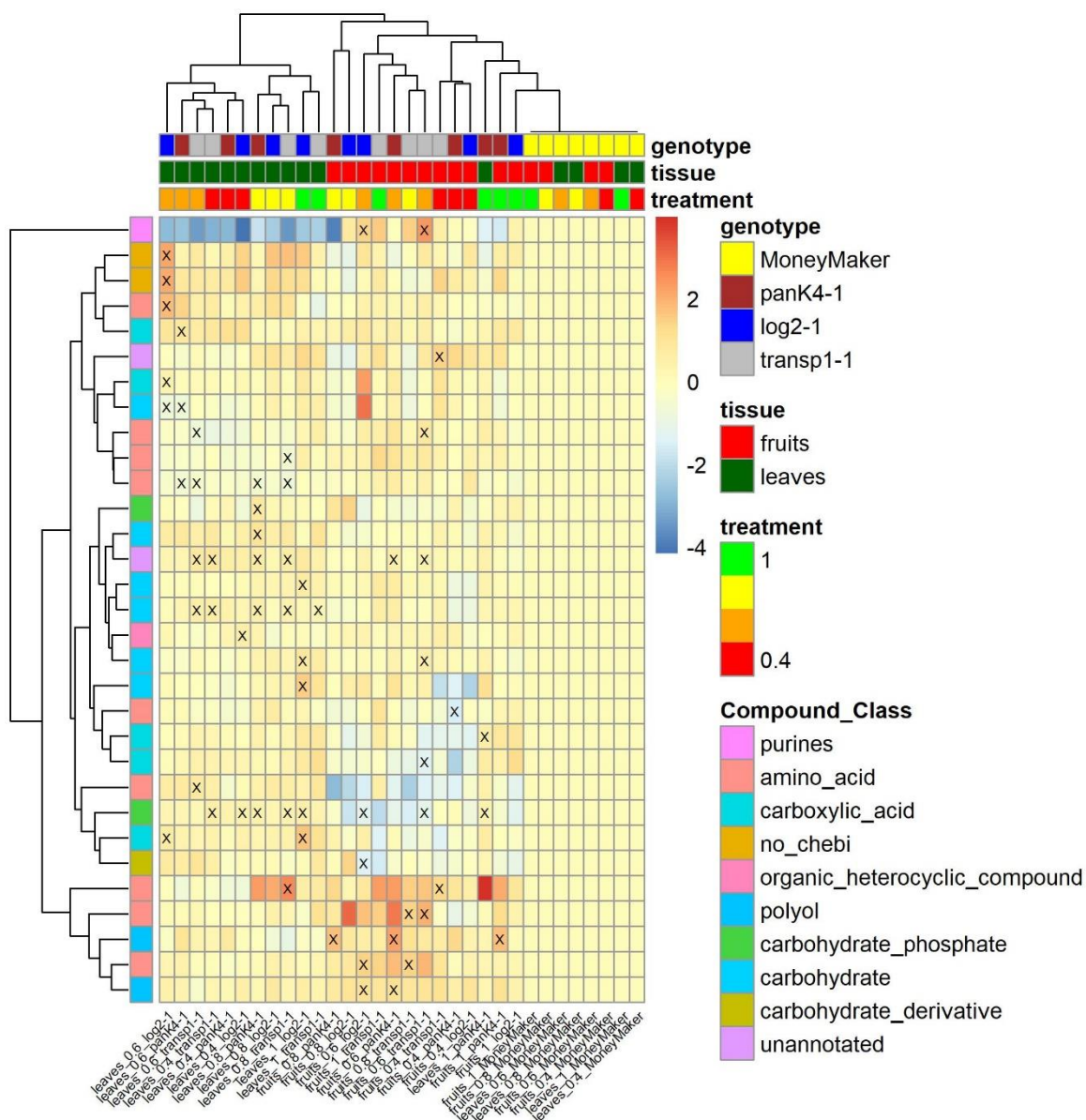
Figure 42: Heatmap of primary metabolites. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue and irrigation condition. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

I then calculated the CV of metabolite means across conditions for MoneyMaker to get an estimation of the general level of metabolite CVs (Figure 43).
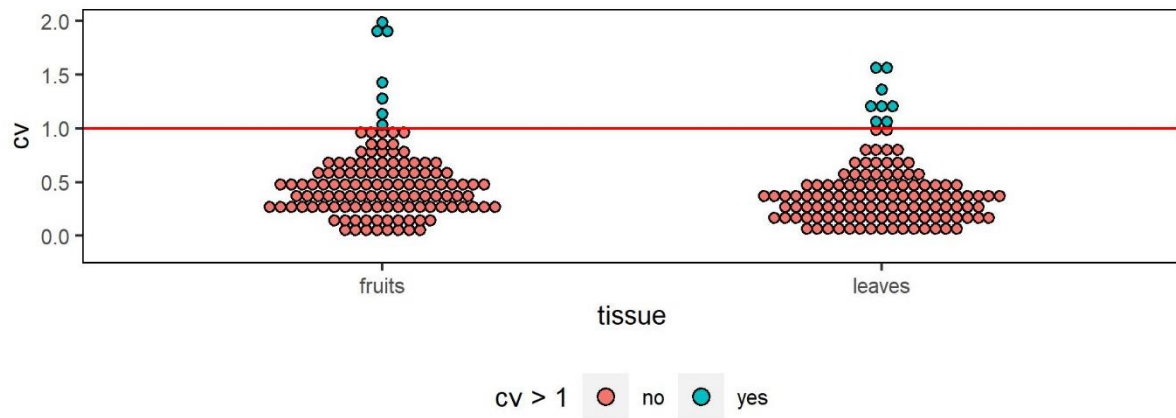
Figure 43: CV of metabolites across conditions for MoneyMaker samples in fruits and leaves. The red horizontal line denotes a CV of 1. CV is grouped into bins of 0.1

As one can see both in leaves and fruits the CV values are relatively low, with most values not surpassing a CV of 1. When jackknifing values across conditions and comparing wild type samples to mutant samples, I can see significant changes in a few more even than for the mean levels of the metabolites (Figure 44).
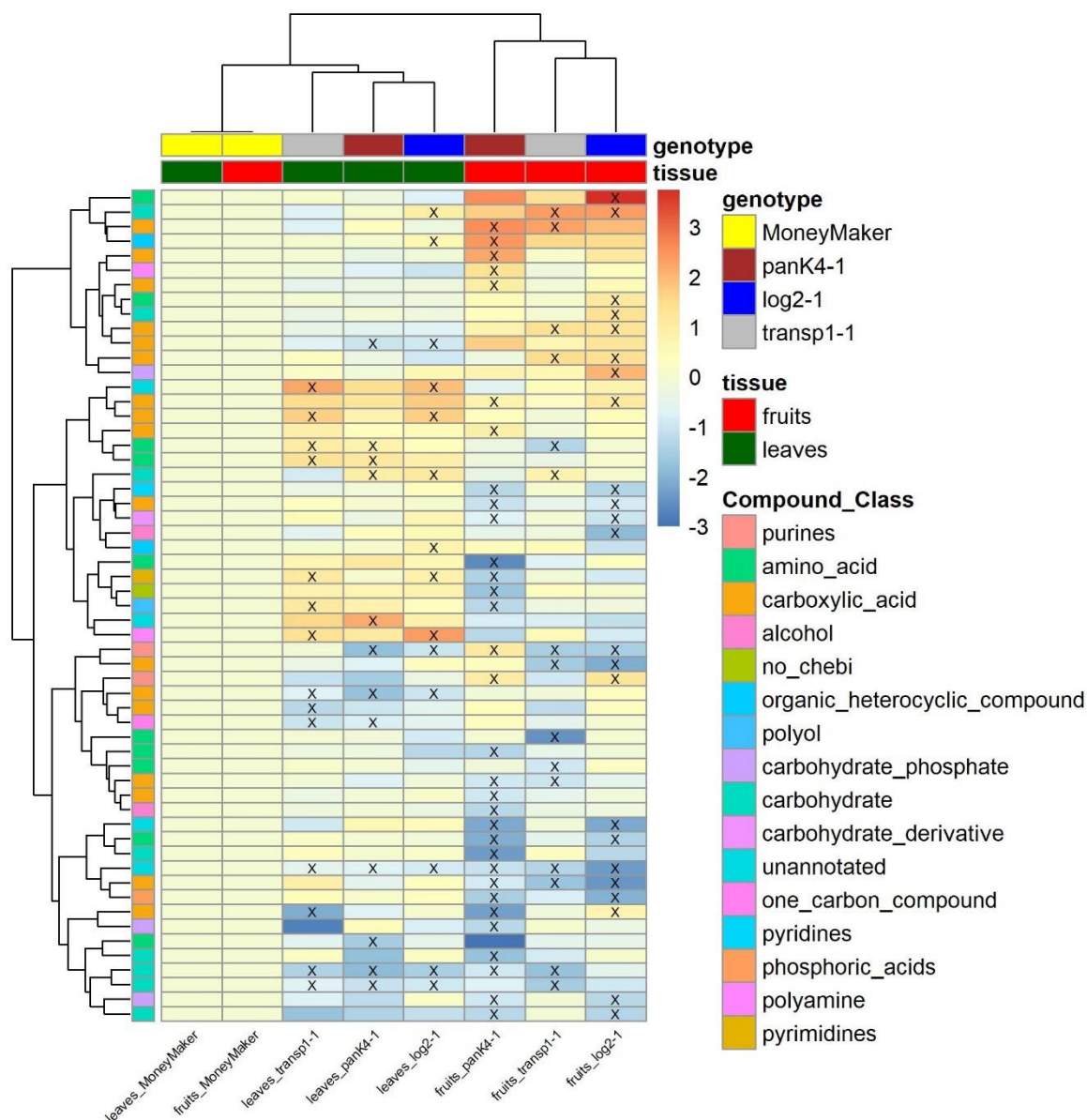
Figure 44: Heatmap of CV of primary metabolites across different conditions. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue and. crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

Although there are some changes in samples from leaves, most changes seem to be occurring in fruit samples. Further on, there seem to be more significant downregulations than upregulations (Figure 44).

Since I was primarily interested in the metabolites, for which cmQTLs were detected on the basis of which I created the gene-edited lines, I directed my focus towards these metabolites. The PanK4-1 candidate gene from the cmQTL of malic acid, was the target for one of our gene-edits, which is why I anticipated malic acid to be decanalized across environments. Although malic acid levels in tomato fruit of the *panK4-1* mutant compared to the wild type show no

significant changes under any irrigation condition, I can see that the CV of malic acid is significantly higher in the mutant than in the wild type (Figure 45).

Figure 45: Average malic acid level under different irrigation conditions (A) and CV of malate across different irrigation conditions of MoneyMaker and mutant plants (B). Panels in A show irrigation condtions from 0.4x to 1x of the standard optimal irrigation amount. Shown is the mean fold-change in relation to MoneyMaker mean at standard optimal conditions or mean CV of malate ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8

I can see that wild type levels of malic acid fluctuate somewhat across conditions, while in the *panK4-1* mutant levels of malic acids seem to decrease towards the lowest irrigation condition, which results in a roughly 2-fold increase of the CV. Meanwhile, CV of uracil, beta-alanine valine and pyruvate are significantly decreased (not shown).

The LOG2 gene was selected as a candidate gene for the canalization of phenylalanine and I therefore checked phenylalanine levels and cross-condition CV in the *log2-1* mutant compared to wild type samples (Figure 46).
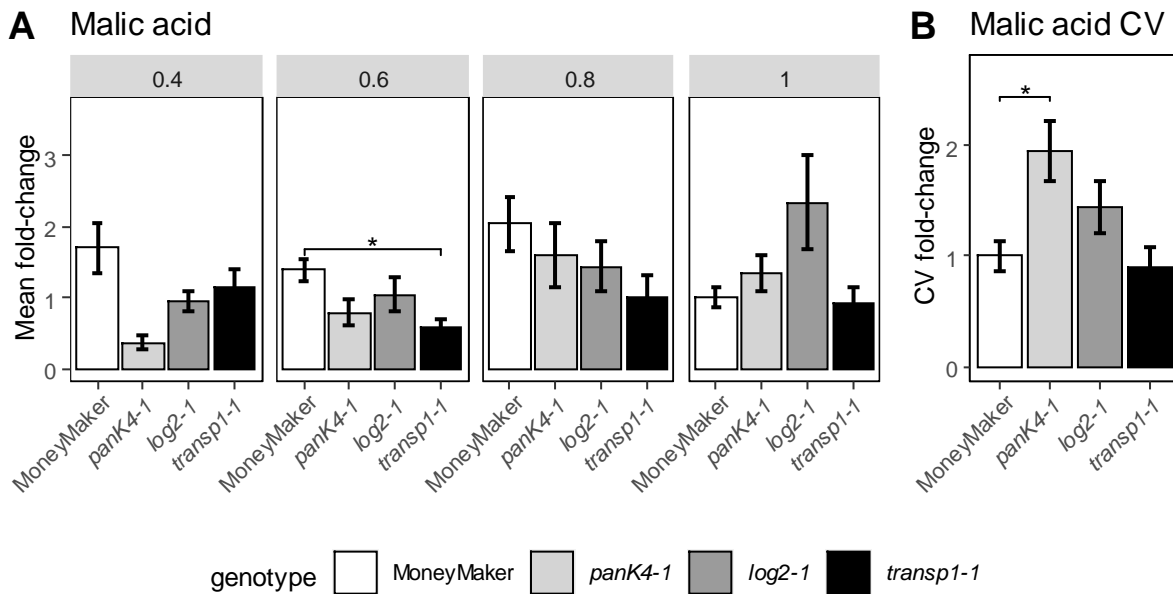
A Phenylalanine

B Phe CV
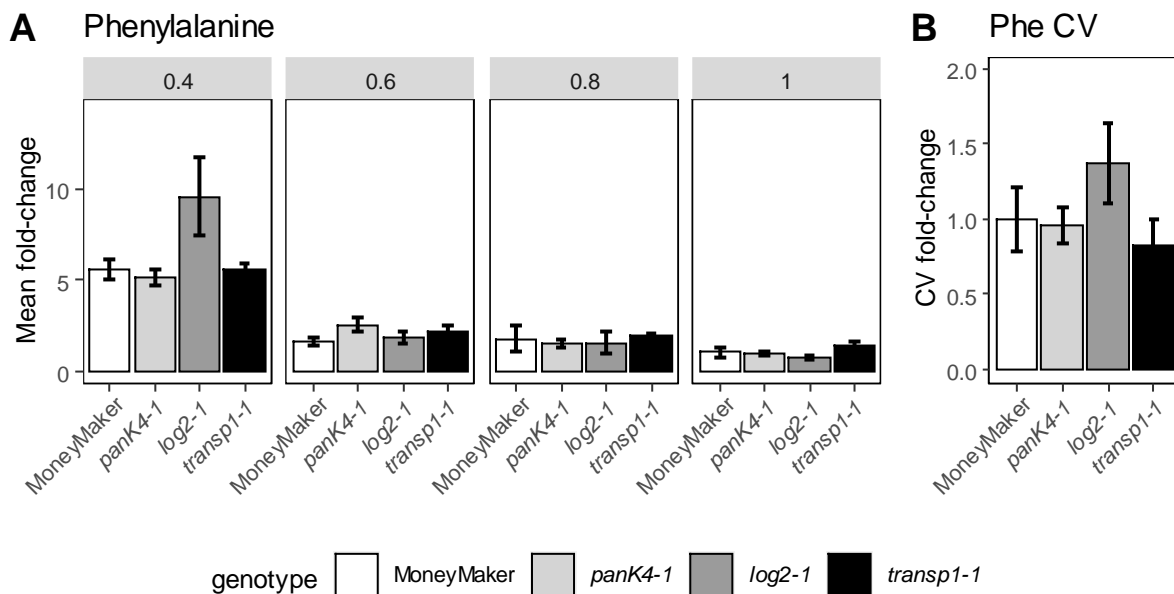
*Figure 46: Average phenylalanine level under different irrigation conditions (A) and CV of phenylalanine across different irrigation conditions of MoneyMaker and mutant plants (B). Panels in A show irrigation condtions from 0.4x to 1x of the standard optimal irrigation amount. Shown is the mean fold-change in relation to MoneyMaker mean at standard optimal conditions or mean CV of malate ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8, Phe: Phenylalanine*

Although I do see an increase in phenylalanine levels in pericarp of *log2-1* plants under the lowest irrigation conditions and therefore a higher CV, none of these changes are statistically significant (Figure 46). Similarly for the *transp1-1* line I see no changes in the CV of glucose-6-phosphate, fructose-6-phosphate or maltose (Figure 47).

A F6P CV **B** G6P CV **C** Maltose CV
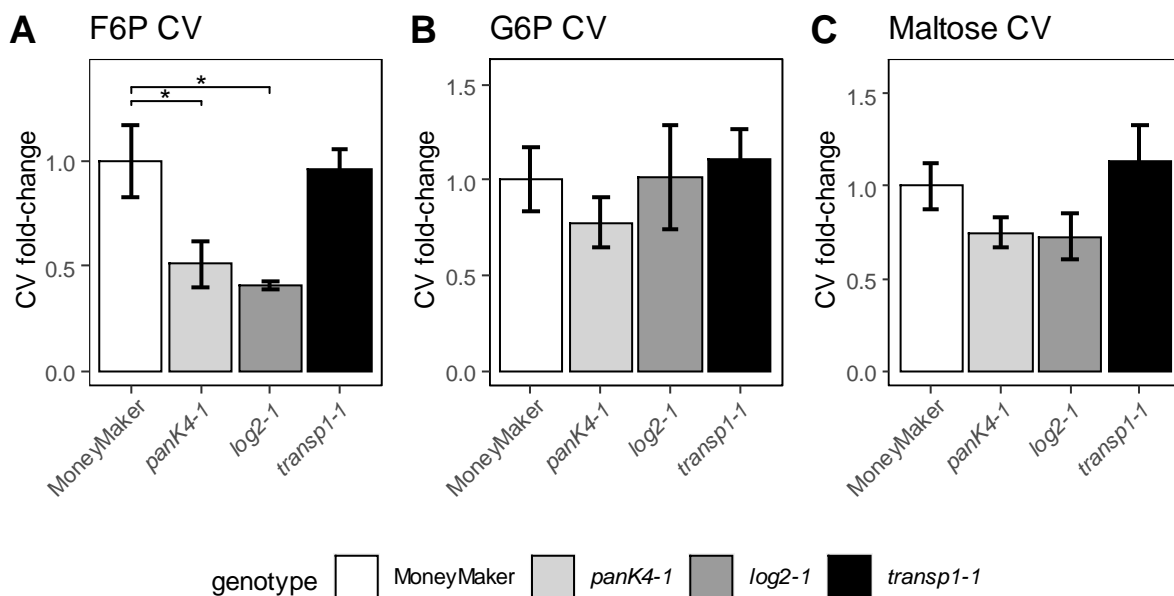
*Figure 47: Mean fold-change of CV of fructose-6-phosphate (A), glucose-6-phosphate (B) and maltose (C) across different irrigation conditions of mutant plants compared to MoneyMaker plants. Shown is the mean fold-change of the CV in relation to MoneyMaker CV ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8, F6P: fructose-6-phosphate; G6P: glucose-6-phosphate*

3.3.5.2.2. Secondary metabolism

Regarding the general level of secondary metabolites, I see most changes between gene-edited lines and wild type plants in the leaves (Supplementary Figure 2, Figure 48).
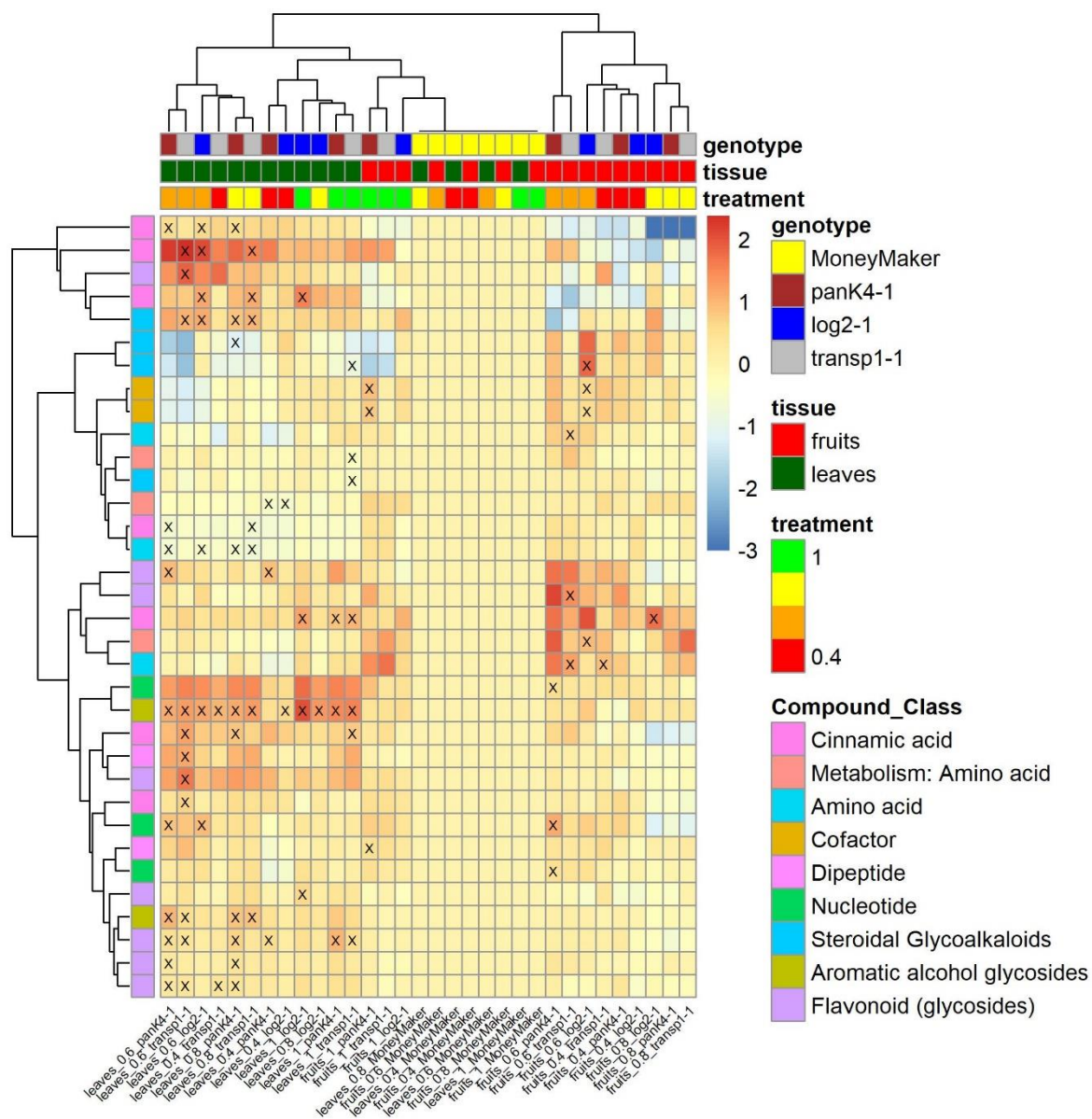
Figure 48: Heatmap of secondary metabolites. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue and irrigation condition. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

I also found that most changes are significant upregulations of metabolites compared to the wild type, rather than downregulations (Figure 48). Also here I calculated the CV of metabolites across conditions for MoneyMaker samples (Figure 49).
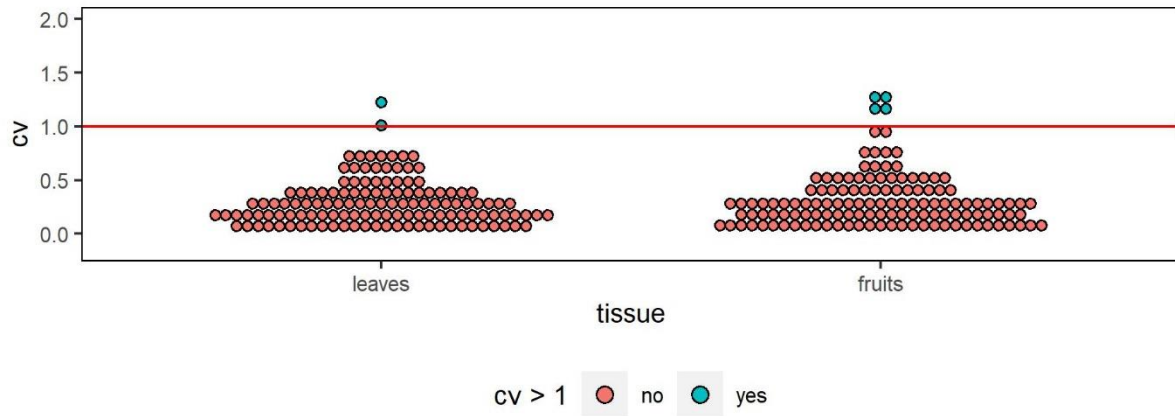
Figure 49: CV of secondary metabolites across conditions for MoneyMaker samples in fruits and leaves. The red horizontal line denotes a CV of 1. CV is grouped into bins of 0.1

As one can see, CV is generally low and only few metabolites have a CV higher than one (Figure 49).

When I look at the CV of these metabolites across the different watering conditions, however, I can see, significant changes in leaves and fruits and I see metabolites with increased CVs, but even more with decreased CVs (Figure 50).
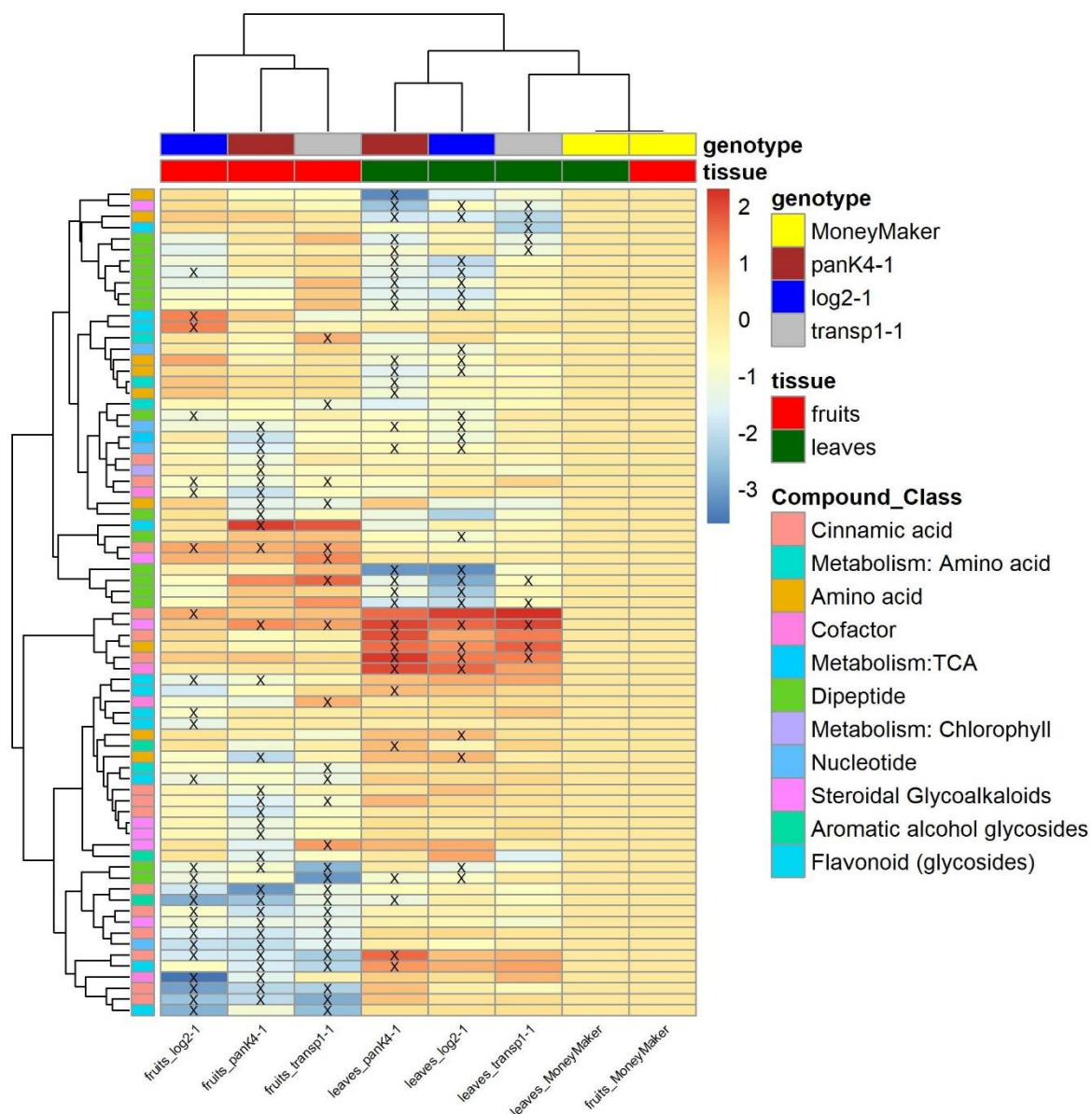
Figure 50: Heatmap of CV of secondary metabolites across different conditions. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue and. crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

In the leaves, I see a set of several dipeptides with significantly reduced CVs in most gene-edited lines compared to the wild type. The same metabolites show only minor changes in the fruits (Figure 50). Other changes in CV do not seem to be dominated by any compound class. Noteworthy, only one of the dipeptides shows significant changes between gene-edited lines and wild type plants for the metabolite level, as well as the CV (Figure 51).
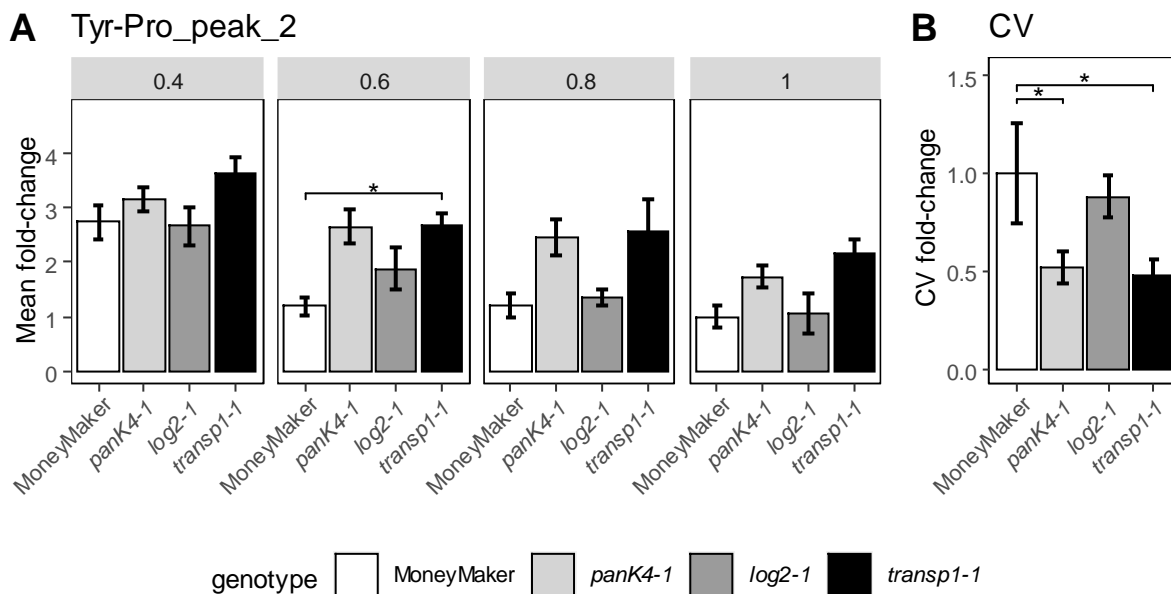
*Figure 51: Average Tyr-Pro dipeptide level under different irrigation conditions (A) and CV of Tyr-Pro dipeptide across different irrigation conditions of MoneyMaker and mutant plants (B). Panels in A show irrigation condtions from 0.4x to 1x of the standard optimal irrigation amount. Shown is the mean fold-change in relation to MoneyMaker mean at standard optimal conditions or mean CV of Tyr-Pro dipeptide ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8*

All other dipeptides only show significant changes in the CV (Figure 52).
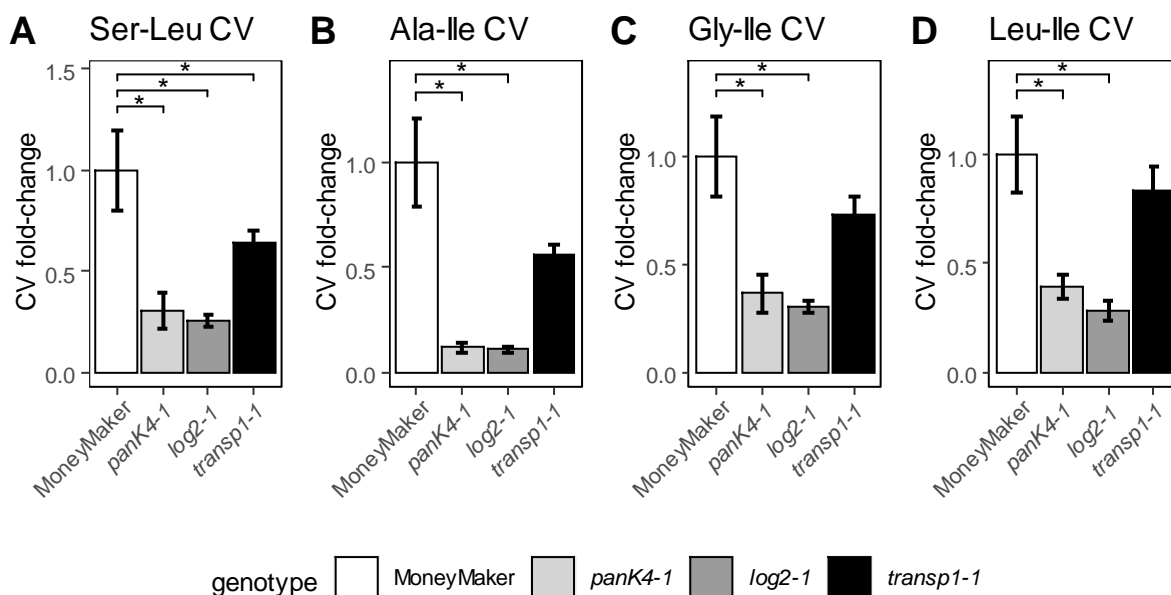
*Figure 52: Mean fold-change of CV of different dipeptides across different irrigation conditions of mutant plants compared to MoneyMaker plants. Shown is the mean fold-change of the CV in relation to MoneyMaker CV ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8*

The LC data give us another shot at estimating phenylalanine. Similarly as observed for the GC-MS data, phenylalanine levels increase towards the lowest irrigation condition in wild type samples but even more strongly in *log2-1* plants, leading to a slightly increased CV (Figure 53). However, the differences are not strong enough to be statistically significant.
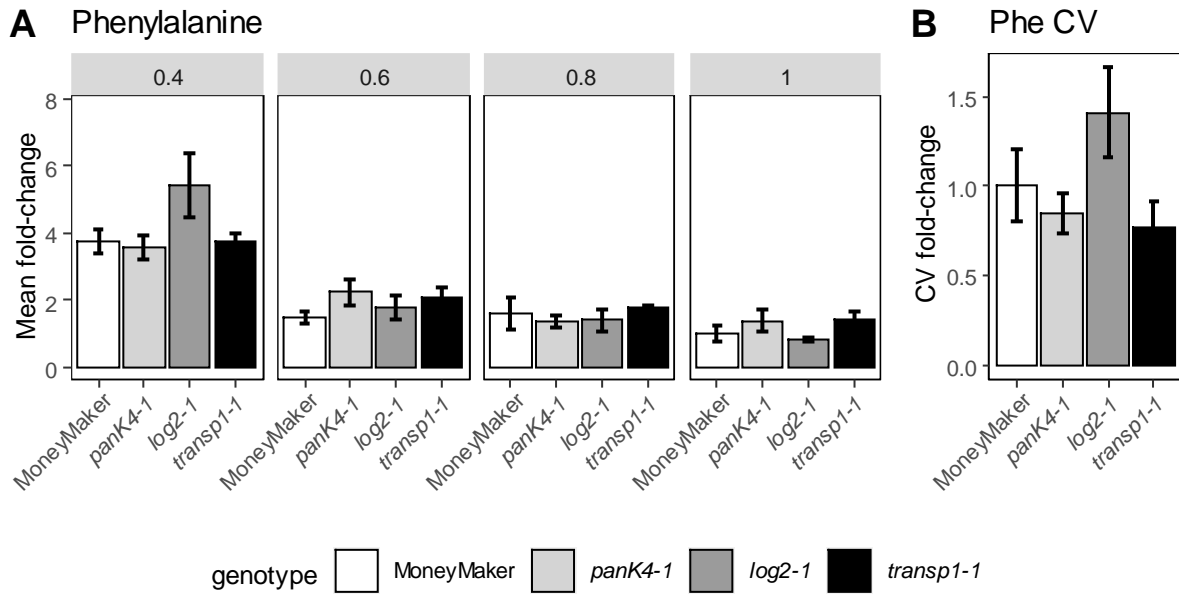
*Figure 53: Average phenylalanine level under different irrigation conditions (A) and CV of phenylalanine across different irrigation conditions of MoneyMaker and mutant plants (B). Panels in A show irrigation condtions from 0.4x to 1x of the standard optimal irrigation amount. Shown is the mean fold-change in relation to MoneyMaker mean at standard optimal conditions or mean CV of malate ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8*

3.3.5.2.3. Lipid metabolism

Again I see most significant changes of the mean level of metabolites in the leaves and most of these changes seem to be increases in the level of a metabolite (Supplementary Figure 3, Figure 54).
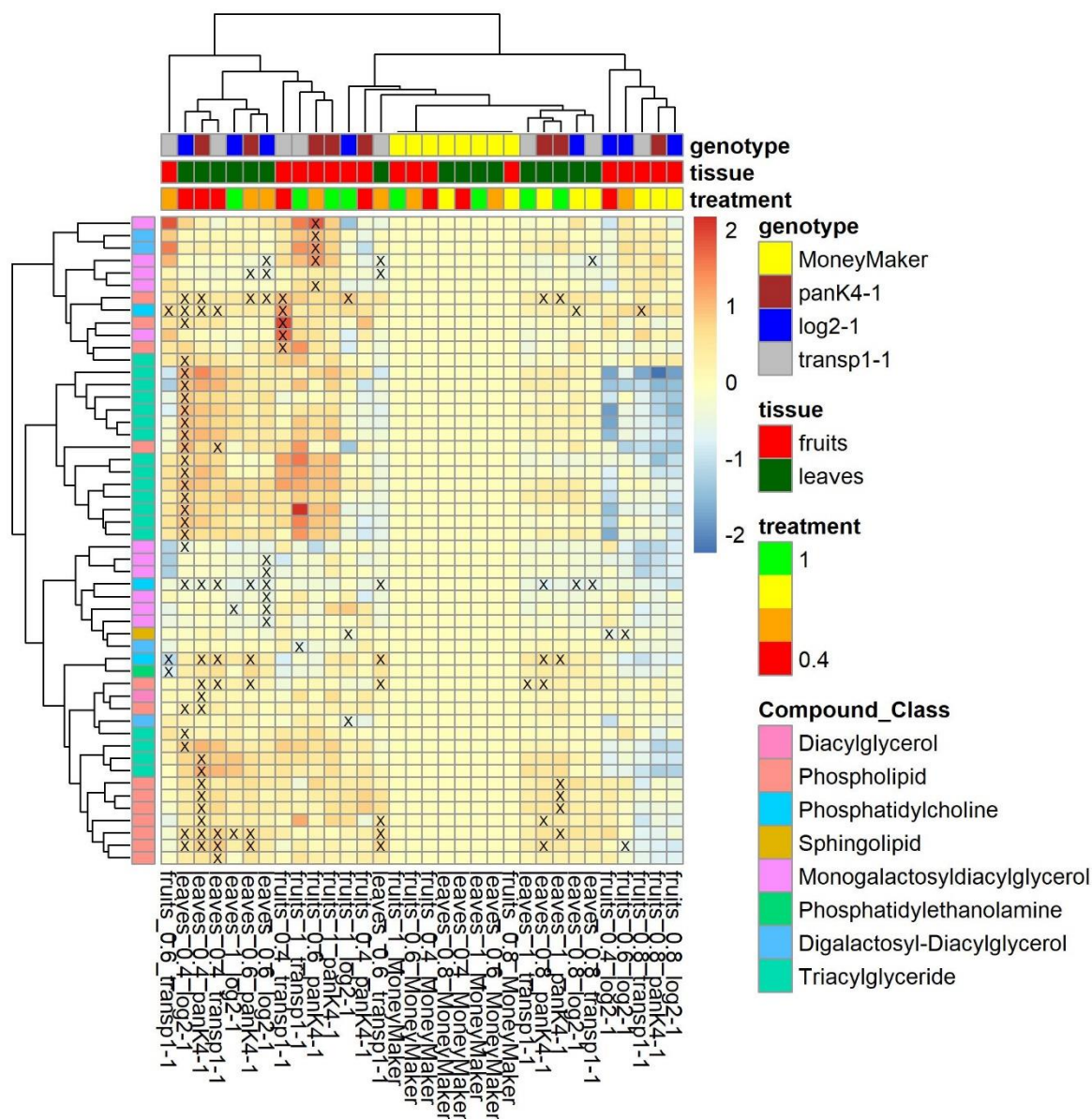
Figure 54: Heatmap of lipids. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue and irrigation condition. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

Another pattern that I observe is that most changes seem to occur under the more severe watering reduction conditions of 40% and 60% (Figure 54). Also a large set of triacylglycerides are significantly upregulated in leaves grown under the 40% watering condition of the *log2-1* line, compared to respective MoneyMaker samples (Figure 55).
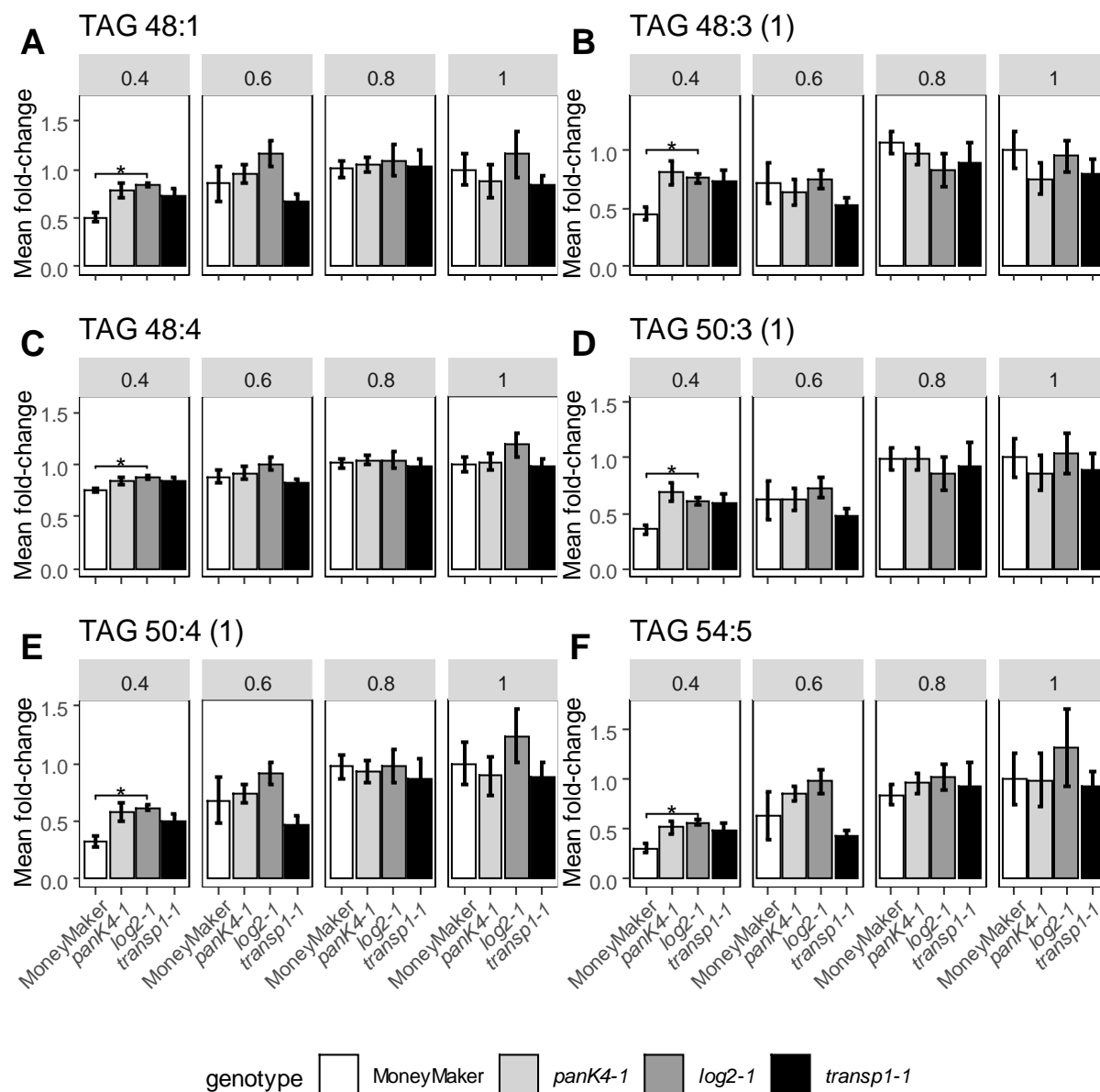
Figure 55: Average TAG levels in leaves of MoneyMaker plants and mutant lines under different irrigation conditions (A) TAG 48:1; (B) TAG 48:3; (C) TAG 48:4; (D) TAG 50:3; (E) TAG 50:4; (F) TAG 54:4 Shown is the mean fold-change in relation to MoneyMaker mean at standard optimal conditions ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8

Again I am calculating the CV lipid values for MoneyMaker samples across all conditions (Figure 56). Interestingly the CV is generally very low and neither leaves nor fruits have any metabolite with a CV higher than 1.
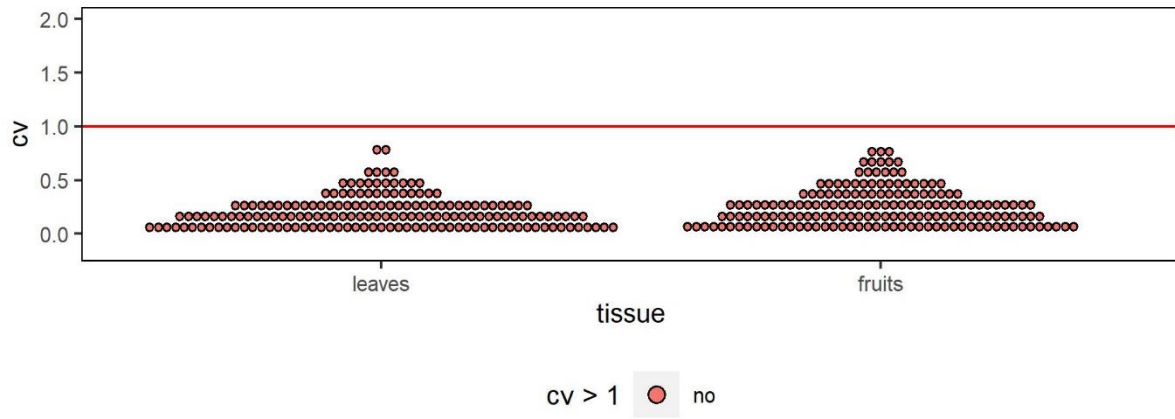
Figure 56: CV of lipids across conditions for MoneyMaker samples in fruits and leaves. The red horizontal line denotes a CV of 1. CV is grouped into bins of 0.1

When looking at the CV of lipids across the different conditions in mutant plants compared to wild type plants I can now see most changes in the fruits (Figure 57).
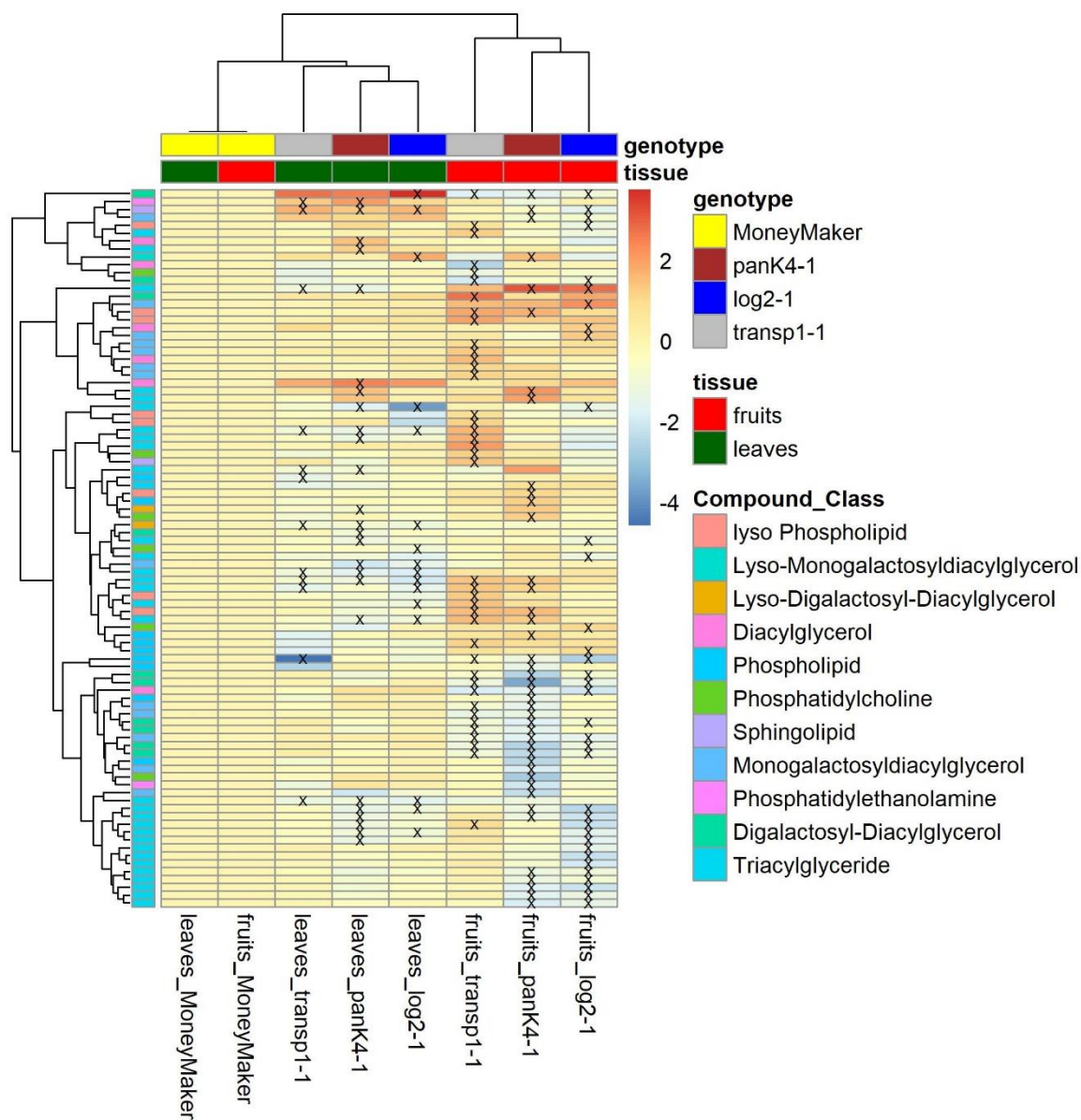
Figure 57: Heatmap of CV of lipids across different conditions. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue and. crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

Among the lipids with significantly changed CV are many TAGs (Figure 57). While most of them show a lower CV in gene-edited plants compared to the wild type, some also show a higher variation. In the leaves, there are four TAGs which show a reduced CV in all gene-edited lines compared to the wild type (Figure 58).
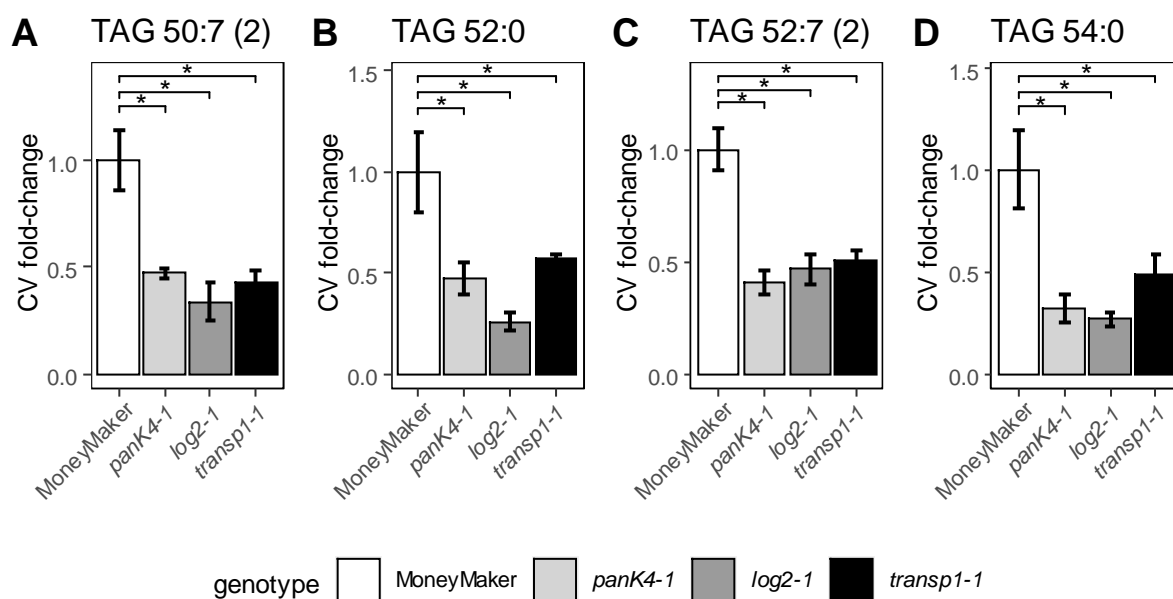
*Figure 58: Mean fold change of TAG CV in leaves of MoneyMaker plants and mutant lines across different irrigation conditions (A) TAG 50:7; (B) TAG 52:0; (C) TAG 52:7; (D) TAG 54:0. Shown is the mean fold-change of the CV in relation to MoneyMaker CV ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8*

In summary, I can say that the metabolic profile between wild type and mutant plants is often similar at normal conditions but significant differences can often be found at stronger drought stress conditions. While significant changes are generally scarce for primary metabolite, they appear both in leaves and fruits. For secondary metabolites and lipids I can find much more significant differences between wild type and mutant plants occurring in the leaves. Concerning the CV, I generally find much more significant differences than for the level of metabolites, for all metabolic classes.

3.3.5.3.   Standard conditions experiment

In the second experiment I only grew plants under optimal conditions, because some lines were still screened for homozygousity at the start of the experiment and sufficient samples of the same zygousity could not have been found for multiple conditions. Further on, it could not be predicted whether plants with a strong phenotype would be able to produce fruit under adverse conditions.

3.3.5.3.1.   Metabolomics normal conditions

3.3.5.3.1.1.   Primary metabolism

The profile of primary metabolites shows some strong changes between the mutant lines and wild type plants (Figure 59).
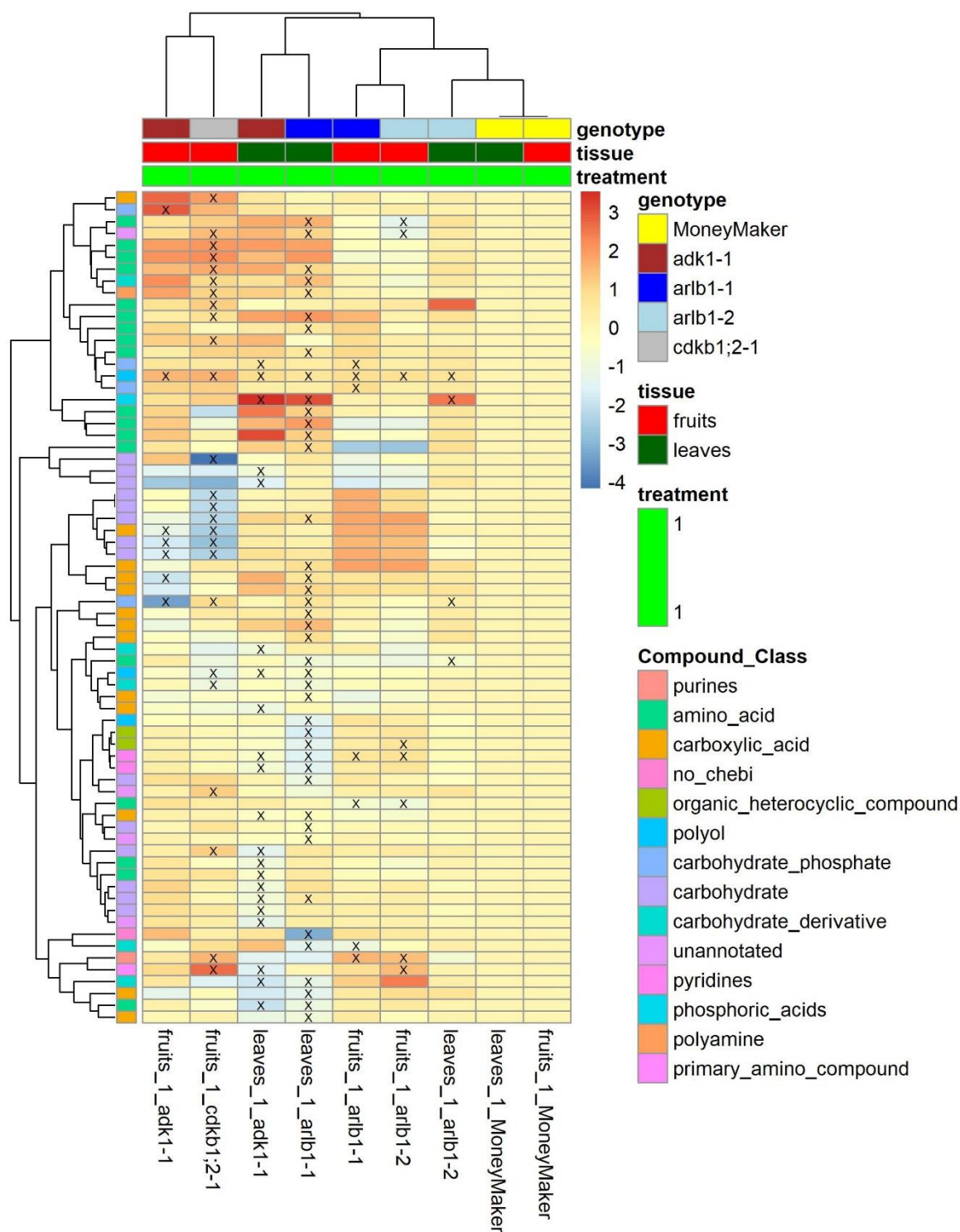
Figure 59: Heatmap of primary metabolites. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

The most significant changes can be found in the leaf samples of *arlb1-1* compared to wild type samples, mostly in amino acids and carboxylic acids (Figure 59). Fruit samples from *cdkb1;2-1* plants also have some decreased levels of different carbohydrates like glucose, fructose, sucrose and levoglucosan (Figure 60).
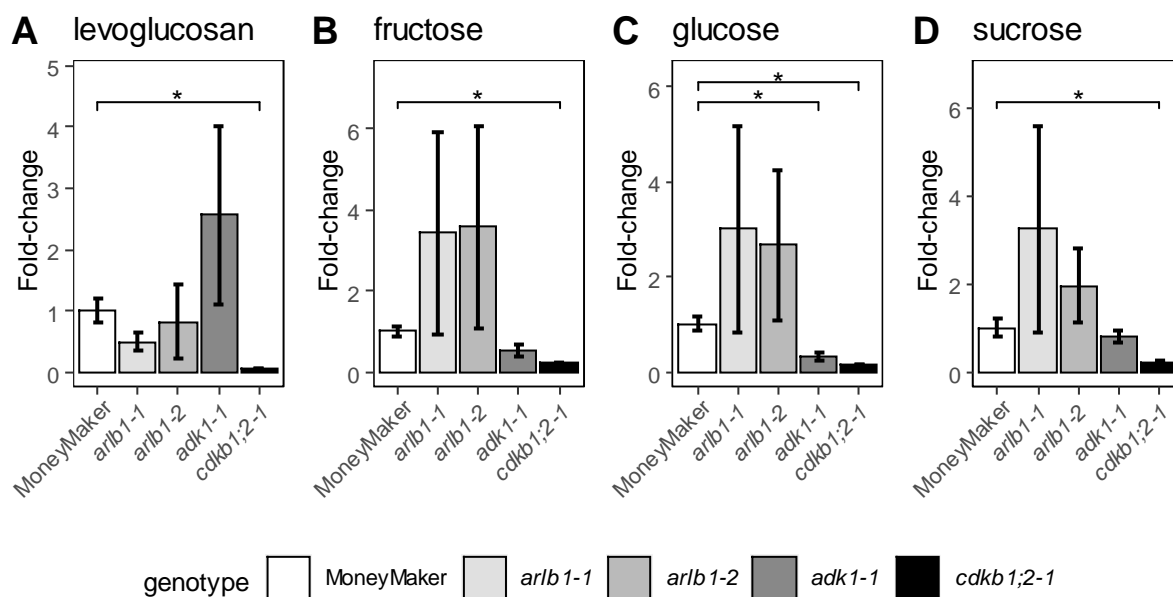
Figure 60: Average metabolite levels in fruits of MoneyMaker plants and mutant lines (A) levoglucosan; (B) fructose; (C) glucose; (D) sucrose. Shown is the mean fold-change in relation to MoneyMaker mean ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8

A different set of carbohydrates is downregulated in the leaves of *adk1-1* mutants compared to wild type plants (Figure 61).
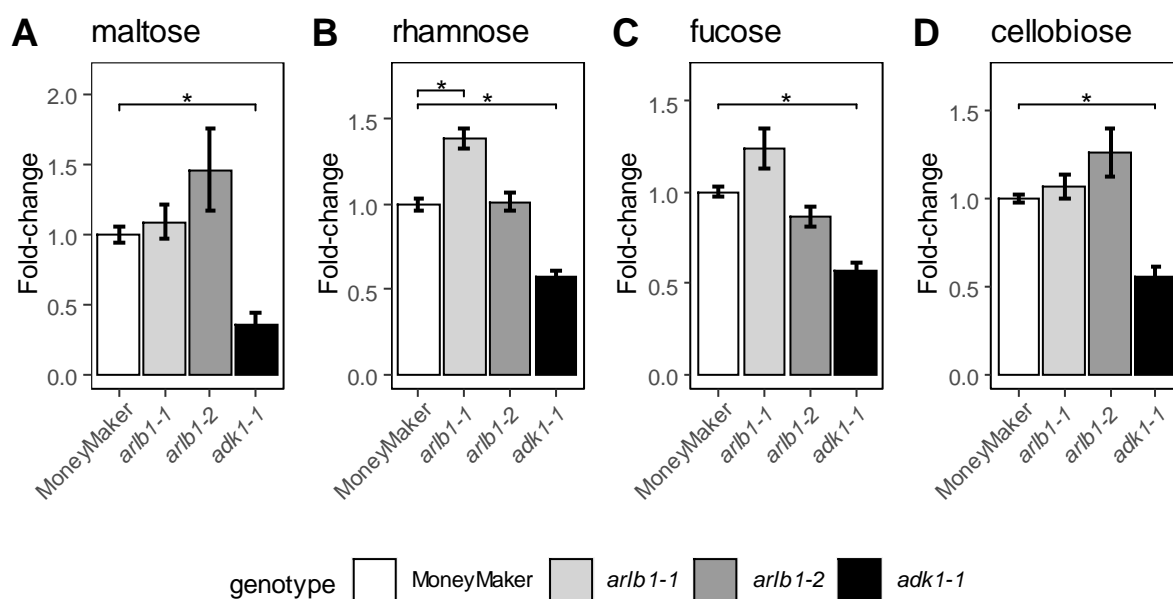
Figure 61: Average metabolite levels in leaves of MoneyMaker plants and mutant lines (A) maltose; (B) rhamnose; (C) fucose; (D) cellobiose. Shown is the mean fold-change in relation to MoneyMaker mean ± standard error; *: p-value ≤ 0.05 in two-sided student's t-test after bonferroni correction, n = 7-8

3.3.5.3.1.2.   Secondary metabolism

In the secondary metabolism, again I see the most changes in leaf samples of the *arlb1-1* line (Figure 62).
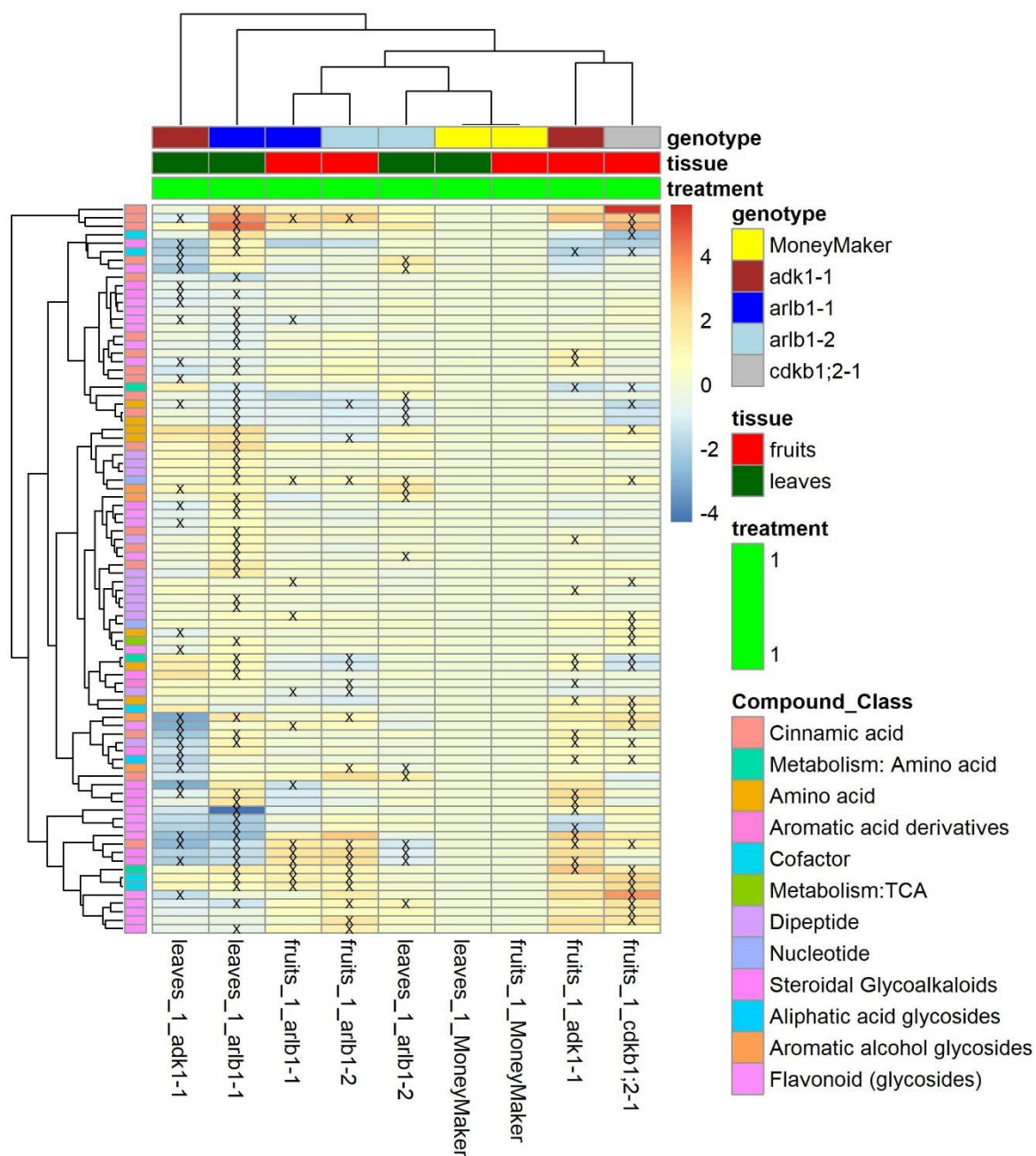
Figure 62: Heatmap of secondary metabolites. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

As some plants from the *arlb1-2* line, turned out to be heterozygous, it contains a mixture of samples from hetero- and homozygous plants. This probably explains, why leaf samples from *arlb1-2* plants show much fewer significant changes in the metabolic profile (Figure 62). Fruit samples from both lines however, generally show a more similar change in the metabolic profile (Figure 62). The line *adk1-1* shows several metabolites, which are strongly downregulated in the leaves. The *cdkb1;2-1* line for which I only had fruit samples shows some upregulated flavonoids in the fruits (Figure 62).

3.3.5.3.1.3. Lipid metabolism

In the lipid profile I can also see many changes in leaves of the *arlb1-1* line, but also of the *adk1-1* line (Figure 63).
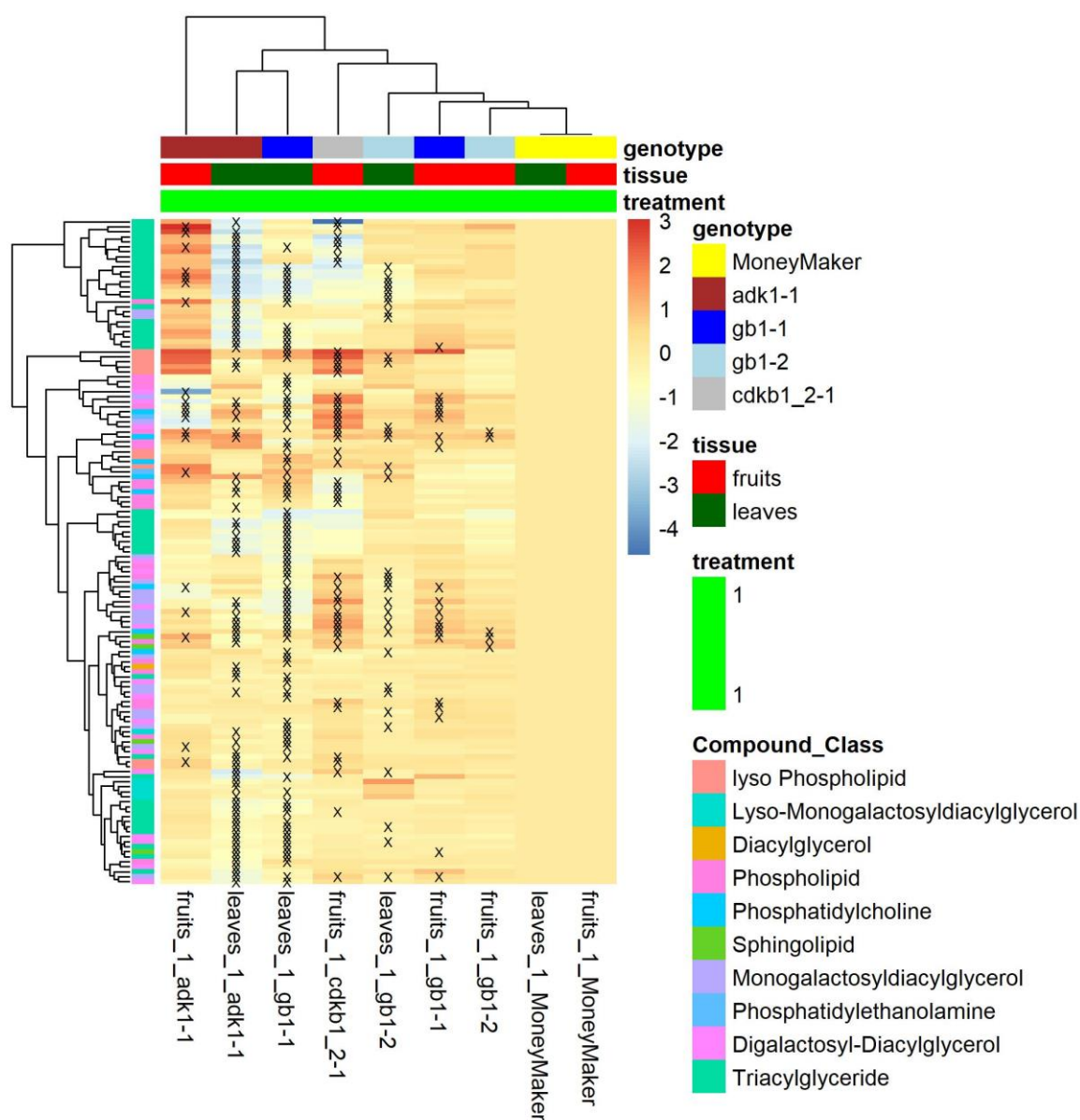
Figure 63: Heatmap of lipids. Values correspond to log2-fold changes in relation to wild-type samples of the same tissue. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a two-sided student's t-test, with trait-wise bonferroni correction; n = 7-8

Several TAGs, which show a significant downregulation in leaves of *adk1-1*, show the opposite response in fruits (Figure 63).

In summary of the metabolomics profiling I can say that I can see many changes in the metabolic profile of these mutant lines, which also showed strong morphological phenotypes, compared to wild type plants.

3.4. Discussion

In this chapter I selected putative candidate genes for which I was able to generate gene-edited lines via CRISPR/Cas9. Unfortunately the transformation and gene-editing efficiency was lower than what was reported in similar experiments (Brooks et al., 2014; S. Sun et al., 2015), but nonetheless I obtained at least one line per construct.

I documented any visible phenotypes, measured several whole plant traits and measured the metabolic profile of fruit and leaf samples to estimate the role of the respective genes in canalization of metabolism.

Several mutant lines did not show any obvious abnormal phenotypes. In the greenhouse experiments the mutant lines performed better in several phenotypic traits of agricultural importance. For example the total fruit weight per plant was significantly different for all mutant lines compared to wild type plants. The average fruit weight also was more robust across different watering conditions, in the *panK4-1* line compared to MoneyMaker plants.

For the metabolomics part, I again needed to find an appropriate normalization strategy to deal with batch variation and machine drift (Alseekh et al., 2018). Since the total sample number here was not too large, I was here able to incorporate many more QC samples, than I had available for the GWAS datasets. For the first experiment with different watering conditions, I always used two QCs at the start of every batch followed by four blocks of 12 samples and two QCs, amounting to a total of 10 QCs per 58 sample batch. This allowed me to use a batch-wise QC-based LOESS signal correction. LOESS signal corrections have successfully been used in literature to account for batch- and drift-variation (Dunn et al., 2011; Fan et al., 2019; Kyle et al., 2021). With a test samples to QC-ratio of roughly 5:1 this dataset, could also be suitable for normalization by a recently developed normalization method, based on the random forest machine learning algorithm (Fan et al., 2019), which may further reduce unwanted variation.

The putative pantothenate kinase (Solyc10g074590) was selected as a candidate gene responsible for the canalization of malate levels across different environments. Indeed, the mutant line *panK4-1* showed a significant increase of the CV of malate in fruit extracts across different watering conditions, which validates the role of the gene in canalization of malate levels. I can only speculate however how exactly this is achieved. The orthologous genes in *Arabidopsis thaliana* code for pantothenate kinases, which are involved in coenzyme A biosynthesis (Tilton et al., 2006). As CoA is an important cofactor which is needed for many metabolic reactions, especially in the TCA cycle, this may be a reason why PanK4 is important

for canalization of malate (Coxon et al., 2005). In plants pantothenate is synthesized from beta-alanine, which can in turn be synthesized from uracil, but also needs contribution from a pathway starting with valine, although a clear enzyme converting keto-pantoate to pantoate has not yet been found in plants (Ottenhof et al., 2004). Interestingly the CV for uracil, valine and beta-alanine were all decreased in *panK4-1* plants, which may indicate a more controlled flux towards pantothenate, to compensate for a reduction in the downstream processing towards coenzyme A. In fact pantothenate levels were slightly lower under all conditions in fruits of the *panK4-1* compared to wild type plants, but not statistically significant.

The gene Solyc10g086080 was chosen as a candidate for canalization of phenylalanine levels. An orthologous gene in *Arabidopsis thaliana* LOG2 was characterized as E3 ubiquitin ligase, interacting with GDU1 to regulate amino acid export from plant cells (Pratelli et al., 2012). Plants overexpressing GDU1 show a selective increase of glutamine in hydathode secretion (Pilot et al., 2004). It may be possible that LOG2 has additional interaction partners, which could be relevant for transport of phenylalanine. Both GC-MS and LC-MS indicated an increase of the mean level of phenylalanine in fruits of *log2-1* plants at the lowest irrigation condition, but neither the mean level nor the CV showed statistically significant differences when compared to wild type samples. It may be possible that either even the most drastic drought condition, was not strong enough or that orthologues may compensate for the reduced activity in the *log2-1* plant. LOG2 is also known under the name ABA-insensitive RING protein 3 (AIRP3) and has been suggested to play a dual role in the ABA-mediated drought stress response and amino acid transport (Kim & Kim, 2013).

A gene annotated to code for a transposon protein (Solyc10g007190) was selected as candidate gene for the canalization of maltose, glucose-6-phosphate and fructose-6-phosphate. The *transp1-1* mutant of this gene did however not show any changes in level or CV of these metabolites and is therefore likely not related in the canalization of them. While metabolism seems not strongly affected in *transp1-1* plants, the days until the first ripe fruit was slightly reduced under all conditions and the ANOVA showed a significant effect. Anecdotally, I also noticed that all later fruits on *transp1-1* plants ripened earlier and more uniformly than fruit on wild type plants, similar to a determinate tomato cultivar. Transposable element insertion polymorphisms (TIPs) in tomato have been shown to be disproportionately located within or adjacent to genes with environmental responses and a GWAS has shown a robust association of these TIPs with variation in major agronomic traits and secondary metabolites (Domínguez et al., 2020).

Aside from these lines which had a wild type-like appearance, some lines also showed strong abnormal phenotypes.

As a second candidate gene for phenylalanine canalization I had chosen Solyc10g086190, annotated as an adenosine kinase. The *adk1-1* line only had few mostly seedless fruits. In *Arabidopsis thaliana*, silencing of orthologous genes also resulted in abnormal growth phenotypes like altered silique size and a changed profile of cytokinin derivates, which is why ADK is suggested to contribute to cytokinin homeostasis (Schoor et al., 2011). The metabolic profile of both fruits and leaves of *adk1-1* plants was strongly altered, which may be a result of the general growth phenotype.

For malate I selected two more candidate genes (Solyc10g074630 and Solyc10g074720). Solyc10g074630 has been shown to be homologous to the ARF-like GTPase AT5G52210 (ARLB) (Paul et al., 2014). ADP-ribosylation factors (ARFs) have known roles in vesicle coating and uncoating in eukaryotes, while related proteins, ARF-like proteins (ARLs), have various functions (Gebbie et al., 2005). Homozygous plants from *arlb1-1* and *arlb1-2* lines showed small outgrowths at leaflet sections as well as scars on the cuticle. Heterozygous plants from *arlb1-2* showed much smaller but still visible outgrowths and wild type-like fruit. The outgrowth phenotype may be comparable to reduced apical dominance, which has been observed in some antisense lines of ARF genes in *Arabidopsis thaliana* (Gebbie et al., 2005). I also saw a strongly altered metabolic profile of *arlb1-1* plants. Especially TAG contents was generally reduced in leaf of mutant plants, while some were increased in fruits.

Solyc10g074720 is annotated as a cell division protein kinase and is orthologous to cyclin-dependent kinase b 1;2. In *Arabidopsis* CDKB1s seem to be crucial for guard mother cell division (J. Yin et al., 2019). The *cdkb1;2* mutant line displayed light-yellow-green leaves. Although no such phenotype seems to have been observed before in *Arabidopsis thaliana*, an unpublished study found the gene differentially expressed in another chlorotic leaf mutant (Gu et al., 2020). In the metabolic profile of *cdkb1;2-1* fruits I could see a decrease of several carbohydrates, like glucose and sucrose, while several amino acids were increased.

For the first set of mutants which appear wild-type like by their general appearance I could at least confirm one candidate gene with a less environmentally robust CV of the target metabolite. The *panK4-1* showed a significantly increased CV of malate across different irrigation conditions. The *log2-1* mutant shows some tendency towards an increased

phenylalanine CV but the results are not statistically significant. The *transp1-1* mutant shows no clear changes of maltose, glucose-6-phosphate or fructose-6-phosphate.

For the second set of mutants with clear morphological abnormalities I do see some differences in the metabolic profile of these lines, but these results are only first estimations. Some lines only produced a few plants and I therefore used technical replicates to have enough values to estimate proper mean values and perform statistical tests. In some cases I also included samples which were originally grown in a different experiment. In the case of the *arlb1-2* line I did not obtain enough homozygous plants and therefore also included heterozygous plants for the metabolic profiling. Given the sometimes drastic phenotypes and the large difference of the estimated values it is however likely that follow-up experiments will confirm the observations I made here. I can of course at this point also not be sure, whether the canalization of the target metabolites across different environments is also affected. It remains to be seen, whether I will at all be able to cultivate these mutants under less than optimal conditions, given the relative difficulty just growing them at optimal conditions.

Applying the appropriate stress conditions to make differences in robustness visible generally proved to be quite challenging. In consultation with a greenhouse staff member, experienced in tomato cultivation, I decided to not lower the irrigation conditions below 40% of the standard optimal conditions. Indeed plants only produced few small fruits at this condition and may likely not have fruited at all under conditions with even lower irrigation. However, as I see the most drastic changes of the metabolic profile at low irrigation conditions and because an increase of irrigation beyond 80% of the standard optimal conditions did not result in any further performance increase it may be worthwhile to consider generally lower watering conditions for future experiments, to get a better resolution of the more drastic effects under low water availability. Another possibility may be to grow all plants under normal conditions until fruit set and then apply even more drastic drought scenarios.

Of course I will try to test any untested promising mutant lines for metabolic canalization. For further investigations it would also be interesting to introduce gene-edits in orthologues of genes, where a modification showed a metabolic or morphological phenotype or a tendency towards one. Creating double mutants, may enable to detect changes in metabolic canalization, which remain obscured in single mutants, although the time consumption would be considerable in tomato.

Chapter 4: CANA1

4.1. Introduction

Through oxygenic photosynthesis, plants are able to capture sunlight, convert it into chemical energy in the form of ATP and NADPH, which can consecutively be used to assimilate carbon dioxide into organic matter (Baslam et al., 2020; Nelson & Ben-Shem, 2004). As mentioned earlier, I can consider this photosynthetic carbon metabolism as one of the most important biochemical processes on the planet, given that its products ultimately support almost all life forms (Y. Wang et al., 2015). In order to fuel the carbon dioxide assimilating Calvin Cycle, ATP and NADPH must first be generated in the photosynthetic electron transport chain (Baslam et al., 2020).

The photosynthetic electron transport chain is constituted by four major protein complexes, photosystem I and II, the cytochrome-b_6f complex and F-ATPase, embedded into the thylakoid membranes of chloroplasts (Nelson & Ben-Shem, 2004). To summarize this highly complex reaction, I can say that captured solar energy is transferred onto reaction centers of photosystem I and II, which induces a cyclic or linear electron flow over several electron carriers to $NADP^+$ and the concomitant transmembrane proton transport, which generates a proton motive force, driving ATP production (Nelson & Ben-Shem, 2004; Tang & Blankenship, 2013).

Due to their complex multi-protein composition, photosystem I and II need to be accurately assembled through a finely controlled mechanism (Lu, 2016; Nellaepalli et al., 2018). Although already more than 30 factors are for example already known to be relevant for the assembly of photosystem II, it is likely that more will be found and the exact function of the discovered ones remains mostly unclear (Eaton-Rye & Sobotka, 2017). The importance of such assembly factors for functioning photosystems can easily be demonstrated by the fact, that some of them were originally discovered by the visual phenotypes like pale or variegated leaves of mutants lacking these factors (Moore et al., 2000; Shimada et al., 2007; Sundberg et al., 1997; Zagari et al., 2017).

A gene, which, if impaired, leads to pale cotyledons in *Arabidopsis thaliana* and variegated leaves in *Lotus japonicus* is the so-called "snowy cotyledons 2" SCO2 (Zagari et al., 2017). SCO2 codes for a protein disulfide isomerase, carrying a conserved C_4-like zinc finger domain, similar to *E. coli* DnaJ (Shimada et al., 2007). *In vitro* studies have shown that this domain is critical for its catalytic activity and that the protein uses glutathione as an electron donor (Muranaka et al., 2012). Further on, SCO has been shown to interact with LHCB1 proteins and

likely also associates to other thylakoid proteins (Shimada et al., 2007; Tanz et al., 2012). Mutants of SCO2 also form chloroplasts with abnormal thylakoids (Shimada et al., 2007). Due to these observations it has been concluded that SCO2 is a factor, important for thylakoid biogenesis and chloroplast development (Albrecht et al., 2008; Tanz et al., 2012).

In the following chapter I am describing the characterization of a variegated tomato mutant line by a multi-omics approach. The line was created by an EMS screening for yield stability, by our collaborator and was found to harbor a point mutation in an orthologous gene of SCO2 (Fisher & Zamir; under revision) . In field trials the mutant displayed an increased yield variability in relation to drought stress. I characterized the mutant with transcriptomics, proteomics and metabolomics and are here presenting our findings.

4.2. Materials and Methods

4.2.1. Genotyping/Sequencing

After plant transformation, shoots from well-rooted putatively transformed tomato plants are cut and transferred to fresh sterile media containing containing agar (6.8% m/v), MS, sucrose (2% m/v) and 125 µg/ml Ticarcillin, to eliminate residual *Agrobacterium tumefaciens*. Plants are cultivated for 4-6 weeks under sterile conditions, until they have rooted again and grown back to the original size. This process is repeated at least 2 more times. At the last transfer step, leaf samples are collected and DNA is extracted, similarly as previously reported (Lu, 2011). A first PCR amplifying the chromosomal rpoB gene of *Agrobacterium tumefaciens* GV2260, to check for residual agrobacterial contamination. A second PCR is performed with primers specific to the Cas9 gene. The third PCR amplifies a region ranging from ~100 bp upstream of the first gRNA to ~100 bp downstream of the second gRNA. PCR fragments are separated on an agarose gel (4% m/v in TAE), to detect large or small indels. If bands from putative genome-edited plants could not be distinguished from bands of wild type plants, PCR fragments were subjected to a heteroduplex mobility assay, as described previously (Bhattacharya & Meir, 2019). If bands were still indiscernible, PCR fragments were send for Sanger sequencing to LGC Genomics (LGC Genomics GmbH, Berlin, Germany). Agrobacterium-free, Cas9-positive, genome-edited plants were transferred to the greenhouse for propagation and investigation.

4.2.2. Growth conditions

Putative transformants were kept under sterile conditions in a tissue culture chamber (York International/Johnson Controls; Cork, Ireland) with ~35 µmol m^{-2} s^{-1} of light in a (16h/8h)-

(22°C/22°C)-(day/night) cycle. Selected transgenic plants were propagated under standard greenhouse conditions for seed production.

Seeds were generally first sown on soil in a young plant phytotron (York International/Johnson Controls; Cork, Ireland) with 150 µmol m^{-2} s^{-1} of light in a (16h/8h)-(22°C/18°C)-(70% RH/70%RH)-(day/night) cycle and transplanted into individual pots after cotyledon were fully expanded. After 4 weeks, seedlings were transplanted into bigger pots and transferred to greenhouse chambers. Plants were fertilized once after transplanting to big pots and once at flowering. Axillary meristems were continuously removed and the primary shoots were trimmed off after development of the 4th truss. During growth, pots are watered multiple times daily by an automated drip irrigation system. Standard greenhouse settings are a (16h/8h)-(22°C/20°C)-(50% RH/50%RH)-(day/night) cycle.

For drought conditions plants received 50% of the standard optimal water amount.

4.2.3. Sample collection

Leaf samples were collected from mature non-senescing plants and pericarp samples were collected from red ripe tomato fruits, when 80%-100% of fruits were ripe.

4.2.4. Metabolite extraction

Frozen plant tissue was ground to powder using a Mixer Mill (Retsch; Haan, Germany) at 30 Hz for 1 min. Aliquots of 50 mg were used for metabolite extraction, as described before (Salem et al., 2016). Here, 1 ml of pre-cooled MTBE-MeOH (3:1; vol/vol) extraction buffer is added to each sample and samples are incubated 10 minutes on an orbital shaker at 4°C. Samples are sonicated for 10 minutes in an ice bath before 0.5 ml of MeOH-H$_2$O (3:1; vol/vol) is added and samples are centrifuged at 14000 rpm for 5 minutes at 4°C, leading to a phase separation. An aliquot of the upper (apolar) phase was taken for lipid analysis, the rest of the upper phase was aspirated with the BVC fluid aspiration system (Vacuubrand Inc; Essex, CT, U.S.A.) and two aliquots of the lower (polar) phase were taken for GC-MS and LC-MS. Extracts were dried in a Scan Speed 40 centrifugal vacuum concentrator (Labogene; Allerød, Denmark) coupled to a Scanvac CoolSafe cryo unit (Labogene; Allerød, Denmark) with 1000 g and 30°C for 3h or overnight. Dried extracts were kept at -80°C until further use.

4.2.5. Pigment measurement

The pigment aliquot was diluted 10-fold with methanol and measured spectrophotometically in an Epoch2 96-well plate reader (Biotek/Agilent; Santa Clara, CA, USA) at 470 nm, 652 nm

and 665 nm. Pigment concentrations were calculated as reported previously with a pathlength correction of 0.51 assuming a pure methanolic extract (Lichtenthaler & Buschmann, 2001; Warren, 2008)

4.2.6. Metabolite profiling

Primary metabolites were analyzed as described before (Lisec et al., 2006). For the analysis via GC-MS-TOF a dried aliquot of the polar phase was derivatized by adding 40 µL of (20 mg * mL^{-1}) methoxyamine hydrochloride and incubating at 37°C first for 2 h followed by addition of 70 µL of MSTFA and continued incubation at 37°C for 30 min. The samples were injected by an autosampler Gerstel Multi-Purpose system (Gerstel GmbH & Co.KG, Mülheim an der Ruhr, Germany) into a gas chromatograph coupled to a time-of-flight mass spectrometer (Leco Pegasus HT TOF-MS) (LECO Corporation; St. Joseph, MI, U.S.A.). Analysis of secondary metabolites followed a previously described protocol (Giavalisco et al., 2009). Here a dried aliquot of the polar phase was resuspended in 150 µl H2O:MeOH (50:50) and samples injected to an Acquity UPLC system (Waters Corporation; Milford, MA, U.S.A.) coupled to an Exactive Orbitrap mass detector (ThermoFisher Scientific; Waltham, MA, U.S.A.) via a heated electrospray source (ThermoFisher Scientific; Waltham, MA, U.S.A.) . Mass spectra were obtained by running samples in negative ionization mode. For analysis of lipophilic compounds a previously established protocol was used (Hummel et al., 2011). An aliquot of the dried organic phase was resuspended in 100 µL of UPLC-grade acetonitrile:isopropanol (70:30) mix, of which 2 µl were injected onto an Acquity UPLC system (Waters Corporation; Milford, MA, U.S.A.) equipped with an RP C8 column (Hummel et al., 2011). Mass spectra were obtained by running samples in positive ionization mode on an Orbitrap high-resolution mass spectrometer: Fourier-transform mass spectrometer (FT-MS) coupled with a linear ion trap (LTQ) Orbitrap XL (ThermoFisher Scientific; Waltham, MA, U.S.A.).

4.2.7. Peak picking/Area calculation

For targeted analysis peaks were picked manually and peak area calculated with Xcalibur Version 4.2.47 (ThermoFisher Scientific; Waltham, MA, U.S.A.). Non-targeted analysis was performed with Genedata Expressionist ® 14.0.5 (Genedata; Basel, Switzerland).

4.2.8. Computational analysis and used packages

All analysis was performed using R statistical software v4.0.2 in the RStudio environment or on a unix-based high performance computing cluster. Code was written in base R or with the help of the packages tidyverse, broom, cowplot, ggpubr (Kassambara, 2020; Robinson et al.,

2021; Wickham et al., 2019; Wickham & RStudio, 2020; Wilke, 2020b). Imputation by QRILC was done with the package imputeLCMD (Lazar, 2015). Heatmaps were created with the pheatmap package (Kolde, 2019). I am using the bootstrap::jackknife function to generate jack-values, which I am using for statistical analysis (original et al., 2019). Bubble plots were created by a custom script but inspired by the package GOplot (Walter et al., 2015).

4.2.9. RNA isolation

RNA isolation was performed similarly to a previously described protocol (Bugos et al., 1995), Plant tissue was ground to a fine powder by a Retsch mill (Retsch, Haan, Germany), 100 mg was aliquoted into a tube. Vortexing after each step, the following chemicals were added consecutively: 280 µl extraction buffer (100 mM Tris HCl pH 9, 200 mM NaCl, 15 mM EDTA pH 8, 0.5% sarcosyl, 8 µl/ml β-mercaptoethanol), 280 µl phenol, 56 µl chloroform-isoamylalcohol (24:1) and 19.6 µl 3M NaOAc (pH 5.2). After 15 minutes incubation on ice, samples were centrifuged for 10 min at 10000 rpm and the supernatant was transferred to a new tube. Now 280 µl phenol-chloroform-isoamylacohol (25:24:1) was added, the sample vortexed, centrifuged for 10 min at 10000 rpm and the upper phase supernatant transferred to a new tube until no white interphase could be detected. RNA was precipitated for 30 minutes – 2 hours. The sample was centrifuged and the pellet washed with 500 µl 80% ethanol/DEPC, before drying the pellet at room temperature. The pellet was resuspended in 200 µl sterile H_2O/DEPC and centrifuged to remove any particulate. In a new tube the supernatant was mixed with 100 µl 8M LiCl, by inverting the tube and RNA was precipitated overnight. The sample was centrifuged 15 min at 10000 rpm at 4°C and the pellet washed with 250 µl 70% ethanol/DEPC. After the supernatant was discarded, the pellet was resuspended in 50 µl H_2O/DEPC. RNA concentration was measured with a NanoDrop[TM] One spectrophotometer (Thermo Fisher Scientific; Waltham, MA, U.S.A.). Residual DNA was digested with the Invitrogen Turbo DNAse free kit (Invitrogen/Thermo Fisher Scientific; Waltham, MA, U.S.A.). Exact RNA quality and quantity was measured with the Bioanalyzer 2100 (Agilent; Santa Clara, CA, U.S.A.).

4.2.10. RNA sequencing

RNA was sequenced as a service of the MPIPZ (Cologne, Germany). Samples were run on a Illumina HiSeq 3000 RNA sequencing machine (Illumina; San Diego, CA, U.S.A.). RNA library preparation included a polyA enrichment. Per sample 20 million paired-end reads were sequenced.

4.2.11. Orthology search

Orthologues between *Arabidopsis thaliana* (TAIR10) and *Solanum lycopersicum* (ITAG4.0) were searched by OrthoFinder v2.5.2 using the DIAMOND ultra-sensitive setting (Emms & Kelly, 2019). Data was obtained from TAIR (https://www.arabidopsis.org/) and Solgenomics (https://www.solgenomics.net/).

4.2.12. RNA read mapping

RNA reads were mapped to the genome with the LSTRAP pipeline as described previously (Proost et al., 2017).

4.2.13. Differential gene expression analysis

Differential gene expression analysis was performed by the package deseq2. Genes were considered differentially expressed, when they had a log2FC >= 1 and an adjusted p-value <= 0.1 as suggested by literature (Conesa et al., 2016; Lamarre et al., 2018). Normalized counts after deseq2 processing were used for further analysis.

4.2.14. Proteomics analysis

Proteins were isolated from the pellet of MTBE-extracted samples, processed and analyzed with minor modifications as described previously without fractionation (Salem et al., 2020; Sokolowska et al., 2019). The amount of denaturation buffer was limited to 50 µL and samples were adjusted to 40 µg of protein. Data was analyzed with the proteus package (Gierlinski et al., 2018) and the help of the package limma for quantile normalization (Ritchie et al., 2015; Smyth & Speed, 2003). Proteins were considered differentially expressed, when they had a log2FC >= 1 and an adjusted p-value <= 0.1.

4.2.15. Gene ontology enrichment analysis

Gene ontology enrichment analysis was performed with the package topGO (Alexa & Rahnenfuhrer, 2021).

4.2.16. Data integration

Transcriptomic and metabolomics data was integrated and visualized with pathview (Luo & Brouwer, 2013).

4.3. Results

4.3.1. Phenotype

Plants from the *cana1* line show a variety of differently pigmented leaves, from completely white, over differently patterned variegated until completely green (Figure 64).

Figure 64: Variegation phenotype of cana1 leaves ranging from (almost) completely green (A) over variegated (B) to completely white (C)

I wanted to quantify the pigment content and therefore extracted pigments and measured them spectrophotometrically (Figure 65).
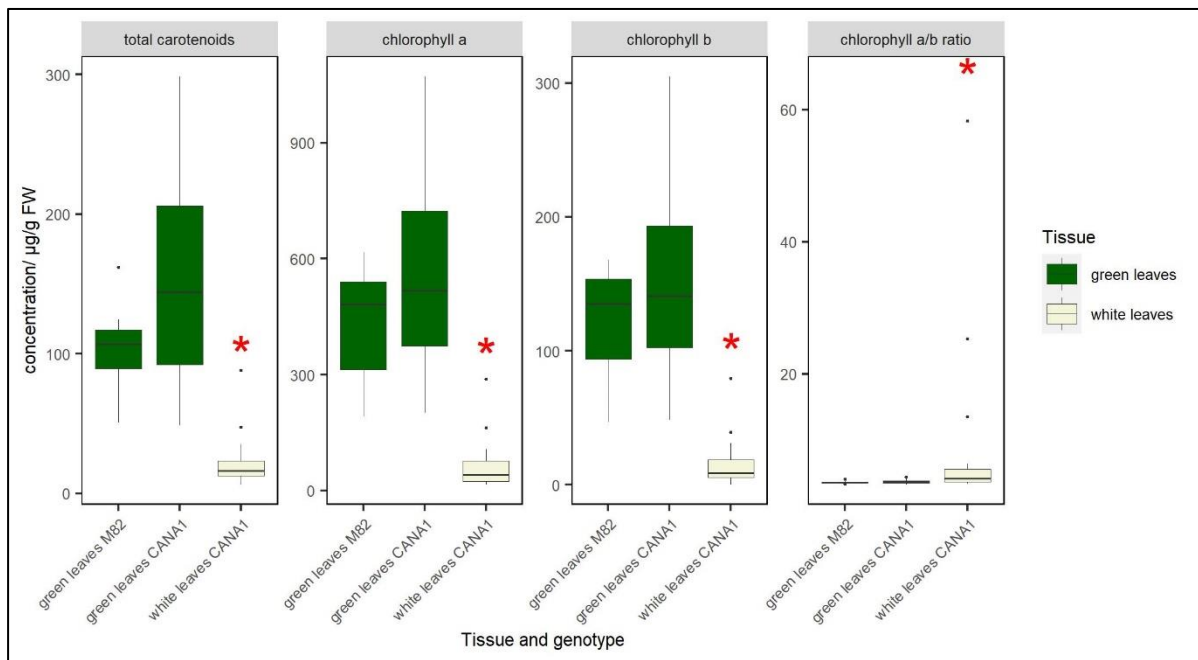
*Figure 65: Content of different pigments and chlorophyll a/b ration in green and white leaves of cana1 compared to wild type leaves; *: p-value ≤ 0.05 in pairwise wilcox-test, n = 9-27*

White leaves of *cana1* plants were almost completely devoid of any chlorophyll and carotenoids, whereas green leaves showed wild type levels of pigments (Figure 65).

I also measured pigment content in stem tissue, where I detected a strongly reduced amount, compared to wild-type stems (Figure 66).
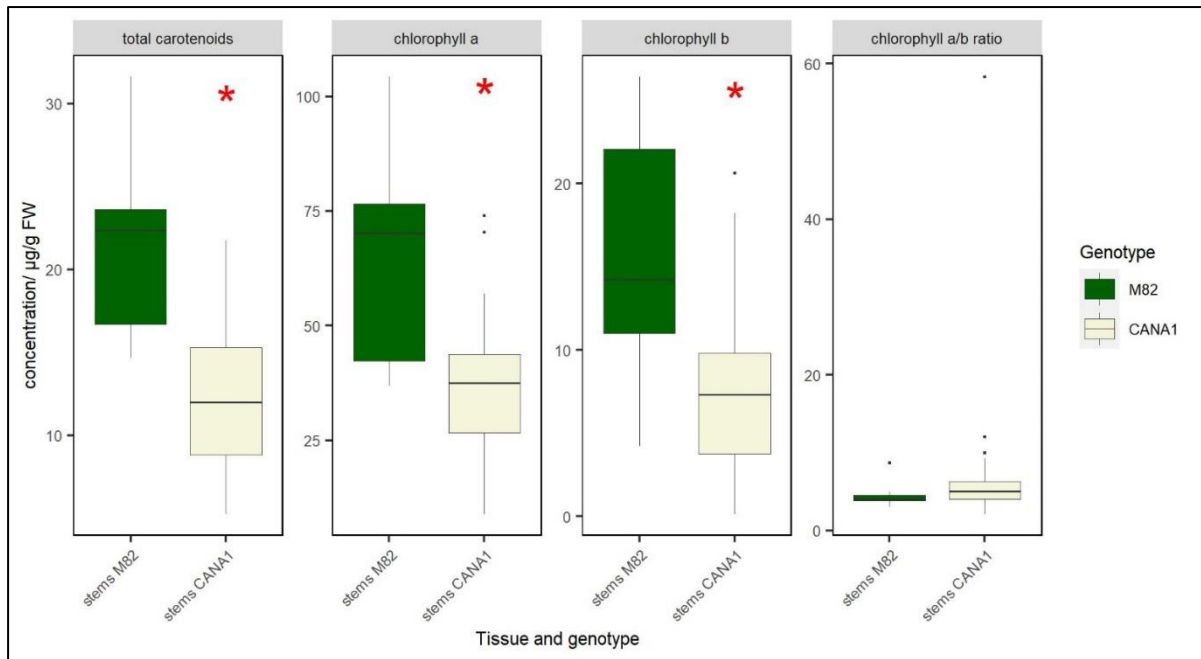
Figure 66: Content of different pigments and chlorophyll a/b ration in stems of cana1 compared to wild type stems; *: p-value ≤ 0.05 in pairwise student's t-test, n = 7-21

4.3.2. RNAseq

To investigate processes at the transcriptional level, I extracted RNA from leaves and stems of *cana1* and wild-type plants and sent it for RNA sequencing.

The PCA nicely separates tissues on the first PC and genotypes on the second PC (Supplementary Figure 4). I performed differential gene expression analysis and discovered many differentially expressed genes (DEGs), both comparing mutant and wild type leaves, but also when comparing white and green leaves of *cana1* plants (Supplementary Figure 5). I found the most DEGs in the comparison of white mutant leaves to green wild-type leaves, with 2284 DEGs exclusive to this comparison (Supplementary Figure 5). I went on, and performed gene ontology enrichment analysis on the different contrasts.
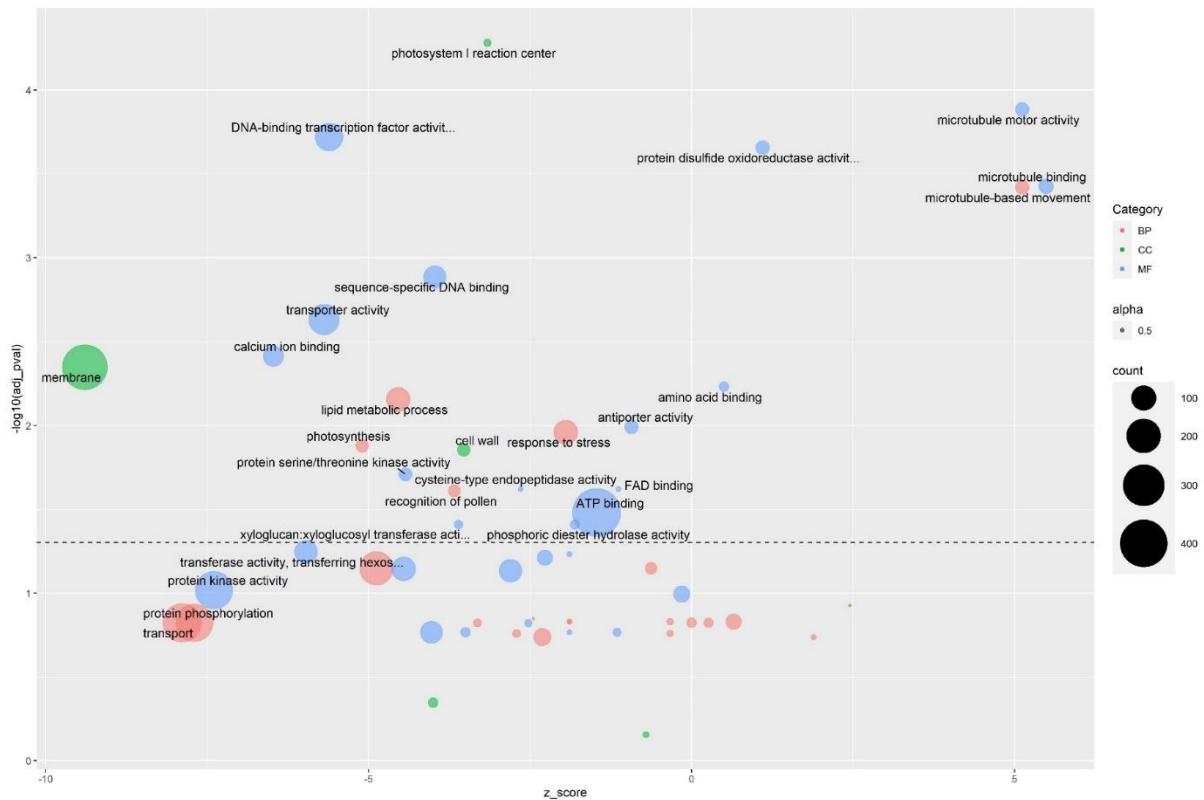
Figure 67: Bubble plot of overrepresented GO terms in transcripts, when comparing cana1 mutant white leaves to wild type green leaves. The x-axis shows the z-score and the y axis shows the –log₁₀(adj. p-value). The dotted horizontal line indicates a p-value of 0.05.

In the comparison of mutant white leaves vs wild-type green leaves I can find many terms related to photosynthesis, but also related to DNA-binding, as well the term "protein disulfide oxidoreductase activity" (Figure 67). When comparing white and green leaves from *cana1* plants I can see an even stronger enrichment of photosynthesis terms, however photosynthesi-related terms barely surpass a p-value threshold of 0.05 in the comparison of green mutant and wild type leaves (Supplementary Figure 6+7). On the other hand the term "aminoacyl-tRNA ligase activity" is strongly enriched in this comparison (Supplementary Figure 7).
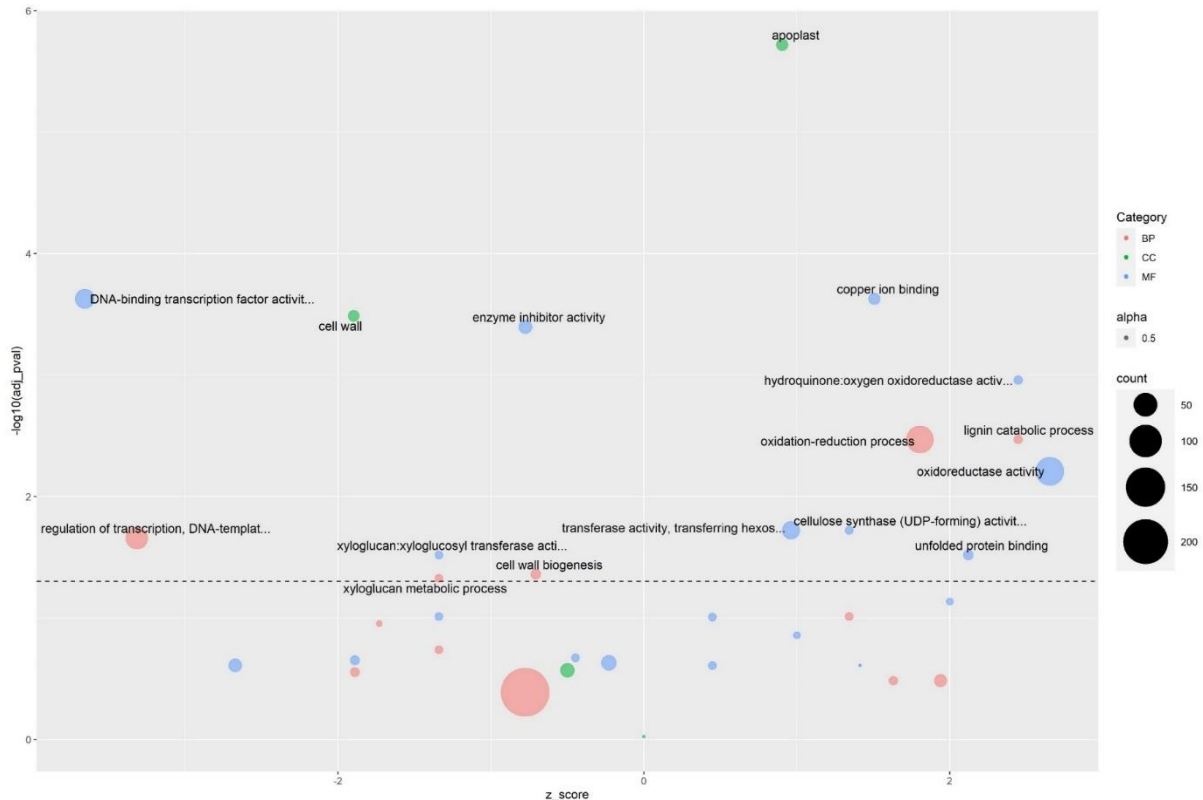
Figure 68: Bubble plot of overrepresented GO terms in transcripts, when comparing cana1 mutant stems to wild type stems. The x-axis shows the z-score and the y axis shows the $-\log_{10}$(adj. p-value). The dotted horizontal line indicates a p-value of 0.05.

When comparing DEGs of stems of mutant and wild type I mostly found terms, which are likely related to the general growth of the stem, like "lignin catabolic process", "cell wall biogenesis" and "xyloglucan metabolic process" (Figure 68). I also found the term "DNA-binding transcription factor activity".

To get a better understanding of which genes are exactly up- or downregulated, I took a closer look at the expression profile of genes related to some of the discovered terms.

First of all I was interested to see, whether the mutated gene showed any changed expression (Figure 69).
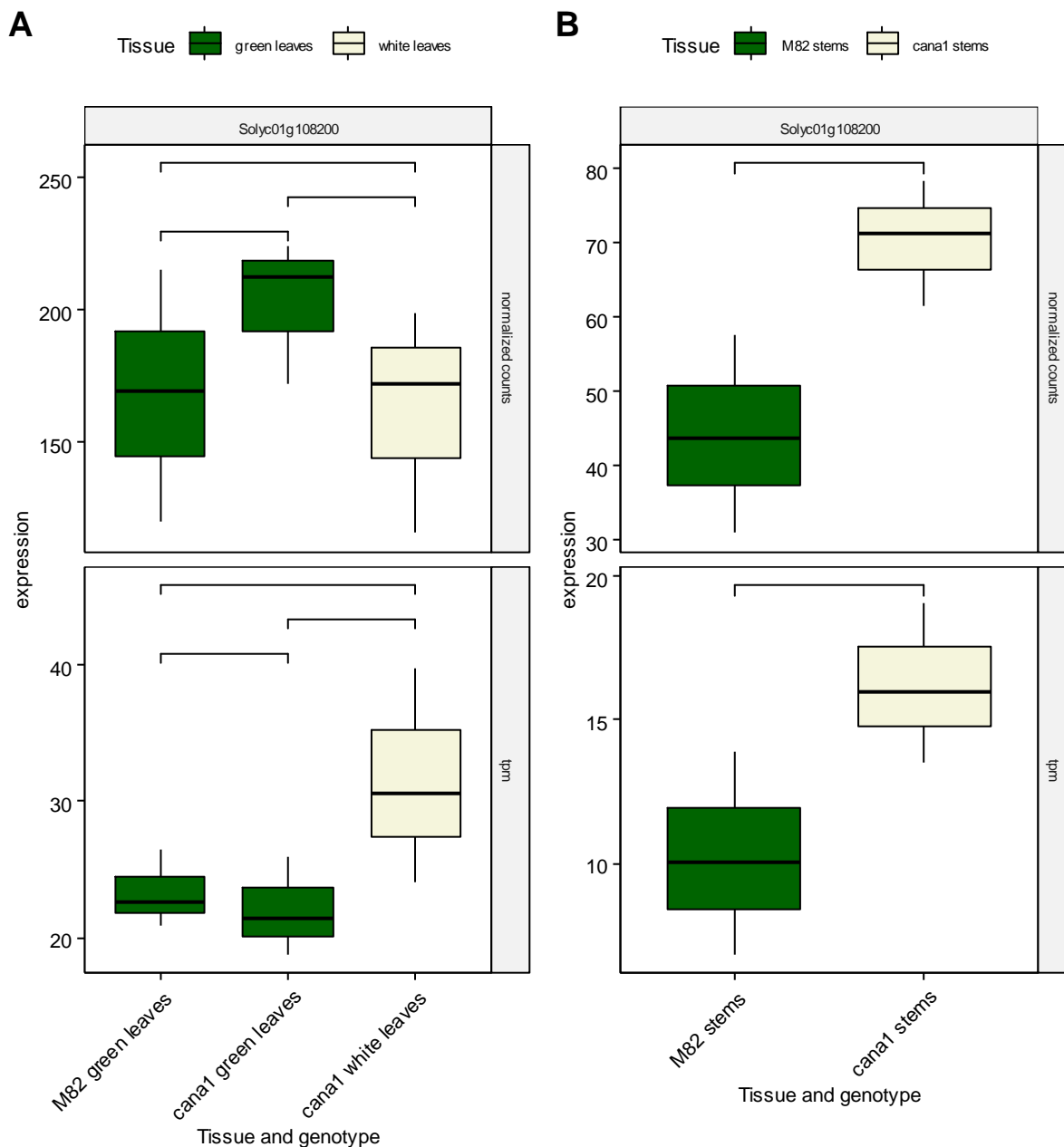
*Figure 69: Expression of Solyc01g108200 (CANA1) in leaves (A) and stems (B) of wild type and cana1 mutant plants. Upper panel shows normalized counts and lower panel tpm values; student's t-test was performed by ggpubr::stat_compare_means, P-values are annotated as *: $p \leq 0.05$; **; $p \leq 0.01$; ***: $p \leq 0.001$; ****: $p \leq 0.0001$ n = 3*

For both tissues, I could not detect any significant changes. If anything, the raw reads of mutant green leaves appear slightly increased, although not significantly (Figure 69).

As the CANA1 gene is a putative photosystem assembly factor (Zagari et al., 2017), I was interested in the expression of other assembly factors and photosystem components. I gathered known genes from *Arabidopsis thaliana* from different sources (Hankamer et al., 1997; Jensen

Chapter 4: CANA1

et al., 2007; Lu, 2016; Shi et al., 2012; Yang et al., 2015)(Supplementary Table 1) and used them to investigate corresponding tomato orthologs. First I investigated the expression of other known assembly factors (Figure 70).
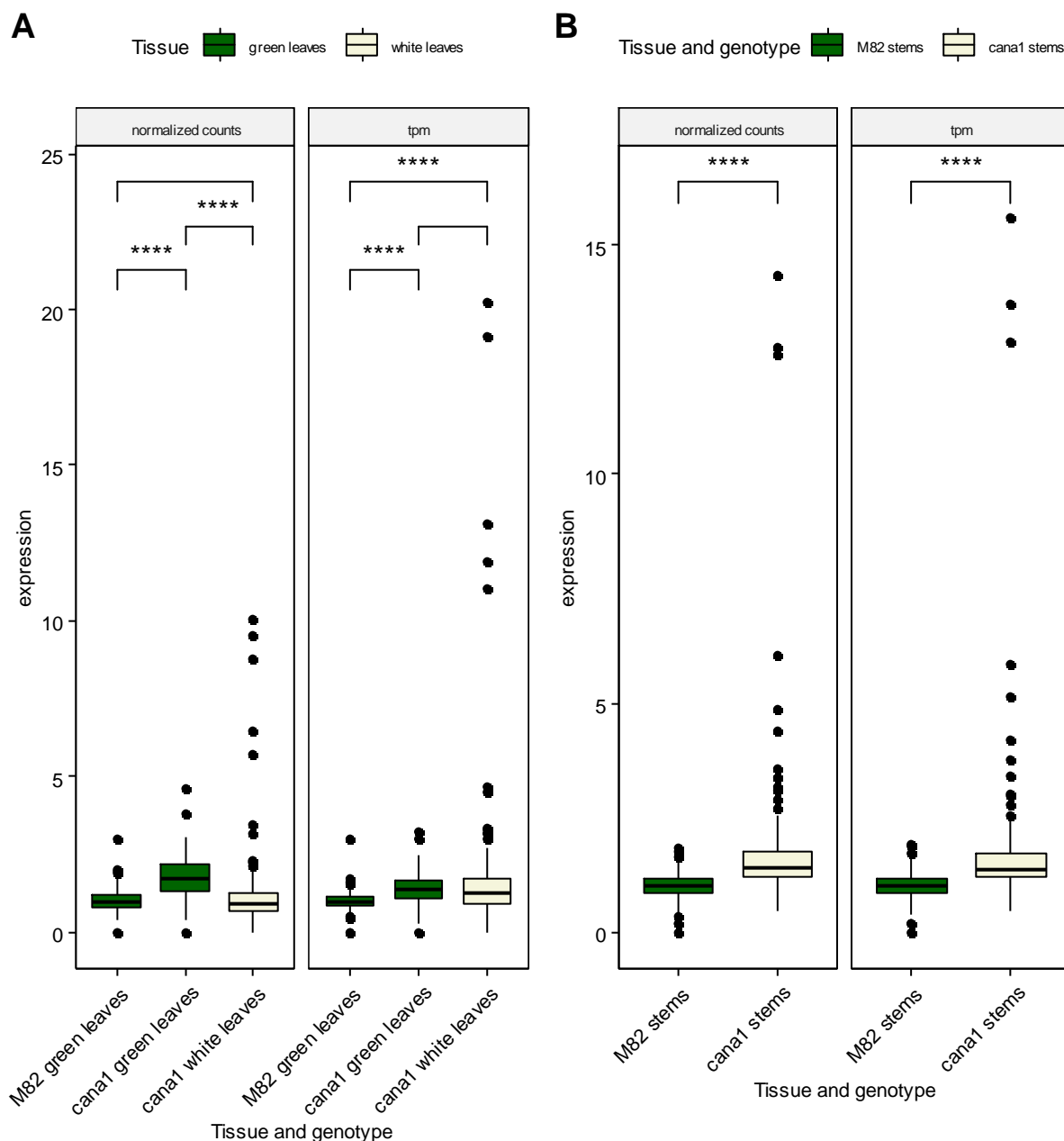
*Figure 70: Expression of assembly factors in leaves (A) and stems (B) of wild type and cana1 mutant plants normalized to wild type levels. Left panel shows normalized counts and right panel tpm values; student's t-test was performed by ggpubr::stat_compare_means, P-values are annotated as *: p ≤0.05; **; p≤0.01; ***: p≤0.001; ****: p≤0.0001.*

One can see that green mutant leaves have an increased expression, when compared to wild-type leaves, as do mutant stems compared to wild-type stems (Figure 70). White mutant leaves however have, based on raw reads, an expression comparable to wild-type leaves, although

some genes seem to have some extreme outliers (Figure 70). I was interested to see, which these genes were and therefore extracted them and checked their expression (Figure 71).
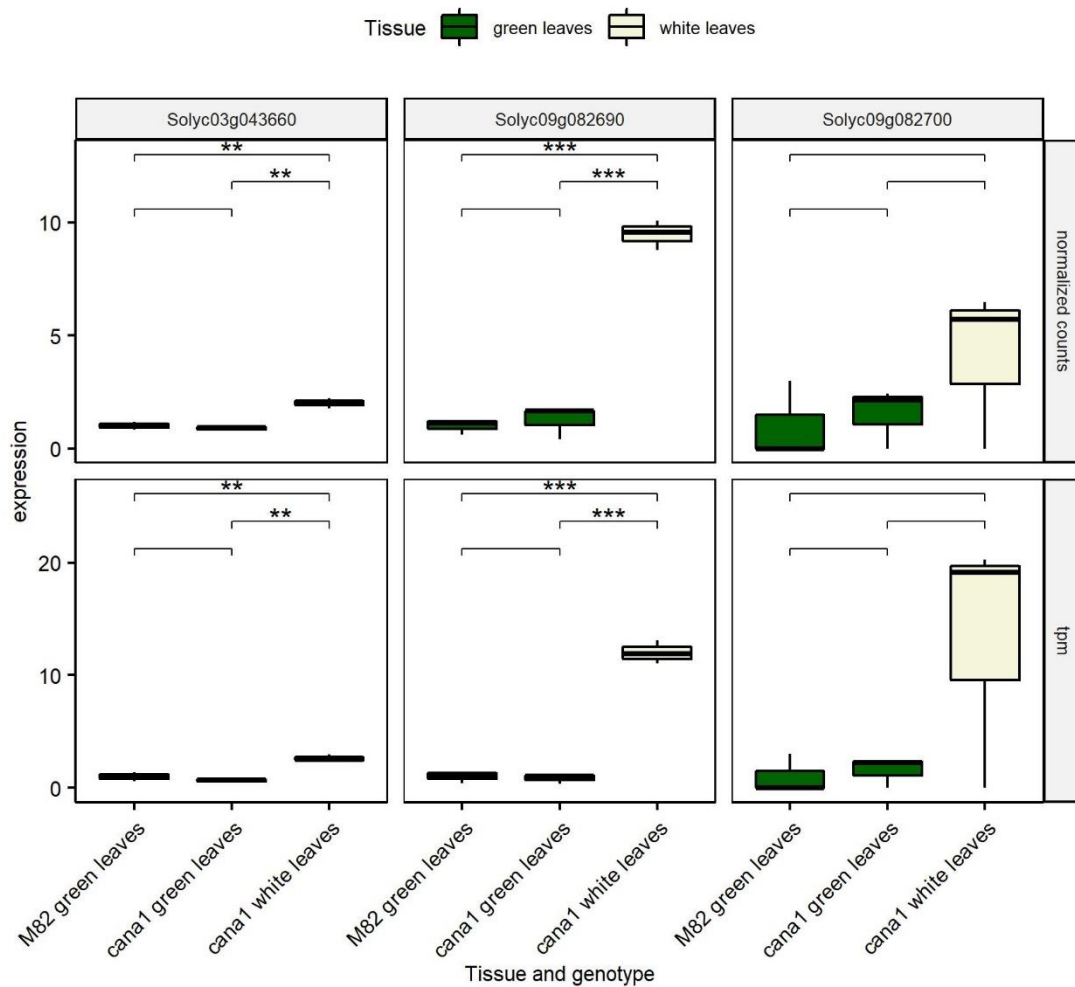
*Figure 71: Expression of Solyc03g043660, Solyc09g082690 and Solyc09g082700 of wild type and cana1 mutant plants normalized to wild type levels. Upper panel shows normalized counts and lower panel tpm values; student's t-test was performed by ggpubr::stat_compare_means, P-values are annotated as *: p ≤ 0.05; **; p≤ 0.01; ***: p≤ 0.001; ****: p≤ 0.0001, n = 3*

The genes with high expression in white leaves, I extracted from the assembly components, were orthologs of early-light inducible proteins (ELIPs) and a DEG protease, which are important for thylakoid development and degradation of thylakoid proteins (Casazza et al., 2005; T. Sun et al., 2019; X. Sun et al., 2010).

If I filter out these genes, I can now see a clearer expression pattern (Figure 72).
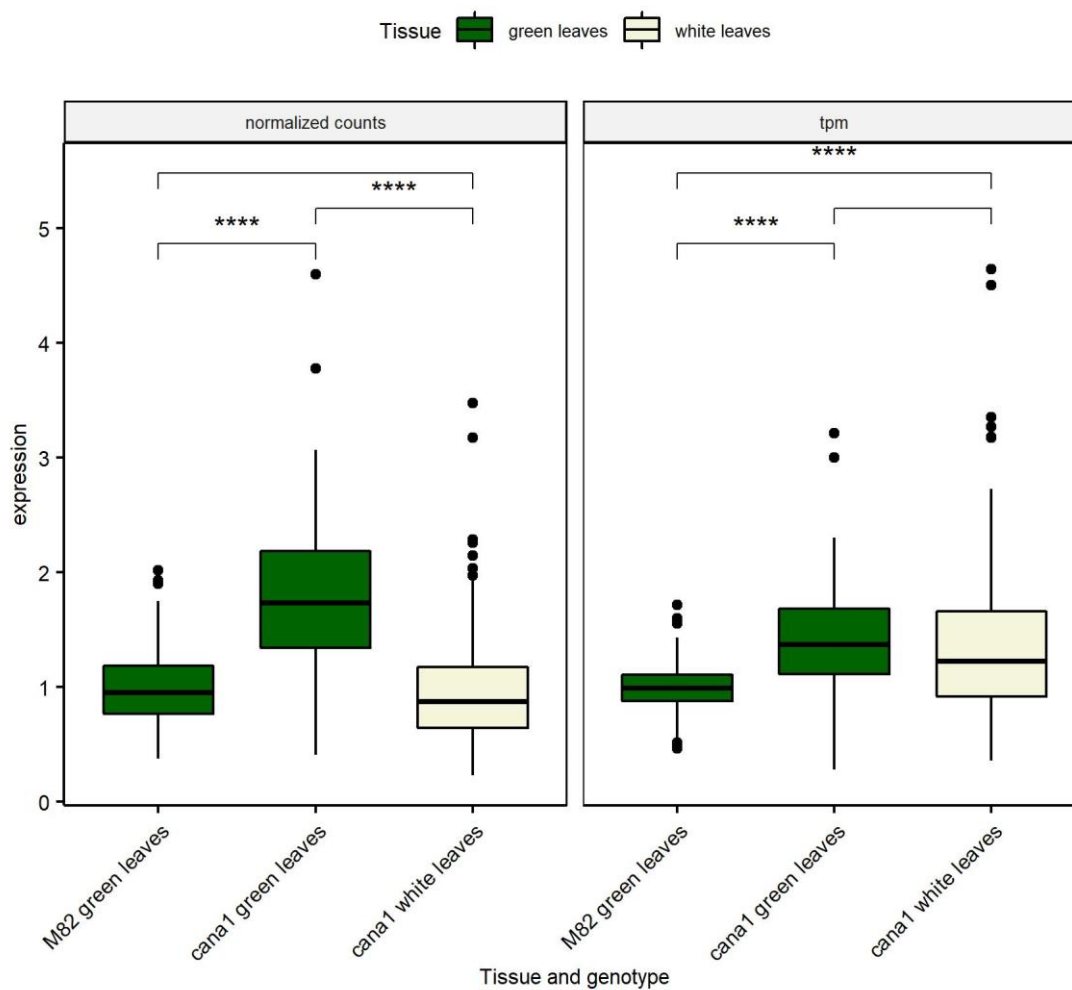
*Figure 72: Expression of assembly factors in leaves of wild type and cana1 mutant plants normalized to wild type levels after removal of ELIP and DEG orthologs. Left panel shows normalized counts and right panel tpm values; student's t-test was performed by ggpubr::stat_compare_means, P-values are annotated as *: p ≤ 0.05; **: p≤ 0.01; ***: p≤ 0.001; ****: p≤ 0.0001.*

I can see a clear upregulation of the remaining photosystem assembly genes in green *cana1* leaves compared to green M82 leaves (Figure 72). If I look at the tpm-based expression profile, I can see that a similar amount of transcripts are produced in white mutant leaves as in green mutant leaves, in relation to the total amount of transcripts (Figure 72).

If I now look at all the genes, which make up components of the final photosystems, I can see a similar pattern (Figure 73).
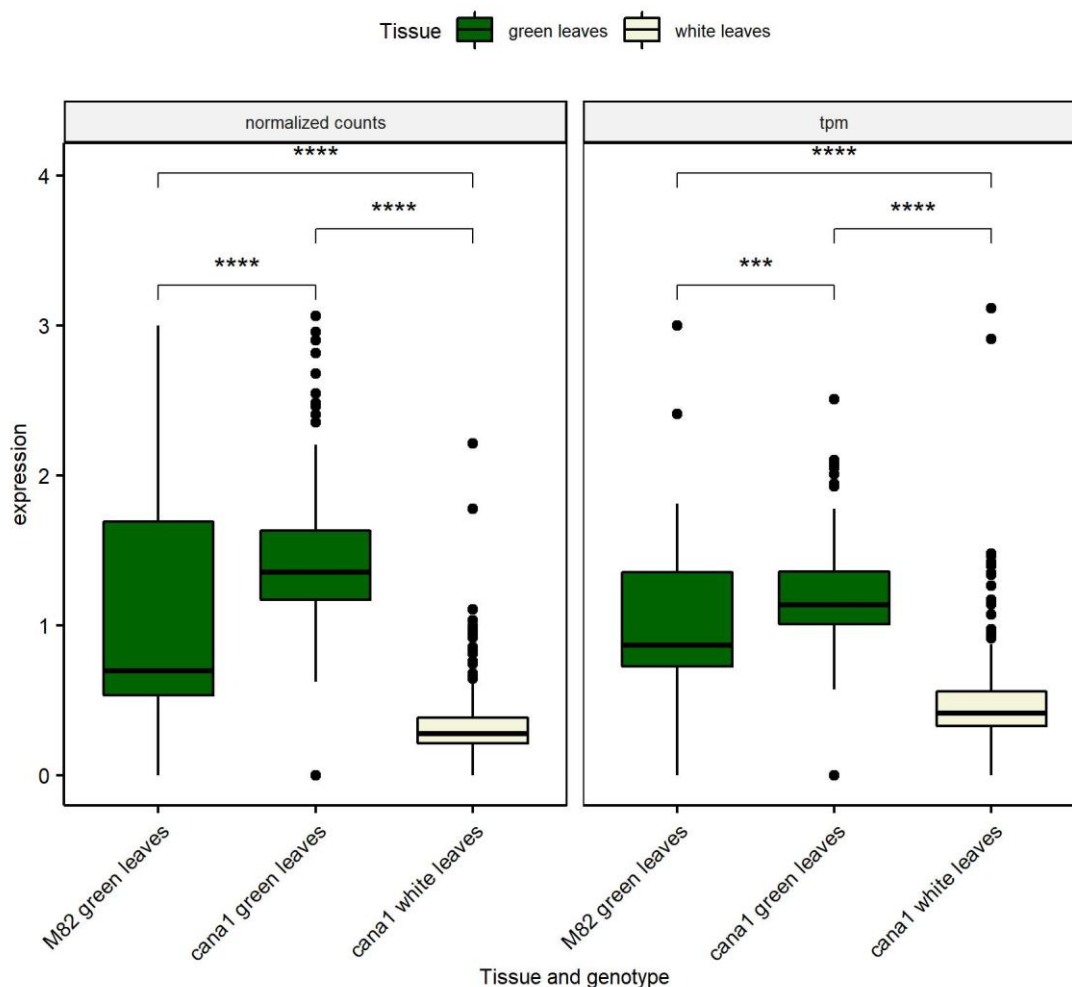
Figure 73: Expression of photosystem components in leaves of wild type and cana1 mutant plants normalized to wild type levels. Left panel shows normalized counts and right panel tpm values; student's t-test was performed by ggpubr::stat_compare_means, P-values are annotated as *: p ≤ 0.05; **: p≤ 0.01; ***: p≤ 0.001; ****: p≤ 0.0001.

Green leaves from *cana1* plants show a significant increase in the expression of photosystem components, while white leaves even show a decrease in expression, when compared to M82 leaves (Figure 73).

As I found the term "protein disulfide oxidoreductase activity" in some of the contrasts of the GO enrichment, I wondered whether any compensatory mechanisms may be elicited by the reduction in function of the disulfide isomerase CANA1 in the mutant tissues. I therefore extracted all DEGs, with the GO term "protein disulfide oxidoreductase activity", as this was the term I found before and also because no other DEG had a GO annotation to the term "protein disulfide isomerase activity". I then compared the summarized relative expression of these genes of *cana1* and wild type plants (Figure 74).
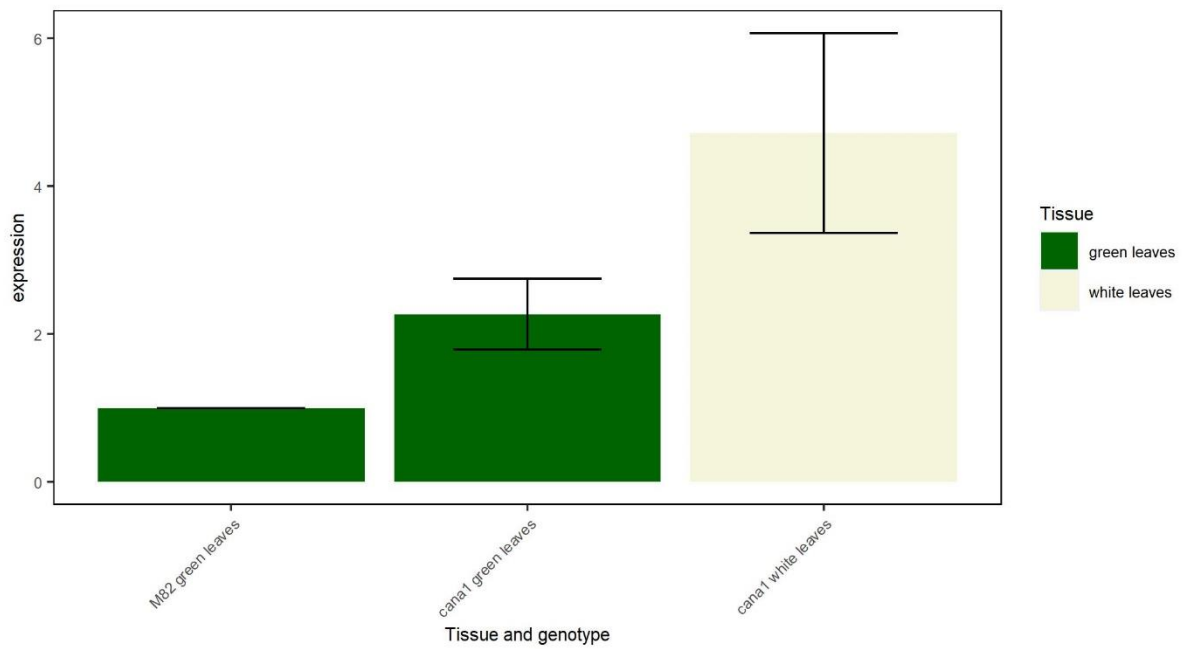
Figure 74: Expression of genes annotated with the GO term protein disulfide oxidoreductase activity in leaves of wild type and cana1 mutant plants normalized to wild type levels.

As I can see, these genes show a 2-fold induction in *cana1* green leaves and a roughly 5-fold induction in *cana1* white leaves, when compared to leaves of M82. Although this is just a summary of many genes and some genes may show a different pattern, I found a group of genes which consistently show an increased expression in white mutant leaves (Figure 75).
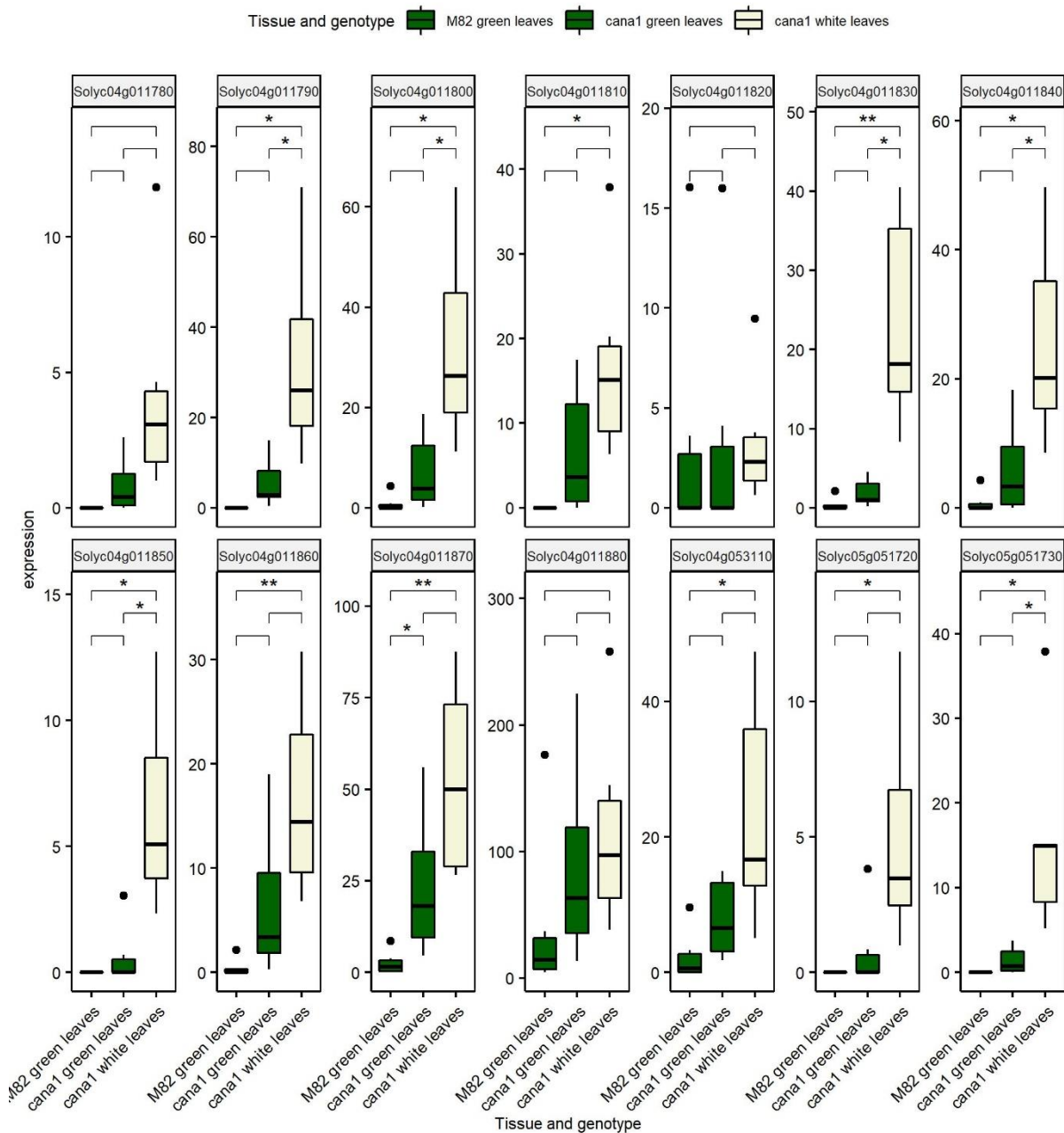
Figure 75: Expression of a subset of genes annotated with the GO term protein disulfide oxidoreductase in leaves of wild type and cana1 mutant plants normalized to wild type levels; student's t-test was performed by ggpubr::stat_compare_means, P-values are annotated as *: p ≤0.05; **; p≤ 0.01; ***: p≤ 0.001; ****: p≤ 0.0001, n=3

Interestingly an orthology search showed that all these 14 genes seem to be orthologs of each other with no orthologs in *Arabidopsis thaliana*. However, TAIR lists the genes GRXS1 (AT1G03020) and GRXS6 (AT3G62930) homologs of these genes. Either way, the genes belong to a Glutaredoxin//Thioredoxin superfamily and seem to be responsible for arsenate detoxification

(https://solcyc.solgenomics.net/gene?orgid=LYCO&id=SOLYC04G011800.1#tab=TU).

4.3.3. Proteomics

I used another aliquot of the plant material I had sent for RNA sequencing, to extract proteins and subject it to proteomics analysis. Also here I investigated differential expression of proteins and analyzed the data with a similar pipeline as the RNAseq data. The results largely confirm our observations of the transcriptomic analyses. For example, when I compare white *cana1* leaves with M82 leaves, I see an enrichment of proteins related to "chloroplast thylakoid membrane" and many other photosynthesis-related terms (Figure 76).
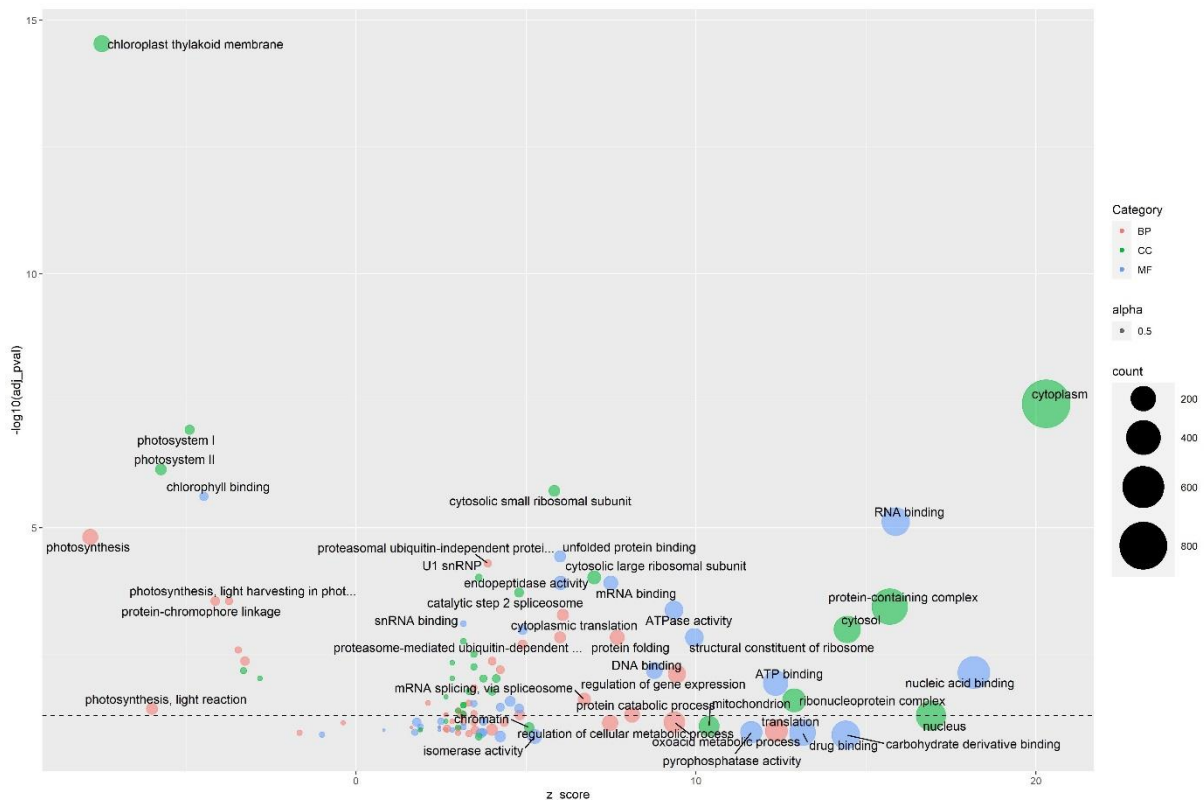
Figure 76: Bubble plot of overrepresented GO terms in proteomics, when comparing cana1 mutant white leaves to wild type green leaves. The x-axis shows the z-score and the y axis shows the –log₁₀(adj. p-value). The dotted horizontal line indicates a p-value of 0.05.

Additionally, I can see other ontology terms enriched, which are related to protein folding or degradation and nucleic acid binding and splicing (Figure 76). The enrichment profile of the contrast between white and green mutant leaves, looks very similar to the previously described one (Supplementary Figure 8). Again I find terms related to photosynthesis, protein folding and degradation, as well as nucleic acid binding. This may be a first hint that the proteomic profile of mutant and wild-type green leaves is very similar. In fact, when I did compare those two tissues, I did not find a single significantly enriched term. I only found one differentially expressed protein (A0A3Q7EIF5), which is annotated as a subunit of the oxygen-evolving system of photosystem II.

To get a visual representation of the expression of photosystem-related genes in green and white mutant leaves compared to green wild-type leaves, I utilized the expression data to generate a KEGG graph (Figure 77).
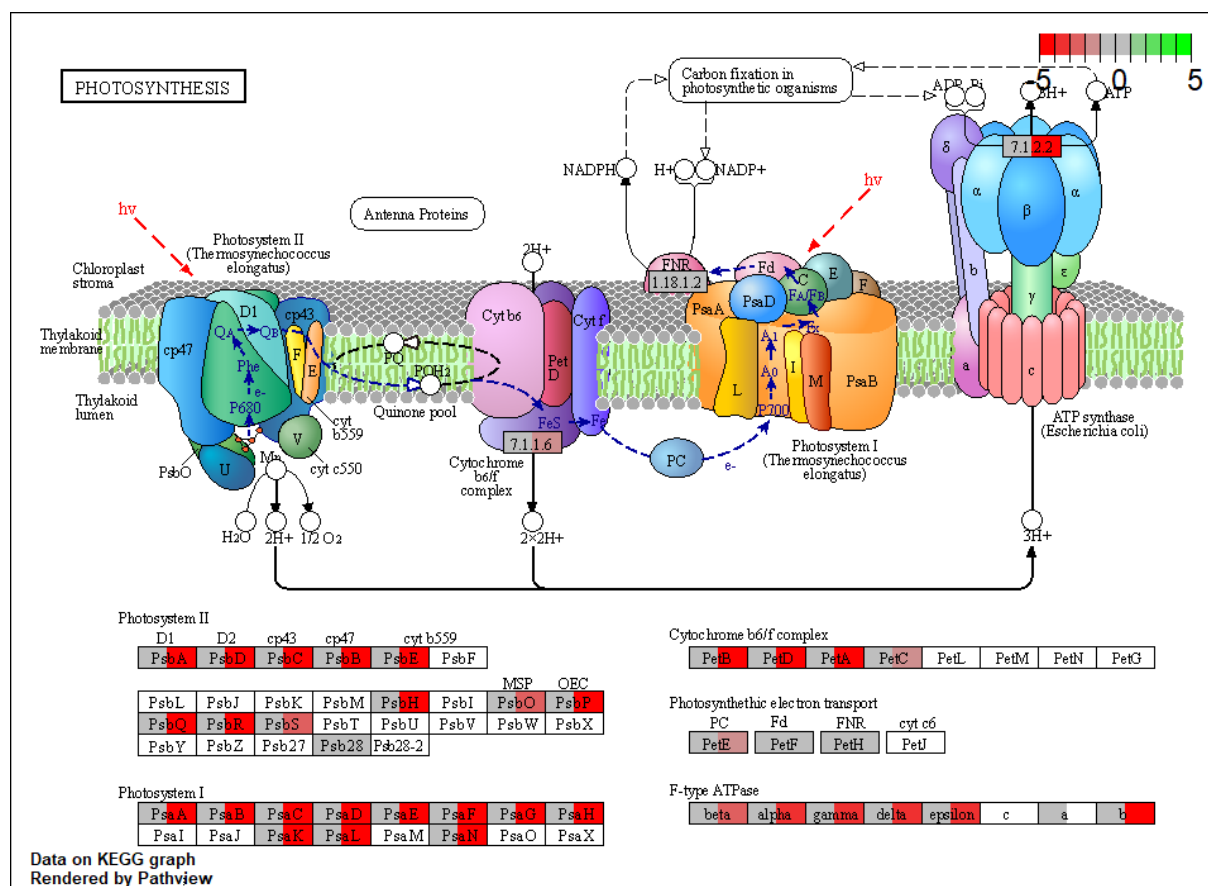
Figure 77: Pathview image of Photosynthesis. Proteins are shown as rectangles. Protein nodes are split in the middle to simultaneously show two comparisons. Changes are the log2-fold change of mutant green leaves (left) or mutant white leaves (right) in comparison to wild type green leaves. Blank nodes could either be not mapped or are not known.

As I can see, almost all proteins, for which I had expression values, show a strong downregulation in mutant white leaves, while no changes can be seen in mutant green leaves (Figure 77). In the comparison of *cana1* stems and M82 stems, I also find many enriched terms related to photosynthesis (Supplementary Figure 9).

4.3.4. Metabolomics

I also extracted metabolites from plants grown in 3 seasons under control and drought conditions. First I investigated primary metabolites (Figure 78). As expected, I see the strongest changes again in the white leaves of *cana1* plants. Many amino acids are strongly upregulated, while a few carboxylic acids are downregulated. The strongest change can be seen in asparagine (Figure 79).
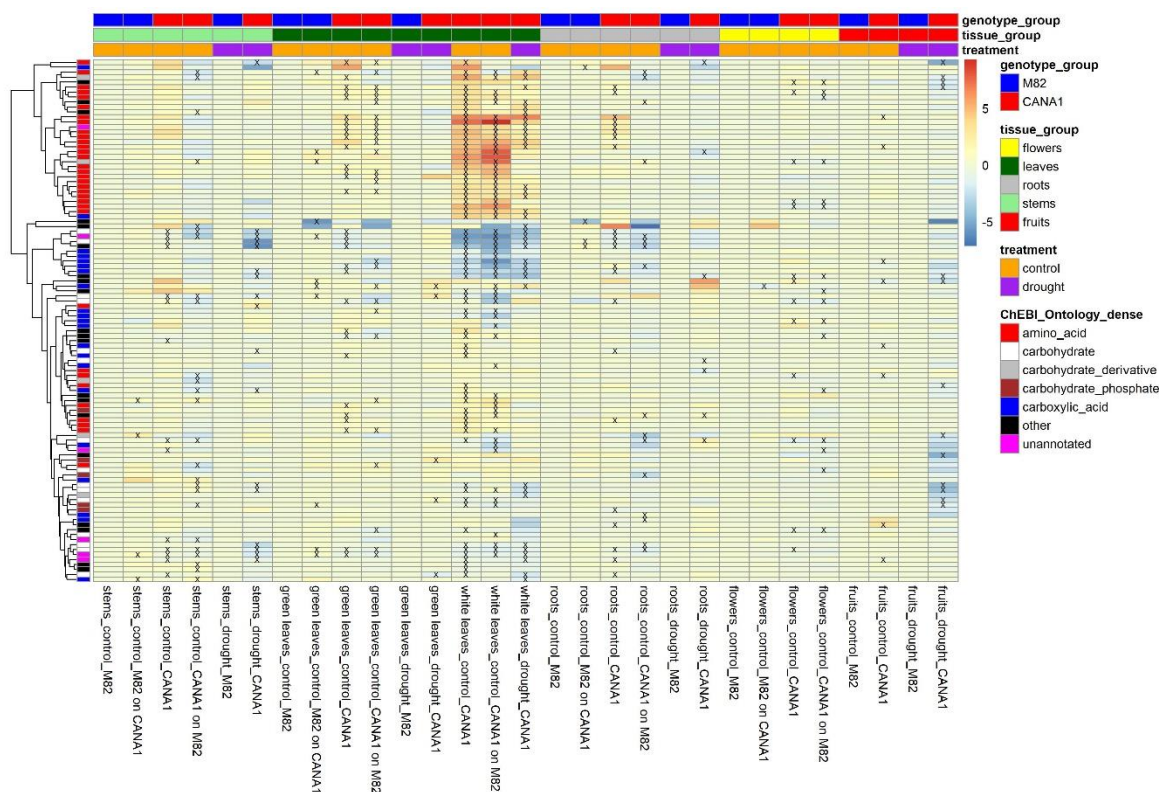
Figure 78: Heatmap of primary metabolites of cana1 and M82 samples as well as samples from cana1 on M82 and M82 on cana1 grafted plants from 3 independent experiments. Values correspond to log2-fold changes in relation to M82 samples of the same tissue and condition. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a pairwise wilcox-test, with trait-wise FDR correction
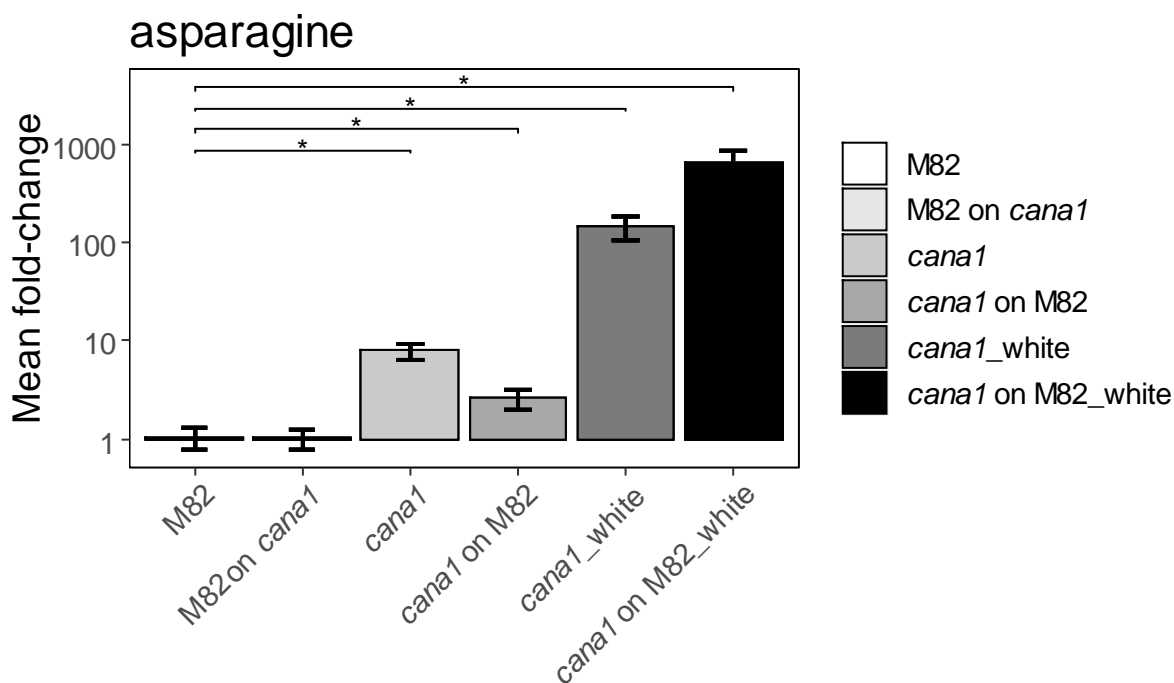
Figure 79: Mean fold-change of asparagine levels in leaves of cana1 and grafted plants relative to M82 levels. Data combines values from 3 independent experiments; *: p-value ≤ 0.05 in pairwise wilcox-test after metabolite-wise FDR-correction. The y-axis was set to a log_{10}-scale to show differences in magnitude

In green leaves the amount of asparagine is already significantly increased, but I can see much stronger changes in white leaves of *cana1* were the content is more than 100-fold increased and white leaves from *cana1* plants grafted onto M82 rootstocks even show a more than 600-fold increase (Figure 79).

When I look at the polar LC-MS data, I see a similar picture but with milder changes (Figure 80). I was also able to detect some amino acids here, which show again a strong increase in white leaves of mutant plants. The strongest change can be detected for the level of tyrosine (Figure 81). Additionally a few flavonoids and dipeptides show an upregulation in white *cana1* leaves, while other flavonoids and some cinnamic acids are downregulated (Figure 80). However, no clear pattern emerges in the expression profile of specialized metabolites.
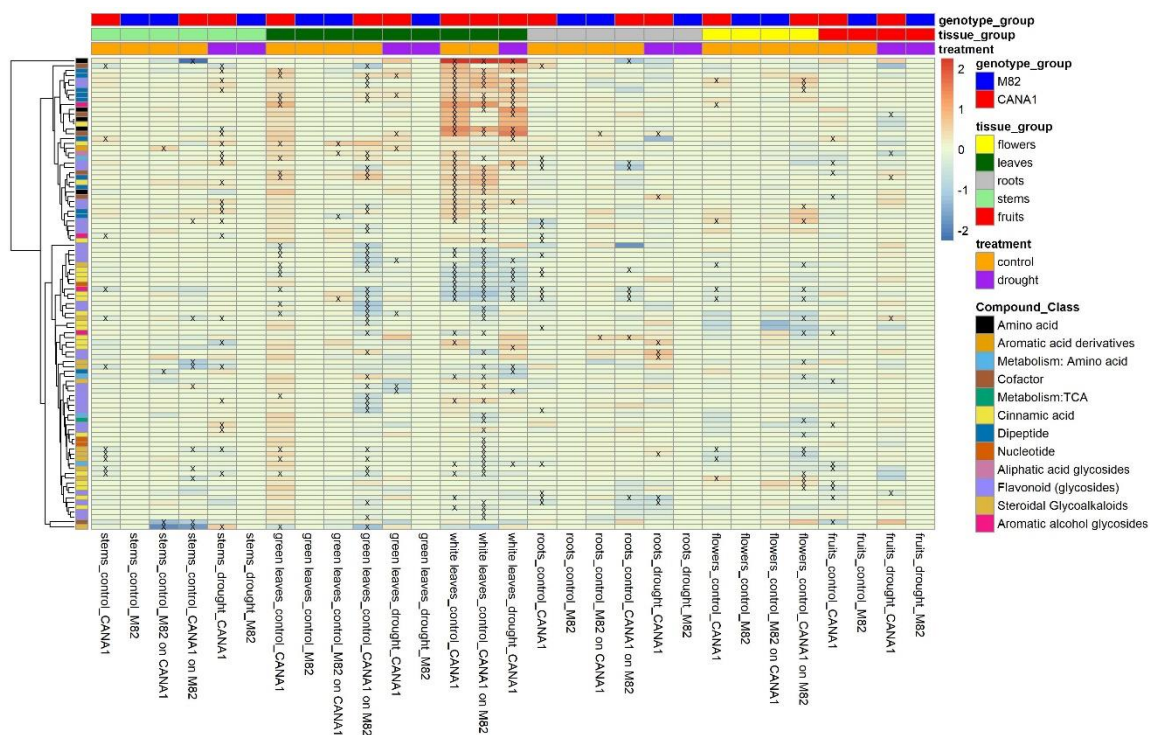
Figure 80: Heatmap of secondary metabolites of cana1 and M82 samples as well as samples from cana1 on M82 and M82 on cana1 grafted plants from 3 independent experiments. Values correspond to log2-fold changes in relation to M82 samples of the same tissue and condition. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a pairwise wilcox-test, with trait-wise FDR correction.
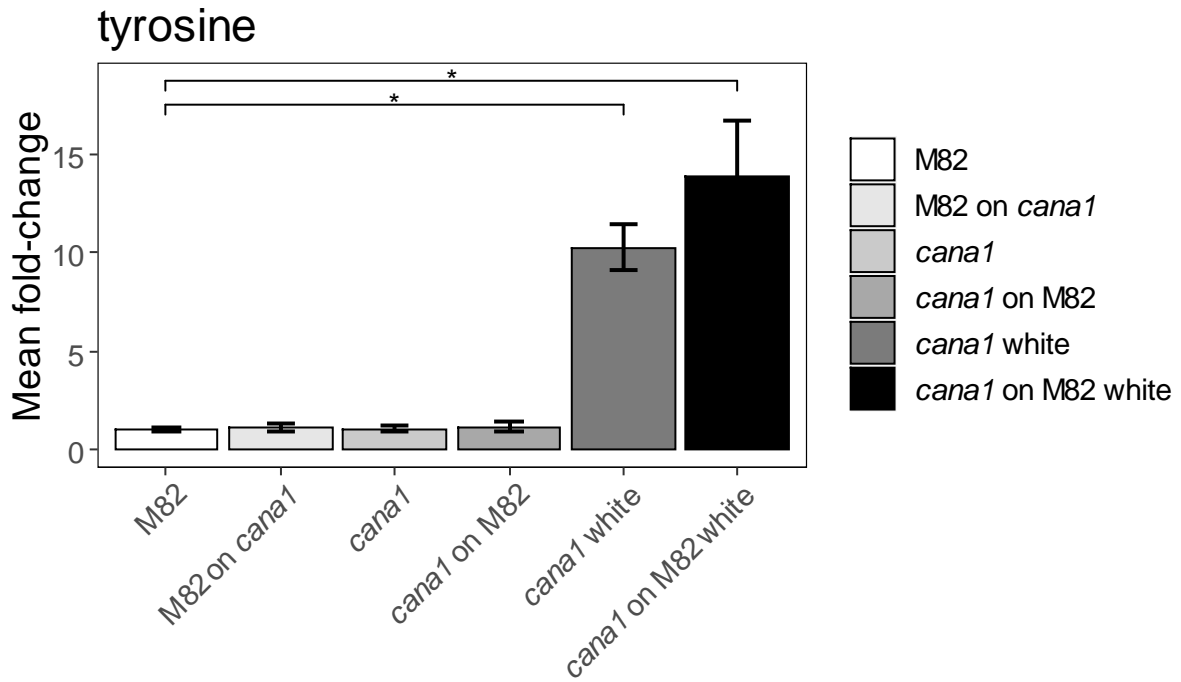
Figure 81: Mean fold-change of tyrosine levels in leaves of cana1 and grafted plants relative to M82 levels. Data combines values from 3 independent experiments; *: p-value ≤0.05 in pairwise wilcox-test after metabolite-wise FDR-correction.

When I look at the lipidomic profile, the pattern looks much clearer (Figure 82).
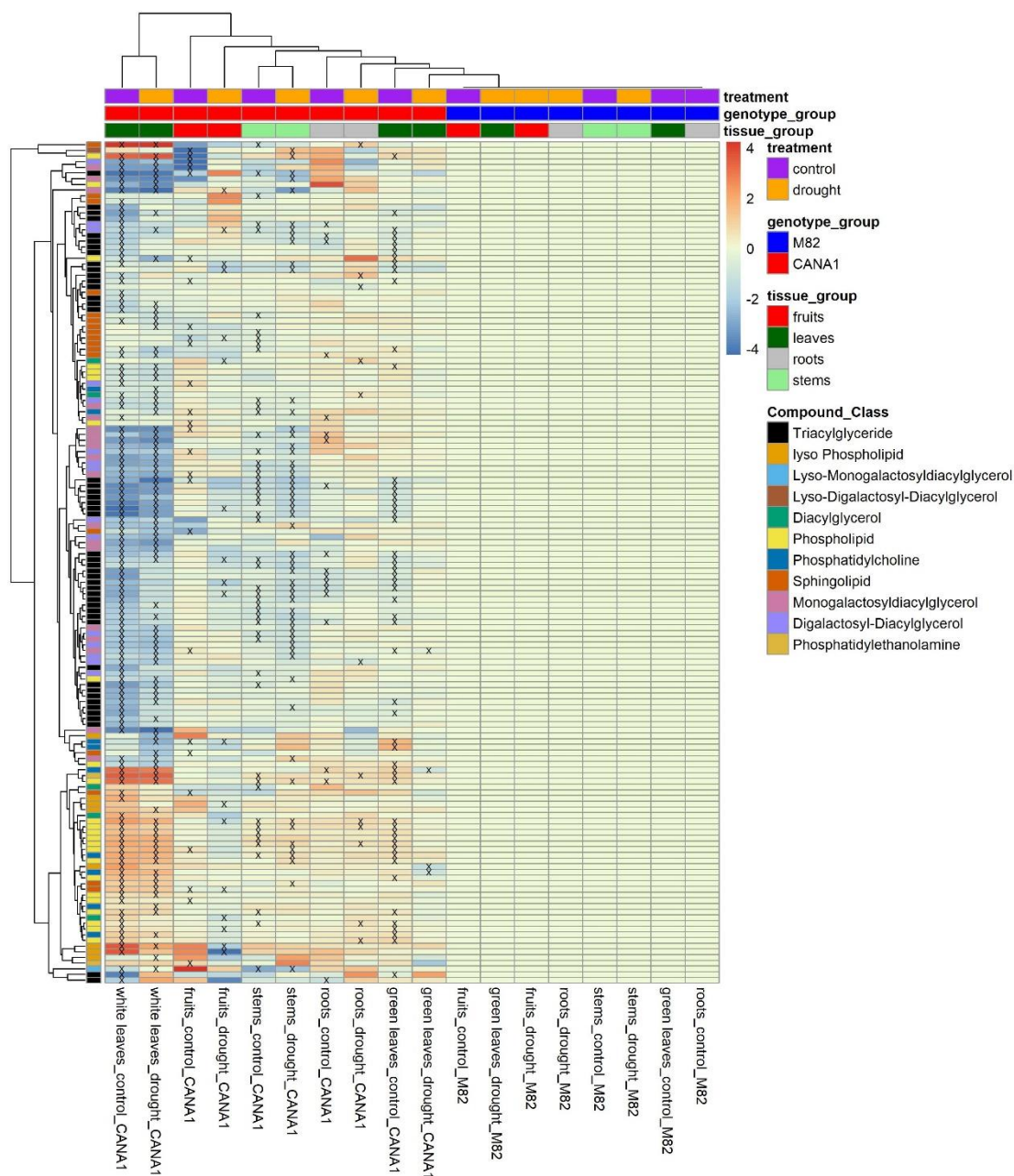
Figure 82: Heatmap of lipophilic compounds of cana1 and M82 samples as from 2 independent experiments. Values correspond to log2-fold changes in relation to M82 samples of the same tissue and condition. Crosses show significant differences in comparison to the respective wild type samples. Significance was tested by a pairwise wilcox-test, with trait-wise FDR correction.

In the white leaves and to a lesser extent also in stems, I can see many TAGs, MGDGs and DGDGs significantly reduced in *cana1* plants (Figure 82). On the other hand, phospholipids, are upregulated in white leaves of mutant plants. The largest upregulation can be found in the glucosylceramide GlcCer d18:1/h16:0, which shows a more than 300-fold induction in white leaves of the mutant (Figure 83).

Figure 83: Mean fold-change of tyrosine levels in leaves of cana1 plants relative to M82 levels. Data combines values from 2 independent experiments; *: p-value ≤ 0.05 in pairwise wilcox-test after metabolite-wise FDR-correction.

4.3.5. Integration

I integrated metabolic and transcriptomic data from a few pathways, which showed a KEGG enrichment, with KEGG graphs. First I investigated starch and sucrose metabolism (Figure 84).

Figure 84: Pathview image of starch and sucrose metabolism. Metabolites are shown as circles and genes are shown as rectangles. Metabolite and gene nodes are split in the middle to simultaneously show two comparisons. Changes are the log2-fold change of mutant green leaves (left) or mutant white leaves (right) in comparison to wild type green leaves. Blank nodes could either be not mapped or are not known.

It can be seen that while *cana1* green leaves show mostly wild type levels of metabolites and transcripts, white leaves of the mutant show strong changes. Sucrose levels are moderately and glucose levels are strongly depleted. I can also see that catabolic genes leading to the generation of glucose are upregulated, while anabolic genes that lead away from glucose are downregulated (Figure 84).

When looking at biosynthesis of different amino acids, the picture looks very different (Figure 85, Figure 86). Many amino acids are moderately and strongly upregulated in green and white leaves of *cana1* plants respectively. As I highlighted earlier asparagine is strongly upregulated but also aspartate, alanine, glutamate and glutamine (Figure 85).
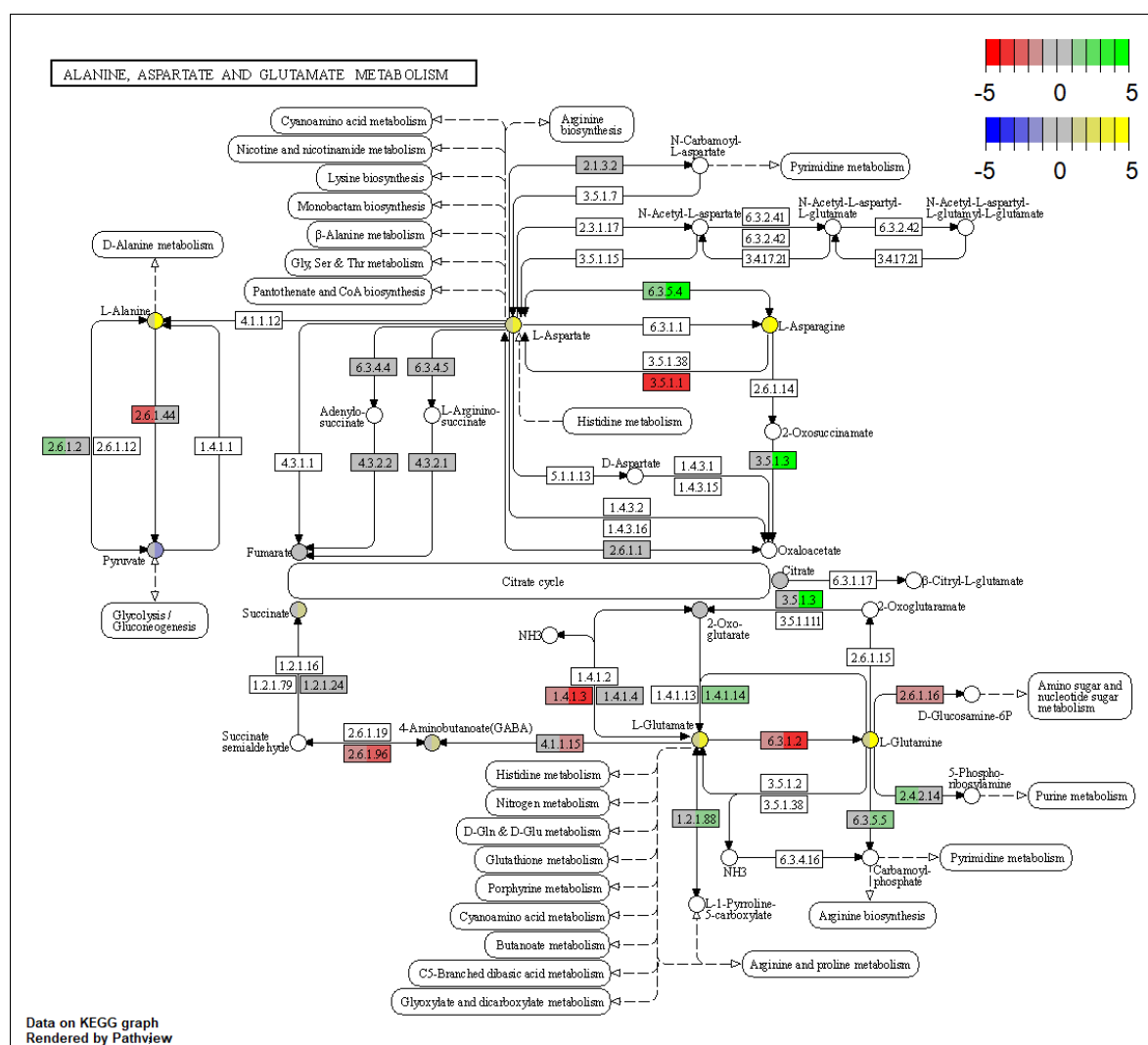
Figure 85 Pathview image of alanine, aspartate and glutamate metabolism. Metabolites are shown as circles and genes are shown as rectangles. Metabolite and gene nodes are split in the middle to simultaneously show two comparisons. Changes are the log2-fold change of mutant green leaves (left) or mutant white leaves (right) in comparison to wild type green leaves. Blank nodes could either be not mapped or are not known.

Valine, leucine and isoleucine are all increased in white *cana1* leaves, while green *cana1* leaves show wild type levels (Figure 86). Transcripts for biosynthetic genes are upregulated in both mutant samples, although in several cases stronger in white leaves.
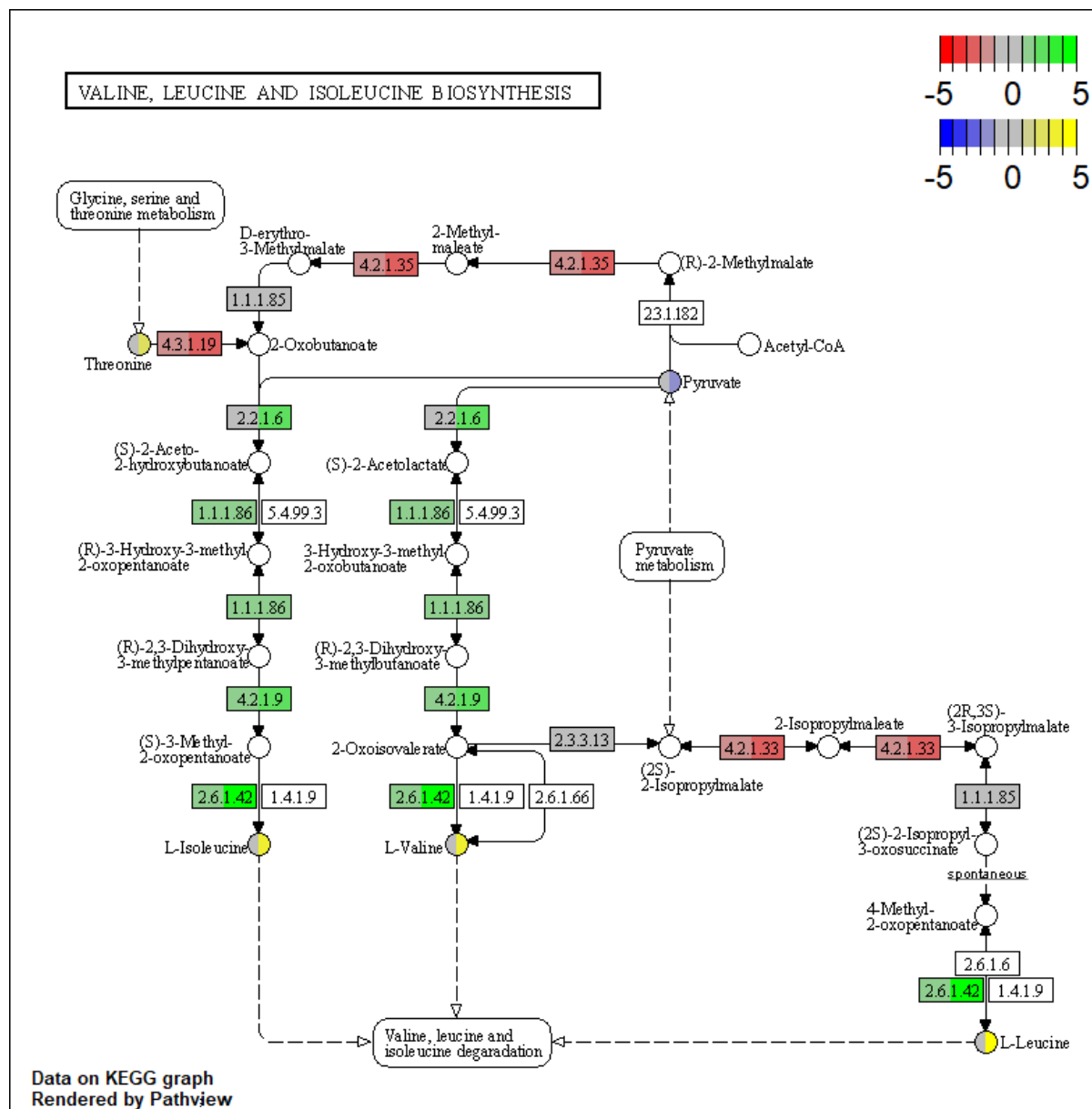
Figure 86: Pathview image valine, leucine and isoleucine metabolism. Metabolites are shown as circles and genes are shown as rectangles. Metabolite and gene nodes are split in the middle to simultaneously show two comparisons. Changes are the log2-fold change of mutant green leaves (left) or mutant white leaves (right) in comparison to wild type green leaves. Blank nodes could either be not mapped or are not known.

4.4. Discussion

In this chapter I characterized the variegation mutant *cana1* that had been shown to display less robustness to yield in a drought stress scenario. The leaf phenotype ranges from completely white leaves over varying degrees of variegation to completely green leaves. These results are similar to the phenotype the *ljsco2* mutants in *Lotus japonicus*, which also displays variegation

on true leaves, but are in contrast to the phenotypes of mutants in *Arabidopsis thaliana* where only cotyledons are pale but not variegated (Zagari et al., 2017). Unsurprisingly white leaves of *cana1* plants were almost completely devoid of chlorophyll a and b. Green leaves of mutant plants however had wild type level photosynthetic pigments.

As expected RNAseq results yielded many gene ontology terms related to photosynthesis enriched, but also terms related to transcription and translation. A deeper investigation of photosynthetic genes, shows that while white mutant leaves show wild-type levels of transcription of several photosystem assembly factors, green leaves show an even higher level. The exception to this are orthologues to ELIPs and DEG7. ELIPs (early light inducible proteins) accumulate during etioplast to chloroplast transition and are suggested to temporarily bind free chlorophylls until antenna proteins are synthesized (Casazza et al., 2005; T. Sun et al., 2019). Since ELIP transcript and protein abundance has been shown to increase under rising light intensity they may also bind chlorophyll released as a result of photoinhibition, which would give them a photoprotective function (Lu, 2016). However, since the light-sensitivity phenotypes of mutant lacking ELIPs appear wild type-like the exact function of these proteins needs to be further investigated (Casazza et al., 2005). In either case due to the importance of ELIPs in early chloroplast development, their upregulation might suggest that white leaf sectors of *cana1* mutant plants get stuck at a pre-thylakoid formation step. Since the stromal DEG7 is known to cleave photodamaged PSII proteins and is therefore suggested to be relevant for PSII repair (X. Sun et al., 2010) its upregulation may be indicative of persisting photodamage in white mutant leaf sectors. The overshoot of transcription in assembly factors in green leaf sectors of *cana1* plants may be a compensation mechanism that finally enables development of properly assembled thylakoids.

Proteomic experiments largely confirm results from RNAseq, with many photosynthesis-related terms as enriched. Further on, I can find almost no difference between green M82 leaves and green *cana1* leaves on a proteomic level. This fits to our theory, that the overcompensation of transcription allows normal thylakoid biogenesis, through adequate levels of photosystem proteins. The mutation in the CANA1 gene is located close to but not within the conserved zinc finger domain, which has been shown crucial for its catalytic activity (Muranaka et al., 2012). It seems therefore likely that the mutant protein retains some functionality and if enough total catalytic activity is available proper thylakoids can be assembled. This fits to a threshold model, proposed in the *immutans* (IM) variegation mutant (D. Wu et al., 1999). This model suggested a threshold of IM activity, needed for chloroplast biogenesis. However, due to the phenotype

of null *im* plants, which are still variegated and not albino, the model was updated to be dependent on a threshold of electron transport (D. Wu et al., 1999). A similar model suggests a threshold of FtsH complexes needed for normal chloroplast development (Aluru et al., 2006). In a similar way it could be possible that enough CANA1 activity is needed to build functioning thylakoids and I have reason to believe that this activity cannot be easily be compensated for by another protein. First of all the only orthologue of CANA1 seems to be a pseudogene. Secondly I could not find a clear candidate in the transcript or protein analysis of green leaf sectors of *cana1* plants that would explain a compensation of CANA1 activity in these cells. Lastly, although anecdotally at this point, I may add that I have another allelic version of the same gene, which carries a point mutation within the conserved zinc finger domain and displays an even more severe phenotype. I also know that our collaborator generated several CRISPR-mediated knock-out lines, one of which showed pale cotyledons but did not survive the seedling stage, suggesting that the protein is crucial for chloroplast development.

In the metabolomics analysis I found a strongly changed metabolite profile in white leaves of mutant plants. I found several amino acids, like asparagine strongly induced in white leaves. Asparagine has been associated to several stress responses before (Lea et al., 2007). Tomato leaves accumulated large quantities of asparagine after bacterial infection (Pérez-García et al., 1998). Meanwhile glucose pools were strongly depleted and transcripts for catabolic enzymes leading to its production were upregulated. As I would normally expect mature leaves to act as source organs, supplying carbon to other parts of the plant (Chang & Zhu, 2017), I could classify white leaves as sinks for carbohydrates, which seems plausible given their inability to photosynthesize properly. The change of the lipidomic profile is mostly characterized by a decrease of MGDGs, DGDGs and TAGs. The decrease of MGDGs and DGDGs can easily be explained by the fact that they make up large fractions of thylakoid membranes (Kobayashi, 2016), which are likely abnormal in white leaves of mutant plants. In connection with that, the accumulation of phospholipids seems plausible, as MGDGs and DGDGs are synthesized over DAG, which is partly generated from phospholipids (J. Li et al., 2020). Therefore the drastic shift of the metabolic profile is likely an effect of the general phenotype and may be indicative of a source-sink shift and a general stress response.

In summary I can say that the CANA1 protein is crucial for chloroplast biogenesis in tomato and likely acts in a similar way as has been found for SCO2 in *Arabidopsis thaliana*. The general difference between *Arabidopsis thaliana* and other plants like *Lotus japonicus* as well as tomato as I described it here, is that mutants show pale cotyledons and variegated true leaves

respectively (Zagari et al., 2017). In *Arabidopsis thaliana* it has been suggested that this effect is caused by two separate pathways for thylakoid formation; a SCO2-dependent transport vesicle-mediated pathway predominantly active during germination and a SRP proteins-dependent pathway predominantly active in developed leaves (Tanz et al., 2012). In conclusion I can hypothesize that in tomato the vesicle-mediated pathway is still needed for thylakoid formation in mature true leaves.

Unfortunately I cannot contribute much mechanistic understanding to the exact function of the protein disulfide isomerase investigated here and the in-depth investigation may exceed our specialized expertise on metablism.I am still also interested in how exactly the leaf variegation has caused the observed yield destabilization. I could hypothesize that the variegation pattern, which seems random at first sight may be actually influenced by external factors during the development of chloroplasts. Given that newly synthesized complexes of PSII seem to be especially vulnerable to photodamage in comparison to mature photosystems (Shevela et al., 2019), it may be possible that other abiotic factors like for example drought stress could destabilize photosystem assembly.

Despite the fact that the mutation of CANA1 of course truly affects yield canalization, in respect to drought stress one may still be cautious in calling CANA1 a true canalization gene. As pointed out previously, mutants may fail to construct a certain phenotype, which leads to a higher variance, but does not point to a mechanism for robustness (Félix & Barkoulas, 2015). Then again, it was precisely the fact that developmental processes seem to bring about a definite final state, that made Waddington to consider them canalized (Waddington, 1942). Either way, it is true that the total amount of photosynthetically active leaf surface area is always dependent on the concrete pattern and distribution of variegation and shows some plant-to-plant variation. In conclusion, I could say that a lack of CANA1 leads to a destabilization of the mean level of photosynthetically active area across the whole leaf surface area, which likely manifests in the observed yield variability. To test if an increased abundance of CANA1 can confer even stronger robustness to yield under normal or stress conditions, I generated transgenic tomato lines overexpressing CANA1 and are currently investigating them.

Chloroplast biogenesis is a complex process influenced by many different factors. As embryos develop into seedling, light is perceived by photoreceptors and organelles need to coordinate, one must consider temporal, cellular and environmental factors that all affect chloroplast biogenesis (Pogson & Albrecht, 2011). Due to this multi-factorial scenario likely a higher

temporal and spatial resolution of the cellular processes is needed to get a better understanding of the exact mechanism of chloroplast biogenesis.

Chapter 5: General discussion

In this work, I have applied multiple approaches to investigate canalization of metabolism and yield. Previous chapters have described them individually and here I am discussing the results in synthesis.

In the first project I calculated the CV of metabolic abundance in leaf tissue of *Arabidopsis thaliana* and *Phaseolus vulgaris* across different conditions. In the datasets for primary metabolites of *Arabidopsis thaliana* I generally found lower CVs for metabolites than for secondary metabolites of *Phaseolus vulgaris*, which may be a characteristic of secondary metabolism, but may also include a species and condition effect. CVs of closely related primary metabolites showed moderate to high correlation values. When using the CV as input for the mixed linear model in a genome wide association approach, I find several canalized metabolic QTL. Several candidate genes with known regulatory roles or unknown function, which may ultimately prove to be of a regulatory nature can be found. However, several metabolic enzymes that are either directly, more distantly or apparently not at all related to the respective metabolites can also be found. Together with the correlation pattern of metabolite CVs this may point to an inherent role of metabolism to its own stability.

In an attempt to validate putative candidates for canalization found in a cmQTL mapping of metabolite variation of tomato pericarp across three seasons, I created CRISPR/Cas9-mediated gene-edited tomato lines. While several plants show no apparent abnormal phenotype, other lines show strong alterations in leaf pigmentation, fruit size and seed number or additional meristems. Unfortunately I was so far not able to generate enough offspring of the latter, to test the variation of metabolism across different conditions, but in comparison to wild type plants they already show strong metabolic shifts under standard conditions. Under the morphological inconspicuous mutants one line which has a gene-edit in pantothenate kinase, chosen as a putative candidate gene for malate variation, indeed shows a higher CV of its target metabolite across different watering conditions. Another mutant of a candidate gene for phenylalanine variation, the orthologue of which is known to be involved in amino acid transport in *Arabidopsis thaliana*, shows some tendency to an increase of phenylalanine CV but no statistically significant results. With both of these genes related to production or transport of metabolites, our second approach also points to at least some relevance of metabolism for its own stability.

Chapter 5: General discussion

In a third project I investigated an EMS tomato mutant with a variegated leaf phenotype, previously shown to have a reduced yield stability in response to drought stress. The orthologous gene is known from *Arabidopsis thaliana*, where it has been characterized as a protein disulfide isomerase crucial for thylakoid biogenesis. When comparing mutant green leaves to wild type green leaves, I can see an upregulation of transcripts of photosystem assembly factors in the former, while the proteomic profile looks wild type-like. I therefore hypothesize the need for a compensation for a partially impaired enzyme. If assembly factors are upregulated, cells may surpass a threshold of enzyme activity or stable protein complexes needed for thylakoid biogenesis and can develop proper chloroplasts. Similar versions of a threshold model have been suggested for other variegation mutants (Aluru et al., 2006; D. Wu et al., 1999). Although metabolism is strongly altered in white sectors of mutant leaves, it is likely an effect of the phenotype, which may alter lipid abundance through abnormal thylakoid membranes and carbon metabolism as an effect of their inability to photosynthesize. As to the yield stability sensitive to drought stress, I can hypothesize it as a function of the variation of total photosynthetically active leaf area. I have yet to design a proper experiment to test this hypothesis.

Based on our results here I can hypothesize that besides previously anticipated regulatory genes (Alseekh et al., 2017), also enzymes or metabolism itself may play a significant role in canalization of metabolism. Together with other features, such robust architectures have been suggested as a key characteristic of robust biological networks and several examples are known (Whitacre, 2012). For example, in gene networks it is known that connectivity of network nodes gives rise to network robustness with most nodes of low connectivity and a few so-called hub nodes with high connectivity (Lachowiec et al., 2016). Similarly, it has been shown that metabolic networks, have many metabolites with low connectivity and a few hub metabolites, like ATP and NADPH, which participate in many reactions (Pfeiffer et al., 2005). Additionally, this metabolic network robustness is mediated by another feature of biological robustness, namely redundancy, which in metabolic networks occurs on two levels, via duplicate genes and alternative pathways (Blank et al., 2005; Whitacre, 2012).

I should admit at this point that because of my specialized focus on and familiarity with metabolism, I may be biased towards finding metabolic genes as causative for metabolic canalization and cannot exclude, that I may have overlooked some genes of a specific function that could explain the variance. It would be interesting to direct my attention towards potential regulatory genes, that were originally found in the tomato cmQTL mapping and hold the

potential to increase metabolic robustness, without affecting the metabolite level (Alseekh et al., 2017), but which I have difficulties detecting here. As mentioned earlier a clear difference between this original approach and the cmGWAS approach presented here is the level and character of the variation. Although genetic variation within a population, which can be exploited by GWAS is suitable to connect genetic to phenotypic variation (Alonso-Blanco et al., 2016), it seems plausible that it can only utilize genetic variation already present in the population. Mapping populations such as the ILs used in the cmQTL mapping carry genetic variation introduced from the wild *Solanum penellii* relative of tomato (Eshed & Zamir, 1995), may however introduce novel genetic variation. In conclusion it could be possible that, although I do see some variation of metabolite CV between ecotypes in the GWAS dataset, it may only be explained by genetic variation residing on rare alleles. As I filtered out all alleles with a minor allele frequency lower than 0.05, in order to prevent additional confounding effects known to be caused by such alleles (Brachi et al., 2011), it is possible, that I missed a few interesting true positives. The observation that mean values of CVs were generally low for most primary metabolites in the GWAS datasets of *Arabidopsis thaliana* and also the data from the cmQTL validation experiment of tomato fruits and leaves, could also point to the fact that metabolism is inherently canalized across different environments. Following that hypothesis, it also seems plausible that some sort of disruption, could decanalize metabolism and enable finding the genetic determinants, like it was shown for the tomato fruit metabolism (Alseekh et al., 2017). Both for the cmQTL mapping in tomato fruits and for our cmGWAS approaches it may be that any results found in such mapping approaches may be species or tissue-specific and thus I may not be able to directly compare them to results from other organisms and organs (Alseekh et al., 2017). Fruits are for example generally considered a sink-tissue while mature leaves can be considered a source-tissue (Chang & Zhu, 2017), so metabolism could be differently canalized in these tissues. Then again, when reconsidering the cmQTL mapping in tomato, I should note that a few metabolic genes were found, which are directly related to their target compound, like phenylalanine ammonia lyase in the constrained cmQTL region of phenylalanine (Alseekh et al., 2017). In conclusion the presence of enzymatic genes directly or more distantly related to the respective target compound, emerges as a common feature at least for *Arabidopsis thaliana* and common bean leaves as well as tomato fruits.

Regarding yield canalization, the gene I characterized here is important for the construction of a regular green leaf phenotype. As mentioned earlier there is no clear consensus in the community, whether mutants with such pleiotropic effects point to a robustness mechanism,

although in my own opinion it goes along the original canalization concept of Waddington, highlighting any developmental phenotype as normally canalized under wild type conditions (Waddington, 1942). Although I did not investigate the variation of metabolism across conditions in the *cana1* mutant, metabolite levels were strongly altered, predominantly in the white leaf sectors. This alteration is most likely an indirect effect of the phenotype. It has been suggested however that hub genes may affect an organisms overall plasticity (Laitinen & Nikoloski, 2018). One such example of a potential hub gene is the molecular chaperone HSP90, which has been found to affect several morphological phenotypes in different eukaryotic species (Sangster et al., 2008). Since this highly connected and evolutionarily conserved protein keeps other proteins poised in a metastable state, it seems plausible, that the lack of it leads to an increase in phenotypic variation (Lachowiec et al., 2016; Queitsch et al., 2002). In conclusion, it could also be that the genes, which resulted in a strong phenotype when edited in the cmQTL validation experiment, may have some relevance for metabolic canalization, although I could not yet test this. Either way it may be difficult to disentangle cause and effect in such pleiotropic phenotypes. It has been suggested to investigate mean and variance along dose-response curves to interpret changes in variance (Félix & Barkoulas, 2015). To achieve different dosages of gene-expression, I could apply a similar CRISPR/Cas9 strategy as employed here, but instead of targeting the exons to create loss-of-function mutants, it may be possible to target the promoter region, to generate knock-down mutants with varying degree of transcript abundance (Elison et al., 2017). However, even more sophisticated methods could be used, which modulate gene expression by altering methylation status of promoter regions via a modified Cas9-system (Kang et al., 2008). Similarly, it would be interesting to see the effects of overexpressing genes with a putative function in canalization, to see if it can increase trait canalization beyond wild type levels. For the CANA1 gene I have generated overexpression lines and will soon analyze the first results.

Of course it would be also interesting to explore possibilities of using other metrics suggested by literature (Laitinen & Nikoloski, 2018) to estimate variation across different environments. I was using the coefficient of variation here. Similarly, so-called mean-scaled variance, which is equal to the square of the CV can also be used to estimate variation (Pélabon et al., 2011). Although the CV has been suggested as an appropriate measure to estimate variation there may be confounding effects of the mean (Laitinen & Nikoloski, 2018). Also as I saw here, the CV is limited by the square-root of the number of values it is calculated over. In the specific case of the cmQTL validation it may be possible to also use the Levene's transformed metabolite

values to test for variation as it was done in the study leading to cmQTLs (Alseekh et al., 2017). Further on Finlay-Wilkinson regression has been suggested and is already being used to study plasticity of plant phenotypes (Laitinen & Nikoloski, 2018; N. Liu et al., 2021).

5.1. Conclusion

The study of canalization or variation is a very interesting and important but also challenging endeavor, as I was hopefully able to show in this work. I gathered evidence pointing to both regulatory genes but even more obviously metabolic genes acting directly or indirectly towards canalization of metabolism. The relevance of metabolism for its own stability fits to previous observations of stability mediated by network structure and architecture. Canalization of yield may be mediated by a proper development of photosynthetic tissue. Although for metabolism and yield stability, further work needs to be done to confirm and mechanistically explain these theories, I hope that my work has made a small contribution towards a better understanding of trait canalization.

References

Abdelhaleem, M. (2010). Helicases: An Overview. In M. M. Abdelhaleem (Ed.), *Helicases: Methods and Protocols* (pp. 1–12). Humana Press. https://doi.org/10.1007/978-1-60327-355-8_1

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nature Communications*, *9*(1), 1911. https://doi.org/10.1038/s41467-018-04252-2

Albrecht, V., Ingenfeld, A., & Apel, K. (2008). Snowy cotyledon 2: The identification of a zinc finger domain protein essential for chloroplast development in cotyledons but not in true leaves. *Plant Mol Biol*, *66*(6), 599–608. https://doi.org/10.1007/s11103-008-9291-y

Alexa, A., & Rahnenfuhrer, J. (2021). *topGO: Enrichment Analysis for Gene Ontology* (2.44.0) [Computer software]. Bioconductor version: Release (3.13). https://doi.org/10.18129/B9.bioc.topGO

Alon, U., Surette, M. G., Barkai, N., & Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, *397*(6715), 168–171. https://doi.org/10.1038/16483

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C. C., Parker, H., Prednis, L., Ansari, Y., Choy, N., … Ecker, J. R. (2003). Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana. *Science*, *301*(5633), 653–657. https://doi.org/10.1126/science.1086391

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491. https://doi.org/10.1016/j.cell.2016.05.063

References

Alseekh, S., & Fernie, A. R. (2018). Metabolomics 20 years on: What have we learned and what hurdles remain? *The Plant Journal*, *94*(6), 933–942. https://doi.org/10.1111/tpj.13950

Alseekh, S., Kostova, D., Bulut, M., & Fernie, A. R. (2021). Genome-wide association studies: Assessing trait characteristics in model and crop plants. *Cellular and Molecular Life Sciences*, *78*(15), 5743–5754. https://doi.org/10.1007/s00018-021-03868-w

Alseekh, S., Tohge, T., Wendenberg, R., Scossa, F., Omranian, N., Li, J., Kleessen, S., Giavalisco, P., Pleban, T., Mueller-Roeber, B., Zamir, D., Nikoloski, Z., & Fernie, A. R. (2015). Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *Plant Cell*, *27*(3), 485–512. https://doi.org/10.1105/tpc.114.132266

Alseekh, S., Tong, H., Scossa, F., Brotman, Y., Vigroux, F., Tohge, T., Ofner, I., Zamir, D., Nikoloski, Z., & Fernie, A. R. (2017). Canalization of Tomato Fruit Metabolism. *Plant Cell*. https://doi.org/10.1105/tpc.17.00367

Alseekh, S., Wu, S., Brotman, Y., & Fernie, A. R. (2018). Guidelines for Sample Normalization to Minimize Batch Variation for Large-Scale Metabolic Profiling of Plant Natural Genetic Variance. *Methods in Molecular Biology (Clifton, N.J.)*, *1778*, 33–46. https://doi.org/10.1007/978-1-4939-7819-9_3

Aluru, M. R., Yu, F., Fu, A., & Rodermel, S. (2006). Arabidopsis variegation mutants: New insights into chloroplast biogenesis. *Journal of Experimental Botany*, *57*(9), 1871–1881. https://doi.org/10.1093/jxb/erj008

Anwar, R., Fatima, T., & Mattoo, A. K. (2019, October 30). *Tomatoes: A Model Crop of Solanaceous Plants*. Oxford Research Encyclopedia of Environmental Science. https://doi.org/10.1093/acrefore/9780199389414.013.223

References

Arouisse, B., Korte, A., van Eeuwijk, F., & Kruijer, W. (2020). Imputation of 3 million SNPs in the Arabidopsis regional mapping population. *The Plant Journal*, *102*(4), 872–882. https://doi.org/10.1111/tpj.14659

Asekova, S., Oh, E., Kulkarni, K. P., Siddique, M. I., Lee, M. H., Kim, J. I., Lee, J.-D., Kim, M., Oh, K.-W., Ha, T.-J., Kim, S.-U., & Cho, K.-S. (2021). An Integrated Approach of QTL Mapping and Genome-Wide Association Analysis Identifies Candidate Genes for Phytophthora Blight Resistance in Sesame (Sesamum indicum L.). *Frontiers in Plant Science*, *12*, 67. https://doi.org/10.3389/fpls.2021.604709

At3g31900—ATP-dependent helicase family protein—Arabidopsis thaliana (Mouse-ear cress)—At3g31900 gene & protein. (n.d.). Retrieved September 22, 2021, from https://www.uniprot.org/uniprot/F4J923

Atkinson, N. J., Lilley, C. J., & Urwin, P. E. (2013). Identification of Genes Involved in the Response of Arabidopsis to Simultaneous Biotic and Abiotic Stresses1[C][W][OPEN]. *Plant Physiology*, *162*(4), 2028–2041. https://doi.org/10.1104/pp.113.222372

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Baslam, M., Mitsui, T., Hodges, M., Priesack, E., Herritt, M. T., Aranjuelo, I., & Sanz-Sáez, Á. (2020). Photosynthesis in a Changing Global Climate: Scaling Up and Scaling Down in Crops. *Frontiers in Plant Science*, *11*, 882. https://doi.org/10.3389/fpls.2020.00882

Bauer, C. R., Li, S., & Siegal, M. L. (2015). Essential gene disruptions reveal complex relationships between phenotypic robustness, pleiotropy, and fitness. *Molecular Systems Biology*, *11*(1), 773-n/a. https://doi.org/10.15252/msb.20145264

Bellucci, E., Benazzo, A., Xu, C., Bitocchi, E., Rodriguez, M., Alseekh, S., Di Vittori, V., Gioia, T., Neumann, K., Cortinovis, G., Frascarelli, G., Murube, E., Trucchi, E., Nanni,

References

L., Ariana, A., Logozzo, G., Shin, J. H., Chaohchih, L., Jiang, L., … Papa, R. (in preparation). *SELECTION AND ADAPTIVE INTROGRESSION GUIDED THE COMPLEX EVOLUTIONARY HISTORY OF THE EUROPEAN COMMON BEANS*.

Ben-Amar, A., Daldoul, S., Reustle, G. M., Krczal, G., & Mliki, A. (2016). Reverse Genetics and High Throughput Sequencing Methodologies for Plant Functional Genomics. *Current Genomics*, *17*(6), 460–475. https://doi.org/10.2174/1389202917666160520102827

Benke, K., & Tomkins, B. (2017). Future food-production systems: Vertical farming and controlled-environment agriculture. *Sustainability: Science, Practice and Policy*, *13*(1), 13–26. https://doi.org/10.1080/15487733.2017.1394054

Bhattacharya, D., & Meir, E. G. V. (2019). A simple genotyping method to detect small CRISPR-Cas9 induced indels by agarose gel electrophoresis. *Scientific Reports*, *9*(1), 4437. https://doi.org/10.1038/s41598-019-39950-4

Bitto, E., Bingman, C. A., Allard, S. T. M., Wesenberg, G. E., & Phillips, G. N. (2005). The structure at 1.7 Å resolution of the protein product of the At2g17340 gene from Arabidopsis thaliana. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, *61*(7), 630–635. https://doi.org/10.1107/S1744309105017690

Blank, L. M., Kuepfer, L., & Sauer, U. (2005). Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biology*, *6*(6), R49. https://doi.org/10.1186/gb-2005-6-6-r49

Blume, C., Ost, J., Mühlenbruch, M., Peterhänsel, C., & Laxa, M. (2019). Low CO2 induces urea cycle intermediate accumulation in Arabidopsis thaliana. *PLOS ONE*, *14*(1), e0210342. https://doi.org/10.1371/journal.pone.0210342

References

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: The missing heritability is in the field. *Genome Biology*, *12*(10), 232. https://doi.org/10.1186/gb-2011-12-10-232

Breitel, D., Brett, P., Alseekh, S., Fernie, A. R., Butelli, E., & Martin, C. (2021). Metabolic engineering of tomato fruit enriched in L-DOPA. *Metabolic Engineering*, *65*, 185–196. https://doi.org/10.1016/j.ymben.2020.11.011

Brooks, C., Nekrasov, V., Lippman, Z. B., & Van Eck, J. (2014). Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. *Plant Physiol*, *166*(3), 1292–1297. https://doi.org/10.1104/pp.114.247577

Broughton, W. J., Hernández, G., Blair, M., Beebe, S., Gepts, P., & Vanderleyden, J. (2003). Beans (Phaseolus spp.) – model food legumes. *Plant and Soil*, *252*(1), 55–128. https://doi.org/10.1023/A:1024146710611

Bugos, R. C., Chiang, V. L., Zhang, X. H., Campbell, E. R., Podila, G. K., & Campbell, W. H. (1995). RNA isolation from plant tissues recalcitrant to extraction in guanidine. *Biotechniques*, *19*(5), 734–737.

Burghardt, L. T., Young, N. D., & Tiffin, P. (2017). A Guide to Genome-Wide Association Mapping in Plants. *Current Protocols in Plant Biology*, *2*(1), 22–38. https://doi.org/10.1002/cppb.20041

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

References

Butelli, E., Titta, L., Giorgio, M., Mock, H.-P., Matros, A., Peterek, S., Schijlen, E. G. W. M., Hall, R. D., Bovy, A. G., Luo, J., & Martin, C. (2008). Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. *Nature Biotechnology*, *26*(11), 1301–1308. https://doi.org/10.1038/nbt.1506

Campos, M. L., Carvalho, R. F., Benedito, V. A., & Pereira Peres, L. E. (2010). Small and remarkable. *Plant Signaling & Behavior*, *5*(3), 267–270.

Canady, M. A., Meglic, V., & Chetelat, R. T. (2005). A library of Solanum lycopersicoides introgression lines in cultivated tomato. *Genome*, *48*(4), 685–697. https://doi.org/10.1139/g05-032

Casazza, A. P., Rossini, S., Rosso, M. G., & Soave, C. (2005). Mutational and expression analysis of ELIP1 and ELIP2 in Arabidopsis thaliana. *Plant Molecular Biology*, *58*(1), 41–51. https://doi.org/10.1007/s11103-005-4090-1

Chang, T.-G., & Zhu, X.-G. (2017). Source–sink interaction: A century old concept under the light of modern molecular systems biology. *Journal of Experimental Botany*, *68*(16), 4417–4431. https://doi.org/10.1093/jxb/erx002

Chen, K., Wang, Y., Zhang, R., Zhang, H., & Gao, C. (2019). CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. *Annual Review of Plant Biology*, *70*(1), 667–697. https://doi.org/10.1146/annurev-arplant-050718-100049

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*. https://doi.org/10.1186/s13059-016-0881-8

Cortinovis, G., Oppermann, M., Neumann, K., Graner, A., Gioia, T., Marsella, M., Alseekh, S., Fernie, A. R., Papa, R., Bellucci, E., & Bitocchi, E. (2021). Towards the Development, Maintenance, and Standardized Phenotypic Characterization of Single-

References

Seed-Descent Genetic Resources for Common Bean. *Current Protocols*, *1*(5), e133. https://doi.org/10.1002/cpz1.133

Coxon, K. M., Chakauya, E., Ottenhof, H. H., Whitney, H. M., Blundell, T. L., Abell, C., & Smith, A. G. (2005). Pantothenate biosynthesis in higher plants. *Biochemical Society Transactions*, *33*(4), 743–746. https://doi.org/10.1042/BST0330743

Cross, J. M., von Korff, M., Altmann, T., Bartzetko, L., Sulpice, R., Gibon, Y., Palacios, N., & Stitt, M. (2006). Variation of Enzyme Activities and Metabolite Levels in 24 Arabidopsis Accessions Growing in Carbon-Limited Conditions. *Plant Physiology*, *142*(4), 1574–1588. https://doi.org/10.1104/pp.106.086629

Curtin, S. J., Tiffin, P., Guhlin, J., Trujillo, D. I., Burghardt, L. T., Atkins, P., Baltes, N. J., Denny, R., Voytas, D. F., Stupar, R. M., & Young, N. D. (2017). Validating Genome-Wide Association Candidates Controlling Quantitative Variation in Nodulation. *Plant Physiology*, *173*(2), 921–931. https://doi.org/10.1104/pp.16.01923

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. *Plant Physiology*, *152*(2), 579–589. https://doi.org/10.1104/pp.109.148817

Delwiche, C. F., & Cooper, E. D. (2015). The Evolutionary Origin of a Terrestrial Flora. *Current Biology*, *25*(19), R899–R910. https://doi.org/10.1016/j.cub.2015.08.029

Diaz, M. (2019). Arabidopsis Thaliana: From Weed to Model Organism. *Current Protocols Essential Laboratory Techniques*, *19*(1), e38. https://doi.org/10.1002/cpet.38

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J. M., Colot, V., & Quadrana, L. (2020). The impact of transposable elements on tomato diversity. *Nature Communications*, *11*(1), 4058. https://doi.org/10.1038/s41467-020-17874-2

References

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S., … Schwen, B. (2021). *data.table: Extension of "data.frame"* (1.14.0) [Computer software]. https://CRAN.R-project.org/package=data.table

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., & Lepiniec, L. (2010). MYB transcription factors in Arabidopsis. *Trends in Plant Science*, *15*(10), 573–581. https://doi.org/10.1016/j.tplants.2010.06.005

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J. D., Halsall, A., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Kell, D. B., & Goodacre, R. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *6*(7), 1060–1083. https://doi.org/10.1038/nprot.2011.335

Eaton-Rye, J. J., & Sobotka, R. (2017). Editorial: Assembly of the Photosystem II Membrane-Protein Complex of Oxygenic Photosynthesis. *Frontiers in Plant Science*, *8*, 884. https://doi.org/10.3389/fpls.2017.00884

Elison, G. L., Song, R., & Acar, M. (2017). A precise genome editing method reveals insights into the activity of eukaryotic promoters. *Cell Reports*, *18*(1), 275–286. https://doi.org/10.1016/j.celrep.2016.12.014

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

References

Eshed, Y., & Zamir, D. (1994). A genomic library of Lycopersicon pennellii in L. esculentum: A tool for fine mapping of genes. *Euphytica*, *79*(3), 175–179. https://doi.org/10.1007/bf00022516

Eshed, Y., & Zamir, D. (1995). An Introgression Line Population of Lycopersicon Pennellii in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated Qtl. *Genetics*, *141*(3), 1147–1162.

Fan, S., Kind, T., Cajka, T., Hazen, S. L., Tang, W. H. W., Kaddurah-Daouk, R., Irvin, M. R., Arnett, D. K., Barupal, D. K., & Fiehn, O. (2019). Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Analytical Chemistry*, *91*(5), 3590–3596. https://doi.org/10.1021/acs.analchem.8b05592

Fang, C., & Luo, J. (2019). Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *The Plant Journal*, *97*(1), 91–100. https://doi.org/10.1111/tpj.14097

Fei, Z., Joung, J.-G., Tang, X., Zheng, Y., Huang, M., Lee, J. M., McQuinn, R., Tieman, D. M., Alba, R., Klee, H. J., & Giovannoni, J. J. (2011). Tomato Functional Genomics Database: A comprehensive resource and analysis package for tomato functional genomics. *Nucleic Acids Research*, *39*(suppl_1), D1156–D1163. https://doi.org/10.1093/nar/gkq991

Félix, M.-A., & Barkoulas, M. (2015). Pervasive robustness in biological systems. *Nature Reviews Genetics*, *16*, 483. https://doi.org/10.1038/nrg3949

Fenwick, D. E., & Oakenfull, D. (1983). Saponin content of food plants and some prepared foods. *Journal of the Science of Food and Agriculture*, *34*(2), 186–191. https://doi.org/10.1002/jsfa.2740340212

Fernie, A. R., & Tohge, T. (2017). The Genetics of Plant Metabolism. *Annual Review of Genetics*, *51*(1), 287–310. https://doi.org/10.1146/annurev-genet-120116-024640

References

Ferreira, H., Pinto, E., & Vasconcelos, M. W. (2021). Legumes as a Cornerstone of the Transition Toward More Sustainable Agri-Food Systems and Diets in Europe. *Frontiers in Sustainable Food Systems*, *5*, 280. https://doi.org/10.3389/fsufs.2021.694121

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, *281*(5374), 237–240. https://doi.org/10.1126/science.281.5374.237

Fisher, J., Bensal, E., & Zamir, D. (2017). Bimodality of stable and plastic traits in plants. *Theor Appl Genet*, *130*(9), 1915–1926. https://doi.org/10.1007/s00122-017-2933-1

Fulop, D., Ranjan, A., Ofner, I., Covington, M. F., Chitwood, D. H., West, D., Ichihashi, Y., Headland, L., Zamir, D., Maloof, J. N., & Sinha, N. R. (2016). A New Advanced Backcross Tomato Population Enables High Resolution Leaf QTL Mapping and Gene Identification. *G3: Genes|Genomes|Genetics*, *6*(10), 3169–3184. https://doi.org/10.1534/g3.116.030536

Fusari, C. M., Kooke, R., Lauxmann, M. A., Annunziata, M. G., Enke, B., Hoehne, M., Krohn, N., Becker, F. F. M., Schlereth, A., Sulpice, R., Stitt, M., & Keurentjes, J. J. B. (2017). Genome-Wide Association Mapping Reveals That Specific and Pleiotropic Regulatory Mechanisms Fine-Tune Central Metabolism and Growth in Arabidopsis[OPEN]. *The Plant Cell*, *29*(10), 2349–2373. https://doi.org/10.1105/tpc.17.00232

Gao, T., Wang, Y., Liu, Y., Ma, M., Li, X., Zhang, D., Ding, K., Li, C., Zou, Y., & Ma, F. (2021). Overexpression of tyrosine decarboxylase (MdTYDC) enhances drought tolerance in Malus domestica. *Scientia Horticulturae*, *289*, 110425. https://doi.org/10.1016/j.scienta.2021.110425

References

García-Fernández, C., Campa, A., Garzón, A. S., Miklas, P., & Ferreira, J. J. (2021). GWAS of pod morphological and color characters in common bean. *BMC Plant Biology*, *21*(1), 184. https://doi.org/10.1186/s12870-021-02967-x

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H., & Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, *468*(7320), 67–71. https://doi.org/10.1038/nature09523

Gebbie, L. K., Burn, J. E., Hocart, C. H., & Williamson, R. E. (2005). Genes encoding ADP-ribosylation factors in Arabidopsis thaliana L. Heyn.; genome analysis and antisense suppression. *Journal of Experimental Botany*, *56*(414), 1079–1091. https://doi.org/10.1093/jxb/eri099

Giavalisco, P., Köhl, K., Hummel, J., Seiwert, B., & Willmitzer, L. (2009). 13C Isotope-Labeled Metabolomes Allowing for Improved Compound Annotation and Relative Quantification in Liquid Chromatography-Mass Spectrometry-based Metabolomic Research. *Analytical Chemistry*, *81*(15), 6546–6551. https://doi.org/10.1021/ac900979e

Gibon, Y., Blaesing, O. E., Hannemann, J., Carillo, P., Höhne, M., Hendriks, J. H. M., Palacios, N., Cross, J., Selbig, J., & Stitt, M. (2004). A Robot-Based Platform to Measure Multiple Enzyme Activities in Arabidopsis Using a Set of Cycling Assays: Comparison of Changes of Enzyme Activities and Transcript Levels during Diurnal Cycles and in Prolonged Darkness. *The Plant Cell*, *16*(12), 3304–3325. https://doi.org/10.1105/tpc.104.025973

Gibon, Y., Bläsing, O. E., Palacios-Rojas, N., Pankovic, D., Hendriks, J. H. M., Fisahn, J., Höhne, M., Günther, M., & Stitt, M. (2004). Adjustment of diurnal starch turnover to short days: Depletion of sugar during the night leads to a temporary inhibition of

References

carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period. *The Plant Journal*, *39*(6), 847–862. https://doi.org/10.1111/j.1365-313X.2004.02173.x

Gierlinski, M., Gastaldello, F., Cole, C., & Barton, G. J. (2018). *Proteus: An R package for downstream analysis of MaxQuant output* (p. 416511). https://doi.org/10.1101/416511

González-Guzmán, M., Apostolova, N., Bellés, J. M., Barrero, J. M., Piqueras, P., Ponce, M. R., Micol, J. L., Serrano, R., & Rodríguez, P. L. (2002). The Short-Chain Alcohol Dehydrogenase ABA2 Catalyzes the Conversion of Xanthoxin to Abscisic Aldehyde. *The Plant Cell*, *14*(8), 1833–1846. https://doi.org/10.1105/tpc.002477

Gu, M., Liu, Y., Cui, M., Wu, H., & Ling, H.-Q. (2020). Requirement and functional redundancy of two large ribonucleotide reductase subunit genes for cell cycle, chloroplast biogenesis in tomato. *BioRxiv*, 2020.05.18.102301. https://doi.org/10.1101/2020.05.18.102301

Gur, A., Semel, Y., Osorio, S., Friedmann, M., Seekh, S., Ghareeb, B., Mohammad, A., Pleban, T., Gera, G., Fernie, A. R., & Zamir, D. (2011). Yield quantitative trait loci from wild tomato are predominately expressed by the shoot. *Theor Appl Genet*, *122*(2), 405–420. https://doi.org/10.1007/s00122-010-1456-9

Gur, A., & Zamir, D. (2015). Mendelizing all Components of a Pyramid of Three Yield QTL in Tomato. *Frontiers in Plant Science*, *6*, 1096. https://doi.org/10.3389/fpls.2015.01096

Hankamer, B., Barber, J., & Boekema, E. J. (1997). Structure and membrane organization of photosystem ii in green plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, *48*(1), 641–671. https://doi.org/10.1146/annurev.arplant.48.1.641

He, Y., Wu, D., Wei, D., Fu, Y., Cui, Y., Dong, H., Tan, C., & Qian, W. (2017). GWAS, QTL mapping and gene expression analyses in Brassica napus reveal genetic control of

References

branching   morphogenesis.   *Scientific   Reports*,   *7*(1),   15971.
https://doi.org/10.1038/s41598-017-15976-4

Hester,   J.   (2020).   *glue: Interpreted   String   Literals*   (1.4.2)   [Computer   software].
https://CRAN.R-project.org/package=glue

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati,
N. W., Platt, A., Sperone, F. G., Vilhjálmsson, B. J., Nordborg, M., Borevitz, J. O., &
Bergelson, J. (2012). Genome-wide patterns of genetic variation in worldwide
Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, *44*(2), 212–
216. https://doi.org/10.1038/ng.1042

Hosmani, P. S., Flores-Gonzalez, M., Geest, H. van de, Maumus, F., Bakker, L. V., Schijlen,
E., Haarst, J. van, Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni,
J. J., Mueller, L. A., & Saha, S. (2019). *An improved de novo assembly and annotation
of the tomato reference genome using single-molecule sequencing, Hi-C proximity
ligation and optical maps* (p. 767764). https://doi.org/10.1101/767764

Hu, Y., Cheng, Z., Heller, L. I., Krasnoff, S. B., Glahn, R. P., & Welch, R. M. (2006).
Kaempferol in Red and Pinto Bean Seed (Phaseolus vulgaris L.) Coats Inhibits Iron
Bioavailability Using an in Vitro Digestion/Human Caco-2 Cell Model. *Journal of
Agricultural   and   Food   Chemistry*,   *54*(24),   9254–9261.
https://doi.org/10.1021/jf0612981

Hummel, J., Segu, S., Li, Y., Irgang, S., Jueppner, J., & Giavalisco, P. (2011). Ultra
Performance Liquid Chromatography and High Resolution Mass Spectrometry for the
Analysis   of   Plant   Lipids.   *Frontiers   in   Plant   Science*,   *2*,   54.
https://doi.org/10.3389/fpls.2011.00054

References

Hundertmark, M., & Hincha, D. K. (2008). LEA (Late Embryogenesis Abundant) proteins and their encoding genes in Arabidopsis thaliana. *BMC Genomics*, *9*, 118. https://doi.org/10.1186/1471-2164-9-118

Hwang, E.-Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., & Cregan, P. B. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics*, *15*, 1. https://doi.org/10.1186/1471-2164-15-1

Jensen, P. E., Bassi, R., Boekema, E. J., Dekker, J. P., Jansson, S., Leister, D., Robinson, C., & Scheller, H. V. (2007). Structure, function and regulation of plant photosystem I. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1767*(5), 335–352. https://doi.org/10.1016/j.bbabio.2007.03.004

Jimenez-Gomez, J. M., Corwin, J. A., Joseph, B., Maloof, J. N., & Kliebenstein, D. J. (2011). Genomic Analysis of QTLs and Genes Altering Natural Variation in Stochastic Noise. *PLoS Genetics*, *7*(9), e1002295. https://doi.org/10.1371/journal.pgen.1002295

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Joshi, V., & Jander, G. (2009). Arabidopsis Methionine γ-Lyase Is Regulated According to Isoleucine Biosynthesis Needs But Plays a Subordinate Role to Threonine Deaminase. *Plant Physiology*, *151*(1), 367–378. https://doi.org/10.1104/pp.109.138651

Kaler, A. S., Gillman, J. D., Beissinger, T., & Purcell, L. C. (2020). Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. *Frontiers in Plant Science*, *10*, 1794. https://doi.org/10.3389/fpls.2019.01794

References

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kassambara, A. (2020). *ggpubr: "ggplot2" Based Publication Ready Plots* (0.4.0) [Computer software]. https://CRAN.R-project.org/package=ggpubr

Kessler, A., & Kalske, A. (2018). Plant Secondary Metabolite Diversity and Species Interactions. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 115–138. https://doi.org/10.1146/annurev-ecolsys-110617-062406

Kikuchi, S., Bheemanahalli, R., Jagadish, K. S. V., Kumagai, E., Masuya, Y., Kuroda, E., Raghavan, C., Dingkuhn, M., Abe, A., & Shimono, H. (2017). Genome-wide association mapping for phenotypic plasticity in rice. *Plant, Cell & Environment*, *40*(8), 1565–1575. https://doi.org/10.1111/pce.12955

Kim, J. H., & Kim, W. T. (2013). The Arabidopsis RING E3 Ubiquitin Ligase AtAIRP3/LOG2 Participates in Positive Regulation of High-Salt and Drought Stress Responses. *Plant Physiology*, *162*(3), 1733–1749. https://doi.org/10.1104/pp.113.220103

Kimura, S., & Sinha, N. (2008). Tomato (Solanum lycopersicum): A Model Fruit-Bearing Crop. *CSH Protocols*, *2008*, pdb.emo105. https://doi.org/10.1101/pdb.emo105

Knapp, A. K., Carroll, C. J. W., & Fahey, T. J. (2014). Patterns and Controls of Terrestrial Primary Production in a Changing World. In R. K. Monson (Ed.), *Ecology and the Environment* (pp. 205–246). Springer. https://doi.org/10.1007/978-1-4614-7501-9_2

Kobayashi, K. (2016). Role of membrane glycerolipids in photosynthesis, thylakoid biogenesis and chloroplast development. *Journal of Plant Research*, *129*(4), 565–580. https://doi.org/10.1007/s10265-016-0827-y

Koester, J., Bussmann, R., & Barz, W. (1984). Malonyl-coenzyme A:Isoflavone 7-O-glucoside-6″-O-malonyltransferase from roots of chick pea (Cicer arietinum L.).

References

Archives of Biochemistry and Biophysics, *234*(2), 513–521. https://doi.org/10.1016/0003-9861(84)90298-4

Kolde, R. (2019). *pheatmap: Pretty Heatmaps* (1.0.12) [Computer software]. https://CRAN.R-project.org/package=pheatmap

Koornneef, M., Alonso-Blanco, C., & Vreugdenhil, D. (2004). NATURALLY OCCURRING GENETIC VARIATION IN ARABIDOPSIS THALIANA. *Annual Review of Plant Biology*, *55*(1), 141–172. https://doi.org/10.1146/annurev.arplant.55.031903.141605

Koornneef, M., & Meinke, D. (2010). The development of Arabidopsis as a model plant. *The Plant Journal*, *61*(6), 909–921. 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

Kuhalskaya, A., Wijesingha Ahchige, M., Perez de Souza, L., Vallarino, J., Brotman, Y., & Alseekh, S. (2020). Network Analysis Provides Insight into Tomato Lipid Metabolism. *Metabolites*, *10*(4), E152. https://doi.org/10.3390/metabo10040152

Kwon, C.-T., Heo, J., Lemmon, Z. H., Capua, Y., Hutton, S. F., Van Eck, J., Park, S. J., & Lippman, Z. B. (2020). Rapid customization of Solanaceae fruit crops for urban agriculture. *Nature Biotechnology*, *38*(2), 182–188. https://doi.org/10.1038/s41587-019-0361-2

Kyle, J. E., Stratton, K. G., Zink, E. M., Kim, Y.-M., Bloodsworth, K. J., Monroe, M. E., Waters, K. M., Webb-Robertson, B.-J. M., Koeller, D. M., & Metz, T. O. (2021). A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases. *Scientific Data*, *8*(1), 114. https://doi.org/10.1038/s41597-021-00894-y

Lachowiec, J., Mason, G. A., Schultz, K., & Queitsch, C. (2018). Redundancy, Feedback, and Robustness in the Arabidopsis thaliana BZR/BEH Gene Family. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00523

References

Lachowiec, J., Queitsch, C., & Kliebenstein, D. J. (2016). Molecular mechanisms governing differential robustness of development and environmental responses in plants. *Ann Bot*, *117*(5), 795–809. https://doi.org/10.1093/aob/mcv151

Laitinen, R. A. E., & Nikoloski, Z. (2018). Genetic basis of plasticity in plants. *Journal of Experimental Botany*, ery404–ery404. https://doi.org/10.1093/jxb/ery404

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M., & Maza, E. (2018). Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00108

Lazar, C. (2015). *imputeLCMD: A collection of methods for left-censored missing data imputation* (2.0) [Computer software]. https://CRAN.R-project.org/package=imputeLCMD

Lea, P. J., Sodek, L., Parry, M. a. J., Shewry, P. R., & Halford, N. G. (2007). Asparagine in plants. *Annals of Applied Biology*, *150*(1), 1–26. https://doi.org/10.1111/j.1744-7348.2006.00104.x

Lehmann, T., & Pollmann, S. (2009). Gene expression and characterization of a stress-induced tyrosine decarboxylase from Arabidopsis thaliana. *FEBS Letters*, *583*(12), 1895–1900. https://doi.org/10.1016/j.febslet.2009.05.017

Lehner, B. (2010). Genes Confer Similar Robustness to Environmental, Stochastic, and Genetic Perturbations in Yeast. *PLOS ONE*, *5*(2), e9035. https://doi.org/10.1371/journal.pone.0009035

Li, J., Liu, L.-N., Meng, Q., Fan, H., & Sui, N. (2020). The roles of chloroplast membrane lipids in abiotic stress responses. *Plant Signaling & Behavior*, *15*(11), 1807152. https://doi.org/10.1080/15592324.2020.1807152

References

Li, P., Li, X., & Jiang, M. (2021). CRISPR/Cas9-mediated mutagenesis of WRKY3 and WRKY4 function decreases salt and Me-JA stress tolerance in Arabidopsis thaliana. *Molecular Biology Reports*, *48*(8), 5821–5832. https://doi.org/10.1007/s11033-021-06541-4

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., & Borevitz, J. O. (2010). Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(49), 21199–21204. https://doi.org/10.1073/pnas.1007431107

Li, Z.-W., Hou, X.-H., Chen, J.-F., Xu, Y.-C., Wu, Q., González, J., & Guo, Y.-L. (2018). Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana. *Genome Biology and Evolution*, *10*(8), 2140–2150. https://doi.org/10.1093/gbe/evy171

Lichtenthaler, H. K., & Buschmann, C. (2001). Chlorophylls and Carotenoids: Measurement and Characterization by UV-VIS Spectroscopy. *Current Protocols in Food Analytical Chemistry*, *1*(1), F4.3.1-F4.3.8. https://doi.org/10.1002/0471142913.faf0403s01

Lin, L.-Z., & Harnly, J. M. (2007). A Screening Method for the Identification of Glycosylated Flavonoids and Other Phenolic Compounds Using a Standard Analytical Approach for All Plant Materials. *Journal of Agricultural and Food Chemistry*, *55*(4), 1084–1096. https://doi.org/10.1021/jf062431s

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., & Fernie, A. R. (2006). Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature Protocols*, *1*(1), 387–396. https://doi.org/10.1038/nprot.2006.59

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., & Chen, L.-L. (2017). CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in Plants. *Molecular Plant*, *10*(3), 530–532. https://doi.org/10.1016/j.molp.2017.01.003

References

Liu, N., Du, Y., Warburton, M. L., Xiao, Y., & Yan, J. (2021). Phenotypic Plasticity Contributes to Maize Adaptation and Heterosis. *Molecular Biology and Evolution*, *38*(4), 1262–1275. https://doi.org/10.1093/molbev/msaa283

Liu, X., Jin, Y., Tan, K., Zheng, J., Gao, T., Zhang, Z., Zhao, Y., Ma, F., & Li, C. (2021). MdTyDc Overexpression Improves Alkalinity Tolerance in Malus domestica. *Frontiers in Plant Science*, *12*, 625890. https://doi.org/10.3389/fpls.2021.625890

Livera, A. M. D., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J. A., Castillo, S., Simpson, J. A., & Speed, T. P. (2015). Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Analytical Chemistry*, *87*(7), 3606–3615. https://doi.org/10.1021/ac502439y

Lu, Y. (2011). Extract Genomic DNA from Arabidopsis Leaves (Can be Used for Other Tissues as Well). *Bio-Protocol*, *1*(13), e90. https://doi.org/10.21769/BioProtoc.90

Lu, Y. (2016). Identification and Roles of Photosystem II Assembly, Stability, and Repair Factors in Arabidopsis. *Frontiers in Plant Science*, *7*(168). https://doi.org/10.3389/fpls.2016.00168

Luo, W., & Brouwer, C. (2013). Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, *29*(14), 1830–1831. https://doi.org/10.1093/bioinformatics/btt285

Macadlo, L. A., Ibrahim, I. M., & Puthiyaveetil, S. (2020). Sigma factor 1 in chloroplast gene transcription and photosynthetic light acclimation. *Journal of Experimental Botany*, *71*(3), 1029–1038. https://doi.org/10.1093/jxb/erz464

Maeda, H. A. (2019). Evolutionary Diversification of Primary Metabolism and Its Contribution to Plant Chemical Diversity. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00881

References

Martin, C. (2013). The interface between plant metabolic engineering and human health. *Current Opinion in Biotechnology*, *24*(2), 344–353. https://doi.org/10.1016/j.copbio.2012.11.005

Martino, H. S. D., Bigonha, S. M., Cardoso, L. de M., Rosa, C. de O. B., Costa, N. M. B., Cárdenas, L. de L. Á. R., & Ribeiro, S. M. R. (2012). Nutritional and Bioactive Compounds of Bean: Benefits to Human Health. In *Hispanic Foods: Chemistry and Bioactive Compounds* (Vol. 1109, pp. 233–258). American Chemical Society. https://doi.org/10.1021/bk-2012-1109.ch015

Massel, K., Lam, Y., Wong, A. C. S., Hickey, L. T., Borrell, A. K., & Godwin, I. D. (2021). Hotter, drier, CRISPR: The latest edit on climate change. *Theoretical and Applied Genetics*, *134*(6), 1691–1709. https://doi.org/10.1007/s00122-020-03764-0

Mata-Nicolás, E., Montero-Pau, J., Gimeno-Paez, E., Garcia-Carpintero, V., Ziarsolo, P., Menda, N., Mueller, L. A., Blanca, J., Cañizares, J., van der Knaap, E., & Díez, M. J. (2020). Exploiting the diversity of tomato: The development of a phenotypically and genetically detailed germplasm collection. *Horticulture Research*, *7*(1), 1–14. https://doi.org/10.1038/s41438-020-0291-7

Menda, N., Strickler, S. R., Edwards, J. D., Bombarely, A., Dunham, D. M., Martin, G. B., Mejia, L., Hutton, S. F., Havey, M. J., Maxwell, D. P., & Mueller, L. A. (2014). Analysis of wild-species introgressions in tomato inbreds uncovers ancestral origins. *BMC Plant Biology*, *14*(1), 287. https://doi.org/10.1186/s12870-014-0287-2

Moore, M., Harrison, M. S., Peterson, E. C., & Henry, R. (2000). Chloroplast Oxa1p Homolog Albino3 Is Required for Post-translational Integration of the Light Harvesting Chlorophyll-binding Protein into Thylakoid Membranes*. *Journal of Biological Chemistry*, *275*(3), 1529–1532. https://doi.org/10.1074/jbc.275.3.1529

References

Muranaka, A., Watanabe, S., Sakamoto, A., & Shimada, H. (2012). Arabidopsis cotyledon chloroplast biogenesis factor CYO1 uses glutathione as an electron donor and interacts with PSI (A1 and A2) and PSII (CP43 and CP47) subunits. *J Plant Physiol*, *169*(12), 1212–1215. https://doi.org/10.1016/j.jplph.2012.04.001

Nadeem, M. A., Yeken, M. Z., Shahid, M. Q., Habyarimana, E., Yılmaz, H., Alsaleh, A., Hatipoğlu, R., Çilesiz, Y., Khawar, K. M., Ludidi, N., Ercişli, S., Aasim, M., Karaköy, T., & Baloch, F. S. (2021). Common bean as a potential crop for future food security: An overview of past, current and future contributions in genomics, transcriptomics, transgenics and proteomics. *Biotechnology & Biotechnological Equipment*, *35*(1), 758–786. https://doi.org/10.1080/13102818.2021.1920462

Nalbantoglu, S. (2019). Metabolomics: Basic Principles and Strategies. In *Molecular Medicine*. IntechOpen. https://doi.org/10.5772/intechopen.88563

Nellaepalli, S., Ozawa, S.-I., Kuroda, H., & Takahashi, Y. (2018). The photosystem I assembly apparatus consisting of Ycf3–Y3IP1 and Ycf4 modules. *Nature Communications*, *9*(1), 2439. https://doi.org/10.1038/s41467-018-04823-3

Nelson, N., & Ben-Shem, A. (2004). The complex architecture of oxygenic photosynthesis. *Nature Reviews Molecular Cell Biology*, *5*(12), 971–982. https://doi.org/10.1038/nrm1525

Niño-González, M., Novo-Uzal, E., Richardson, D. N., Barros, P. M., & Duque, P. (2019). More Transporters, More Substrates: The Arabidopsis Major Facilitator Superfamily Revisited. *Molecular Plant*, *12*(9), 1182–1202. https://doi.org/10.1016/j.molp.2019.07.003

Niron, H., Barlas, N., Salih, B., & Türet, M. (2020). Comparative Transcriptome, Metabolome, and Ionome Analysis of Two Contrasting Common Bean Genotypes in Saline

References

Conditions. *Frontiers in Plant Science*, *11*, 2007. https://doi.org/10.3389/fpls.2020.599501

Nunes-Nesi, A., Carrari, F., Gibon, Y., Sulpice, R., Lytovchenko, A., Fisahn, J., Graham, J., Ratcliffe, R. G., Sweetlove, L. J., & Fernie, A. R. (2007). Deficiency of mitochondrial fumarase activity in tomato plants impairs photosynthesis via an effect on stomatal function. *The Plant Journal*, *50*(6), 1093–1106. https://doi.org/10.1111/j.1365-313X.2007.03115.x

Ofner, I., Lashbrooke, J., Pleban, T., Aharoni, A., & Zamir, D. (2016). Solanum pennellii backcross inbred lines (BILs) link small genomic bins with tomato traits. *The Plant Journal*, *87*(2), 151–160. https://doi.org/10.1111/tpj.13194

O'Malley, R. C., Barragan, C. C., & Ecker, J. R. (2015). A User's Guide to the Arabidopsis T-DNA Insertional Mutant Collections. *Methods in Molecular Biology (Clifton, N.J.)*, *1284*, 323–342. https://doi.org/10.1007/978-1-4939-2444-8_16

original, S., StatLib, from, & Leisch, by R. T. R. port by F. (2019). *bootstrap: Functions for the Book "An Introduction to the Bootstrap"* (2019.6) [Computer software]. https://CRAN.R-project.org/package=bootstrap

Ospina, R., & Marmolejo-Ramos, F. (2019). Performance of Some Estimators of Relative Variability. *Frontiers in Applied Mathematics and Statistics*, *5*, 43. https://doi.org/10.3389/fams.2019.00043

Ottenhof, H. H., Ashurst, J. L., Whitney, H. M., Saldanha, S. A., Schmitzberger, F., Gweon, H. S., Blundell, T. L., Abell, C., & Smith, A. G. (2004). Organisation of the pantothenate (vitamin B5) biosynthesis pathway in higher plants. *The Plant Journal*, *37*(1), 61–72. https://doi.org/10.1046/j.1365-313X.2003.01940.x

Pao, S. S., Paulsen, I. T., & Saier, M. H. (1998). Major Facilitator Superfamily. *Microbiology and Molecular Biology Reviews*, *62*(1), 1–34.

References

Park, J., Lee, S., Park, G., Cho, H., Choi, D., Umeda, M., Choi, Y., Hwang, D., & Hwang, I. (2021). CYTOKININ-RESPONSIVE GROWTH REGULATOR regulates cell expansion and cytokinin-mediated cell cycle progression. *Plant Physiology*, *186*(3), 1734–1746. https://doi.org/10.1093/plphys/kiab180

Paul, P., Simm, S., Mirus, O., Scharf, K.-D., Fragkostefanakis, S., & Schleiff, E. (2014). The Complexity of Vesicle Transport Factors in Plants Examined by Orthology Search. *PLoS ONE*, *9*(5), e97745. https://doi.org/10.1371/journal.pone.0097745

Pélabon, C., Armbruster, W. S., & Hansen, T. F. (2011). Experimental evidence for the Berg hypothesis: Vegetative traits are more sensitive than pollination traits to environmental variation. *Functional Ecology*, *25*(1), 247–257. https://doi.org/10.1111/j.1365-2435.2010.01770.x

Peralta, I. E., Spooner, D., & Knapp, S. (2008). Taxonomy of Wild Tomatoes and Their Relatives (Solanum sect. Lycopersicoides, sect. Juglandifolia, sect. Lycopersicon; Solanaceae). *Systematic Botany Monographs*, *84*, 1–186. https://doi.org/10.2307/25027972

Pérez-García, A., Pereira, S., Pissarra, J., Gutiérrez, A. G., Cazorla, F. M., Salema, R., de Vicente, A., & Cánovas, F. M. (1998). Cytosolic localization in tomato mesophyll cells of a novel glutamine synthetase induced in response to bacterial infection or phosphinothricin treatment. *Planta*, *206*(3), 426–434.

Peterhansel, C., Horst, I., Niessen, M., Blume, C., Kebeish, R., Kürkcüoglu, S., & Kreuzaler, F. (2010). Photorespiration. *The Arabidopsis Book / American Society of Plant Biologists*, *8*, e0130. https://doi.org/10.1199/tab.0130

Pfeiffer, T., Soyer, O. S., & Bonhoeffer, S. (2005). The Evolution of Connectivity in Metabolic Networks. *PLoS Biology*, *3*(7), e228. https://doi.org/10.1371/journal.pbio.0030228

References

Pilot, G., Stransky, H., Bushey, D. F., Pratelli, R., Ludewig, U., Wingate, V. P. M., & Frommer, W. B. (2004). Overexpression of GLUTAMINE DUMPER1 Leads to Hypersecretion of Glutamine from Hydathodes of Arabidopsis Leaves. *The Plant Cell*, *16*(7), 1827–1840. https://doi.org/10.1105/tpc.021642

Pitt, J. J. (2009). Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry. *The Clinical Biochemist Reviews*, *30*(1), 19–34.

Pogson, B. J., & Albrecht, V. (2011). Genetic Dissection of Chloroplast Biogenesis and Development: An Overview. *Plant Physiology*, *155*(4), 1545–1551. https://doi.org/10.1104/pp.110.170365

Pott, D. M., Osorio, S., & Vallarino, J. G. (2019). From Central to Specialized Metabolism: An Overview of Some Secondary Compounds Derived From the Primary Metabolism for Their Role in Conferring Nutritional and Organoleptic Characteristics to Fruit. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00835

Pratelli, R., Guerra, D. D., Yu, S., Wogulis, M., Kraft, E., Frommer, W. B., Callis, J., & Pilot, G. (2012). The Ubiquitin E3 Ligase LOSS OF GDU2 Is Required for GLUTAMINE DUMPER1-Induced Amino Acid Secretion in Arabidopsis. *Plant Physiology*, *158*(4), 1628–1642. https://doi.org/10.1104/pp.111.191965

Proost, S., Krawczyk, A., & Mutwil, M. (2017). LSTrAP: Efficiently combining RNA sequencing data into co-expression networks. *BMC Bioinformatics*, *18*(1), 444. https://doi.org/10.1186/s12859-017-1861-z

Qu, L.-J., & Zhu, Y.-X. (2006). Transcription factor families in Arabidopsis: Major progress and outstanding issues for future research. *Current Opinion in Plant Biology*, *9*(5), 544–549. https://doi.org/10.1016/j.pbi.2006.07.005

Queitsch, C., Sangster, T. A., & Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, *417*(6889), 618–624. https://doi.org/10.1038/nature749

References

Quesneville, H. (2020). Twenty years of transposable element analysis in the Arabidopsis thaliana genome. *Mobile DNA*, *11*(1), 28. https://doi.org/10.1186/s13100-020-00223-x

Radhakrishnan, G. V., Keller, J., Rich, M. K., Vernié, T., Mbadinga Mbadinga, D. L., Vigneron, N., Cottret, L., Clemente, H. S., Libourel, C., Cheema, J., Linde, A.-M., Eklund, D. M., Cheng, S., Wong, G. K. S., Lagercrantz, U., Li, F.-W., Oldroyd, G. E. D., & Delaux, P.-M. (2020). An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nature Plants*, *6*(3), 280–289. https://doi.org/10.1038/s41477-020-0613-7

Raggi, L., Caproni, L., Carboni, A., & Negri, V. (2019). Genome-Wide Association Study Reveals Candidate Genes for Flowering Time Variation in Common Bean (Phaseolus vulgaris L.). *Frontiers in Plant Science*, *10*, 962. https://doi.org/10.3389/fpls.2019.00962

Rambla, J. L., Medina, A., Fernández-del-Carmen, A., Barrantes, W., Grandillo, S., Cammareri, M., López-Casado, G., Rodrigo, G., Alonso, A., García-Martínez, S., Primo, J., Ruiz, J. J., Fernández-Muñoz, R., Monforte, A. J., & Granell, A. (2017). Identification, introgression, and validation of fruit volatile QTLs from a red-fruited wild tomato species. *Journal of Experimental Botany*, *68*(3), 429–442. https://doi.org/10.1093/jxb/erw455

Rambla, J. L., Tikunov, Y. M., Monforte, A. J., Bovy, A. G., & Granell, A. (2014). The expanded tomato fruit volatile landscape. *Journal of Experimental Botany*, *65*(16), 4613–4623. https://doi.org/10.1093/jxb/eru128

Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature Communications*, *6*(1), 5989. https://doi.org/10.1038/ncomms6989

References

Rébeillé, F., Jabrin, S., Bligny, R., Loizeau, K., Gambonnet, B., Van Wilder, V., Douce, R., & Ravanel, S. (2006). Methionine catabolism in Arabidopsis cells is initiated by a γ-cleavage process and leads to S-methylcysteine and isoleucine syntheses. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15687–15692. https://doi.org/10.1073/pnas.0606195103

Reckling, M., Ahrends, H., Chen, T.-W., Eugster, W., Hadasch, S., Knapp, S., Laidig, F., Linstädter, A., Macholdt, J., Piepho, H.-P., Schiffers, K., & Döring, T. F. (2021). Methods of yield stability analysis in long-term field experiments. A review. *Agronomy for Sustainable Development*, *41*(2), 27. https://doi.org/10.1007/s13593-021-00681-4

Reem, N. T., & Van Eck, J. (2019). Application of CRISPR/Cas9-Mediated Gene Editing in Tomato. In Y. Qi (Ed.), *Plant Genome Editing with CRISPR Systems: Methods and Protocols* (pp. 171–182). Springer New York. https://doi.org/10.1007/978-1-4939-8991-1_13

Ren, J.-L., Zhang, A.-H., Kong, L., & Wang, X.-J. (2018). Advances in mass spectrometry-based metabolomics for investigation of metabolites. *RSC Advances*, *8*(40), 22335–22350. https://doi.org/10.1039/C8RA01574K

Reyero-Saavedra, M. del R., Qiao, Z., Sánchez-Correa, M. del S., Díaz-Pineda, M. E., Reyes, J. L., Covarrubias, A. A., Libault, M., & Valdés-López, O. (2017). Gene Silencing of Argonaute5 Negatively Affects the Establishment of the Legume-Rhizobia Symbiosis. *Genes*, *8*(12), 352. https://doi.org/10.3390/genes8120352

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, D., Hayes, A., Couch [aut, S., cre, Patil, I., Chiu, D., Gomez, M., Demeshev, B., Menne, D., Nutter, B., Johnston, L., Bolker, B., Briatte, F., Arnold, J., Gabry, J., Selzer,

References

L., Simpson, G., Preussner, J., Hesselberth, J., … Lee, J. (2021). *broom: Convert Statistical Objects into Tidy Tibbles* (0.7.9) [Computer software]. https://CRAN.R-project.org/package=broom

Rohde, B., Hans, J., Martens, S., Baumert, A., Hunziker, P., & Matern, U. (2008). Anthranilate N-methyltransferase, a branch-point enzyme of acridone biosynthesis. *The Plant Journal*, *53*(3), 541–553. https://doi.org/10.1111/j.1365-313X.2007.03360.x

Rook, F., Corke, F., Card, R., Munz, G., Smith, C., & Bevan, M. W. (2001). Impaired sucrose-induction mutants reveal the modulation of sugar-induced starch biosynthetic gene expression by abscisic acid signalling. *The Plant Journal: For Cell and Molecular Biology*, *26*(4), 421–433. https://doi.org/10.1046/j.1365-313x.2001.2641043.x

Rosado-Souza, L., David, L. C., Drapal, M., Fraser, P. D., Hofmann, J., Klemens, P. A. W., Ludewig, F., Neuhaus, H. E., Obata, T., Perez-Fons, L., Schlereth, A., Sonnewald, U., Stitt, M., Zeeman, S. C., Zierer, W., & Fernie, A. R. (2019). Cassava Metabolomics and Starch Quality. *Current Protocols in Plant Biology*, *4*(4), e20102. https://doi.org/10.1002/cppb.20102

Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S., & Wilson, J. (2016). A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. *Metabolomics*, *12*(3). https://doi.org/10.1007/s11306-016-0972-2

Salem, M. A., Juppner, J., Bajdzienko, K., & Giavalisco, P. (2016). Protocol: A fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample. *Plant Methods*, *12*, 45. https://doi.org/10.1186/s13007-016-0146-2

Salem, M. A., Yoshida, T., Souza, L. P. de, Alseekh, S., Bajdzienko, K., Fernie, A. R., & Giavalisco, P. (2020). An improved extraction method enables the comprehensive

References

analysis of lipids, proteins, metabolites and phytohormones from a single sample of leaf tissue under water-deficit stress. *The Plant Journal*, *103*(4), 1614–1632. https://doi.org/10.1111/tpj.14800

Sangster, T. A., Salathia, N., Undurraga, S., Milo, R., Schellenberg, K., Lindquist, S., & Queitsch, C. (2008). HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(8), 2963–2968. https://doi.org/10.1073/pnas.0712200105

Schauer, N., Semel, Y., Balbo, I., Steinfath, M., Repsilber, D., Selbig, J., Pleban, T., Zamir, D., & Fernie, A. R. (2008). Mode of Inheritance of Primary Metabolic Traits in Tomato. *The Plant Cell*, *20*(3), 509–523. https://doi.org/10.1105/tpc.107.056523

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., & Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology*, *24*(4), 447–454. https://doi.org/10.1038/nbt1192

Schoor, S., Farrow, S., Blaschke, H., Lee, S., Perry, G., von Schwartzenberg, K., Emery, N., & Moffatt, B. (2011). Adenosine Kinase Contributes to Cytokinin Interconversion in Arabidopsis. *Plant Physiology*, *157*(2), 659–672. https://doi.org/10.1104/pp.111.181560

Seki, H., Sawai, S., Ohyama, K., Mizutani, M., Ohnishi, T., Sudo, H., Fukushima, E. O., Akashi, T., Aoki, T., Saito, K., & Muranaka, T. (2011). Triterpene Functional Genomics in Licorice for Identification of CYP72A154 Involved in the Biosynthesis of Glycyrrhizin[C][W][OA]. *The Plant Cell*, *23*(11), 4112–4123. https://doi.org/10.1105/tpc.110.082685

References

Serino, G., & Marzi, D. (2018). Arabidopsis thaliana as an Experimental Organism. In *ELS* (pp. 1–9). American Cancer Society. https://doi.org/10.1002/9780470015902.a0002031.pub3

Shevela, D., Ananyev, G., Vatland, A. K., Arnold, J., Mamedov, F., Eichacker, L. A., Dismukes, G. C., & Messinger, J. (2019). 'Birth defects' of photosystem II make it highly susceptible to photodamage during chloroplast biogenesis. *Physiologia Plantarum*, *166*(1), 165–180. https://doi.org/10.1111/ppl.12932

Shi, L.-X., Hall, M., Funk, C., & Schröder, W. P. (2012). Photosystem II, a growing complex: Updates on newly discovered components and low molecular mass proteins. *Biochimica Et Biophysica Acta*, *1817*(1), 13–25. https://doi.org/10.1016/j.bbabio.2011.08.008

Shimada, H., Mochizuki, M., Ogura, K., Froehlich, J. E., Osteryoung, K. W., Shirano, Y., Shibata, D., Masuda, S., Mori, K., & Takamiya, K. (2007). Arabidopsis cotyledon-specific chloroplast biogenesis factor CYO1 is a protein disulfide isomerase. *Plant Cell*, *19*(10), 3157–3169. https://doi.org/10.1105/tpc.107.051714

Shimizu, H., Peng, L., Myouga, F., Motohashi, R., Shinozaki, K., & Shikanai, T. (2008). CRR23/NdhL is a Subunit of the Chloroplast NAD(P)H Dehydrogenase Complex in Arabidopsis. *Plant and Cell Physiology*, *49*(5), 835–842. https://doi.org/10.1093/pcp/pcn058

Smyth, G. K., & Speed, T. (2003). Normalization of cDNA microarray data. *Methods (San Diego, Calif.)*, *31*(4), 265–273. https://doi.org/10.1016/s1046-2023(03)00155-5

Sokolowska, E. M., Schlossarek, D., Luzarowski, M., & Skirycz, A. (2019). PROMIS: Global Analysis of PROtein-Metabolite Interactions. *Current Protocols in Plant Biology*, *4*(4), e20101. https://doi.org/10.1002/cppb.20101

References

Souza, L. P. de, Scossa, F., Proost, S., Bitocchi, E., Papa, R., Tohge, T., & Fernie, A. R. (2019). Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (Phaseolus vulgaris) metabolome. *The Plant Journal*, *97*(6), 1132–1153. https://doi.org/10.1111/tpj.14178

Sugikawa, Y., Ebihara, S., Tsuda, K., Niwa, Y., & Yamazaki, K. (2005). Transcriptional coactivator MBF1s from Arabidopsis predominantly localize in nucleolus. *Journal of Plant Research*, *118*(6), 431–437. https://doi.org/10.1007/s10265-005-0238-y

Sukumaran, S., & Yu, J. (2014). Association Mapping of Genetic Resources: Achievements and Future Perspectives. In R. Tuberosa, A. Graner, & E. Frison (Eds.), *Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources* (pp. 207–235). Springer Netherlands. https://doi.org/10.1007/978-94-007-7572-5_9

Sulpice, R., Flis, A., Ivakov, A. A., Apelt, F., Krohn, N., Encke, B., Abel, C., Feil, R., Lunn, J. E., & Stitt, M. (2014). Arabidopsis Coordinates the Diurnal Regulation of Carbon Allocation and Growth across a Wide Range of Photoperiods. *Molecular Plant*, *7*(1), 137–155. https://doi.org/10.1093/mp/sst127

Sulpice, R., Trenkamp, S., Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E.-T., Tschoep, H., Steinhauser, M. C., Guenther, M., Hoehne, M., Rohwer, J. M., Altmann, T., Fernie, A. R., & Stitt, M. (2010). Network Analysis of Enzyme Activities and Metabolite Levels and Their Relationship to Biomass in a Large Panel of Arabidopsis Accessions. *The Plant Cell*, *22*(8), 2872–2893. https://doi.org/10.1105/tpc.110.076653

Sun, S., Kang, X.-P., Xing, X.-J., Xu, X.-Y., Cheng, J., Zheng, S.-W., & Xing, G.-M. (2015). Agrobacterium-mediated transformation of tomato (Lycopersicon esculentum L. cv. Hezuo 908) with improved efficiency. *Biotechnology & Biotechnological Equipment*, *29*(5), 861–868. https://doi.org/10.1080/13102818.2015.1056753

References

Sun, T., Zhou, F., Huang, X.-Q., Chen, W.-C., Kong, M.-J., Zhou, C.-F., Zhuang, Z., Li, L., & Lu, S. (2019). ORANGE Represses Chloroplast Biogenesis in Etiolated Arabidopsis Cotyledons via Interaction with TCP14. *The Plant Cell*, *31*(12), 2996–3014. https://doi.org/10.1105/tpc.18.00290

Sun, X., Fu, T., Chen, N., Guo, J., Ma, J., Zou, M., Lu, C., & Zhang, L. (2010). The Stromal Chloroplast Deg7 Protease Participates in the Repair of Photosystem II after Photoinhibition in Arabidopsis1[W][OA]. *Plant Physiology*, *152*(3), 1263–1273. https://doi.org/10.1104/pp.109.150722

Sundberg, E., Slagter, J. G., Fridborg, I., Cleary, S. P., Robinson, C., & Coupland, G. (1997). ALBINO3, an Arabidopsis nuclear gene essential for chloroplast differentiation, encodes a chloroplast protein that shows homology to proteins present in bacterial membranes and yeast mitochondria. *The Plant Cell*, *9*(5), 717–730.

Sweetlove, L. J., Nielsen, J., & Fernie, A. R. (2017). Engineering central metabolism – a grand challenge for plant biologists. *The Plant Journal*, *90*(4), 749–763. https://doi.org/doi:10.1111/tpj.13464

Takuno, S., Terauchi, R., & Innan, H. (2012). The Power of QTL Mapping with RILs. *PLOS ONE*, *7*(10), e46545. https://doi.org/10.1371/journal.pone.0046545

Tan, Y., Zhou, J., Wang, J., & Sun, L. (2020). The Genetic Architecture for Phenotypic Plasticity of the Rice Grain Ionome. *Frontiers in Plant Science*, *11*, 12. https://doi.org/10.3389/fpls.2020.00012

Tang, K.-H., & Blankenship, R. E. (2013). Photosynthetic Electron Transport. In G. C. K. Roberts (Ed.), *Encyclopedia of Biophysics* (pp. 1868–1873). Springer. https://doi.org/10.1007/978-3-642-16712-6_20

Tanz, S. K., Kilian, J., Johnsson, C., Apel, K., Small, I., Harter, K., Wanke, D., Pogson, B., & Albrecht, V. (2012). The SCO2 protein disulphide isomerase is required for thylakoid

References

biogenesis and interacts with LCHB1 chlorophyll a/b binding proteins which affects chlorophyll biosynthesis in Arabidopsis seedlings. *The Plant Journal*, *69*(5), 743–754. https://doi.org/10.1111/j.1365-313X.2011.04833.x

Teshima, T., Yamada, N., Yokota, Y., Sayama, T., Inagaki, K., Koeduka, T., Uefune, M., Ishimoto, M., & Matsui, K. (2020). Suppressed Methionine γ-Lyase Expression Causes Hyperaccumulation of S-Methylmethionine in Soybean Seeds1[OPEN]. *Plant Physiology*, *183*(3), 943–956. https://doi.org/10.1104/pp.20.00254

Thoen, M. P. M., Davila Olivas, N. H., Kloth, K. J., Coolen, S., Huang, P.-P., Aarts, M. G. M., Bac-Molenaar, J. A., Bakker, J., Bouwmeester, H. J., Broekgaarden, C., Bucher, J., Busscher-Lange, J., Cheng, X., Fradin, E. F., Jongsma, M. A., Julkowska, M. M., Keurentjes, J. J. B., Ligterink, W., Pieterse, C. M. J., … Dicke, M. (2017). Genetic architecture of plant stress resistance: Multi-trait genome-wide association mapping. *New Phytologist*, *213*(3), 1346–1362. https://doi.org/10.1111/nph.14220

Tieman, D. M., Zeigler, M., Schmelz, E. A., Taylor, M. G., Bliss, P., Kirst, M., & Klee, H. J. (2006). Identification of loci affecting flavour volatile emissions in tomato fruits. *Journal of Experimental Botany*, *57*(4), 887–896. https://doi.org/10.1093/jxb/erj074

Tilton, G. B., Wedemeyer, W. J., Browse, J., & Ohlrogge, J. (2006). Plant coenzyme A biosynthesis: Characterization of two pantothenate kinases from Arabidopsis. *Plant Molecular Biology*, *61*(4), 629–642. https://doi.org/10.1007/s11103-006-0037-4

Toubiana, D., Xue, W., Zhang, N., Kremling, K., Gur, A., Pilosof, S., Gibon, Y., Stitt, M., Buckler, E. S., Fernie, A. R., & Fait, A. (2016). Correlation-Based Network Analysis of Metabolite and Enzyme Profiles Reveals a Role of Citrate Biosynthesis in Modulating N and C Metabolism in Zea mays. *Frontiers in Plant Science*, *7*, 1022. https://doi.org/10.3389/fpls.2016.01022

References

Vaddepalli, P., Fulton, L., Wieland, J., Wassmer, K., Schaeffer, M., Ranf, S., & Schneitz, K. (2017). The cell wall-localized atypical β-1,3 glucanase ZERZAUST controls tissue morphogenesis in Arabidopsis thaliana. *Development*, *144*(12), 2259–2269. https://doi.org/10.1242/dev.152231

Van Eck, J., Keen, P., & Tjahjadi, M. (2019). Agrobacterium tumefaciens-Mediated Transformation of Tomato. *Methods in Molecular Biology (Clifton, N.J.)*, *1864*, 225–234. https://doi.org/10.1007/978-1-4939-8778-8_16

Waddington, C. H. (1942). CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS. *Nature*, *150*, 563. https://doi.org/10.1038/150563a0

Walter, W., Sánchez-Cabo, F., & Ricote, M. (2015). GOplot: An R package for visually combining expression data with functional analysis. *Bioinformatics*, *31*(17), 2912–2914. https://doi.org/10.1093/bioinformatics/btv300

Wang, S., Alseekh, S., Fernie, A. R., & Luo, J. (2019). The Structure and Function of Major Plant Metabolite Modifications. *Molecular Plant*, *12*(7), 899–919. https://doi.org/10.1016/j.molp.2019.06.001

Wang, Y., Stessman, D. J., & Spalding, M. H. (2015). The CO2 concentrating mechanism and photosynthetic carbon assimilation in limiting CO2: How Chlamydomonas works against the gradient. *The Plant Journal*, *82*(3), 429–448. https://doi.org/10.1111/tpj.12829

Warren, C. R. (2008). Rapid Measurement of Chlorophylls with a Microplate Reader. *Journal of Plant Nutrition*, *31*(7), 1321–1332. https://doi.org/10.1080/01904160802135092

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, *8*(1), 663. https://doi.org/10.1038/s41598-017-19120-0

References

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., & Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature Communications*, *5*(1), 1–10. https://doi.org/10.1038/ncomms4438

Whitacre, J. (2012). Biological Robustness: Paradigms, Mechanisms, and Systems Principles. *Frontiers in Genetics*, *3*, 67. https://doi.org/10.3389/fgene.2012.00067

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & RStudio. (2020). *modelr: Modelling Functions that Work with the Pipe* (0.1.8) [Computer software]. https://CRAN.R-project.org/package=modelr

Wilke, C. O. (2020a). *ggtext: Improved Text Rendering Support for "ggplot2"* (0.1.1) [Computer software]. https://CRAN.R-project.org/package=ggtext

Wilke, C. O. (2020b). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2"* (1.1.1) [Computer software]. https://CRAN.R-project.org/package=cowplot

Wing, I. S., De Cian, E., & Mistry, M. N. (2021). Global vulnerability of crop yields to climate change. *Journal of Environmental Economics and Management*, *109*, 102462. https://doi.org/10.1016/j.jeem.2021.102462

Wu, D., Wright, D. A., Wetzel, C., Voytas, D. F., & Rodermel, S. (1999). The IMMUTANS Variegation Locus of Arabidopsis Defines a Mitochondrial Alternative Oxidase Homolog That Functions during Early Chloroplast Biogenesis. *The Plant Cell*, *11*(1), 43–55. https://doi.org/10.1105/tpc.11.1.43

References

Wu, S., Alseekh, S., Cuadros-Inostroza, A., Fusari, C. M., Mutwil, M., Kooke, R., Keurentjes, J. B., Fernie, A. R., Willmitzer, L., & Brotman, Y. (2016). Combined Use of Genome-Wide Association Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in Arabidopsis thaliana. *PLoS Genetics*, *12*(10), e1006363. https://doi.org/10.1371/journal.pgen.1006363

Wu, S., Tohge, T., Cuadros-Inostroza, A., Tong, H., Tenenboim, H., Kooke, R., Meret, M., Keurentjes, J. B., Nikoloski, Z., Fernie, A. R., Willmitzer, L., & Brotman, Y. (2018). Mapping the Arabidopsis Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions. *Molecular Plant*, *11*(1), 118–134. https://doi.org/10.1016/j.molp.2017.08.012

Yang, H., Liu, J., Wen, X., & Lu, C. (2015). Molecular mechanism of photosystem I assembly in oxygenic organisms. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1847*(9), 838–848. https://doi.org/10.1016/j.bbabio.2014.12.011

Ye, J., Li, W., Ai, G., Li, C., Liu, G., Chen, W., Wang, B., Wang, W., Lu, Y., Zhang, J., Li, H., Ouyang, B., Zhang, H., Fei, Z., Giovannoni, J. J., Ye, Z., & Zhang, Y. (2019). Genome-wide association analysis identifies a natural variation in basic helix-loop-helix transcription factor regulating ascorbate biosynthesis via D-mannose/L-galactose pathway in tomato. *PLOS Genetics*, *15*(5), e1008149. https://doi.org/10.1371/journal.pgen.1008149

Yin, J., Zhang, X., Zhang, G., Wen, Y., Liang, G., & Chen, X. (2019). Aminocyclopropane-1-carboxylic acid is a key regulator of guard mother cell terminal division in Arabidopsis thaliana. *Journal of Experimental Botany*, *70*(3), 897–908. https://doi.org/10.1093/jxb/ery413

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-

References

accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2020.10.007

Zagari, N., Sandoval-Ibañez, O., Sandal, N., Su, J., Rodriguez-Concepcion, M., Stougaard, J., Pribil, M., Leister, D., & Pulido, P. (2017). SNOWY COTYLEDON 2 Promotes Chloroplast Development and Has a Role in Leaf Variegation in Both Lotus japonicus and Arabidopsis thaliana. *Molecular Plant*, *10*(5), 721–734. https://doi.org/10.1016/j.molp.2017.02.009

Zargar, S. M., Raatz, B., Sonah, H., MuslimaNazir, Bhat, J. A., Dar, Z. A., Agrawal, G. K., & Rakwal, R. (2015). Recent advances in molecular marker techniques: Insight into QTL mapping, GWAS and genomic selection in plants. *Journal of Crop Science and Biotechnology*, *18*(5), 293–308. https://doi.org/10.1007/s12892-015-0037-5

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, *136*(4), 1457–1468. https://doi.org/10.1093/genetics/136.4.1457

Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., Huang, S., Tieman, D. M., Klee, H. J., & Causse, M. (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nature Communications*, *10*(1), 1534. https://doi.org/10.1038/s41467-019-09462-w

Zhu, F., Alseekh, S., Koper, K., Tong, H., Nikoloski, Z., Naake, T., Liu, H., Yan, J., Brotman, Y., Wen, W., Maeda, H., Cheng, Y., & Fernie, A. R. (2021). Genome-wide association of the metabolic shifts underpinning dark-induced senescence in Arabidopsis. *The Plant Cell*, *koab251*. https://doi.org/10.1093/plcell/koab251

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., Cao, X., Han, X., Wang, X., van der Knaap, E., Zhang, Z., Cui, X., Klee, H., Fernie, A. R., Luo, J., & Huang, S. (2018). Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell*, *172*(1–2), 249-261 e12. https://doi.org/10.1016/j.cell.2017.12.019

Supplementary data

Supplementary data is provided as separate files retrievable together with the online version of this thesis on the publication server of the university library of the University of Potsdam (https://publishup.uni-potsdam.de/54884).