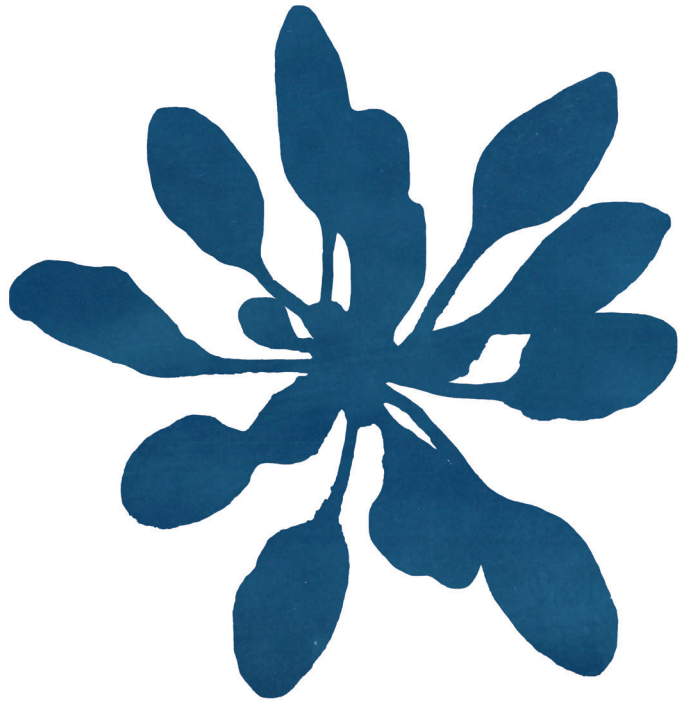
Linking the gene regulatory network with the functional physical structure of whole-genome engineered Arabidopsis mutants : an HR-MAS NMR-based metabolomics approach

Augustijn, D.



Introduction

The changing climate is a challenge for both current and future generations. To address these challenges, countries adopted the Paris agreement in 2015.[1] Two major goals of this agreement are that the global temperature increase stays well below 2 degrees Celsius above pre-industrial levels and to strive for an increase of not more than 1.5 degrees Celsius by the end of the century. It is also stated in the Paris agreement that the adaptions to accomplish the goals are not allowed to affect food production. This is also in agreement with the United Nations Sustainable Development Goals (SDGs) to end all forms of hunger and malnutrition in 2030 and to achieve food security.[2] Also, sustainable food production is promoted by the SDGs.

The challenge for the upcoming decades is to look for new resources, including biological resources, to meet both the demand in energy and food for the rising world population which is expected to reach 9.8 billion in 2050.[3] But in the meantime, it is also important to keep the 2 degree Celsius scenario (2DS) on track. In the 2DS, the goals are limiting the global temperature increase to 2 degree Celsius, but also reduce the CO_2 emission by 60% by 2050 in comparison to 2013.[4] One way of doing this is to make use of bioenergy, which is defined as the energy derived from the conversion of biomass.[5] Bioenergy can be used directly as fuel or processed into liquids or gases. The International Energy Agency (IEA) yearly examines the progress of renewable energy technologies to meet the 2DS targets. According to the recent report "Tracking clean energy progress 2017", the progress in bioenergy is not on track to reach the 2DS targets.[4] It is thus important to improve the yield of biomass resources, not only for food but also for bioenergy.

One way to produce more biomass is to enhance the productivity of agriculture. New scientific and technology tools can help in the challenge to produce sustainable agricultural products with increased yield and with a minimal environmental footprint. Such agricultural products are also called smart crops.[6,7]

Model plant organisms that can help in the development of smart crops are studied extensively by the research community.[8] The most used plant model is *Arabidopsis thaliana*. This is a small flowering rosette plant from the *Brassicaceae* family with a life cycle of 6 weeks. The plant has a small genome of 135 Mbp in total, which has been sequenced in 2000 as the first plant genome.[9,10] Transformations to obtain transgenic plants is an easy procedure.[11] *Arabidopsis* is widely used to understand molecular principles of plant development, cell biology, metabolism, physiology, genetics and epigenetics of plants. This is expanded in the last years to the field of systems biology.[9,10,12] The principal challenge in systems biology is to link the elements of the gene regulatory network with the functional physical structure and find novel molecular-based paths to plant development with enhanced yield. In this thesis, we will show a new unbiased method how this can be achieved.

1.1 Development of smart crop varieties

The development of smart crops can be accelerated by technologies to perform genome editing, in contrast to traditional breeding methods such as cross-breeding.[13] In particular, artificial transcription factors (ATFs) can be used to regulate the expression of target genes. Popular genome-editing tools using ATFs are clustered regularly interspaced short

palindromic repeats/Cas9 (CRISPR/Cas9), transcription activator-like effector nucleases (TALENs) and zinc-fingers (ZFs).[14-17]

In this study, *Arabidopsis* mutants are obtained using zinc finger artificial transcription factor (ZF-ATF)-mediated whole genome interrogation (Figure 1.1). This technology does not require a-priory knowledge or hypotheses and is therefore unbiased. New plants can be identified with rare mutations and phenotypes. ZF-ATFs consist of a DNA-binding domain from the zinc-finger proteins linked to an effector domain which is either an activator or a repressor domain.[14,16,18,19] The zinc protein is approximately 30 amino acids long, with a conserved ββα configuration and a backbone of conjugated cysteine (Cys) and histidine (His) residues and a zinc ion. The Cys_2-His_2 zinc finger domain is a very common DNA-binding motif in eukaryotes and binds to three base pairs in the major groove of the DNA.[14,16,18,19]

For the development of the mutants, an array of three zinc-fingers was used which recognizes the DNA sequence of 5'-GNN-3' (Figure 1.1). There are sixteen 5'-GNN-3'-binding ZFs available[20], which are leading to 4096 possible combinations of three zinc-fingers which will recognize 9 base pairs in the genome.[21] The effector domain used to develop the mutants for this study is the VP16 activator domain from the Herpes Simplex virus.[14,16,18,19] The artificial transcription factor library which originated by fusion of the 4096 three zinc-finger combinations with the VP16 domain is used to prepare transformed *Arabidopsis thaliana* plants using the floral dip method using *Agrobacterium*.[11,22]

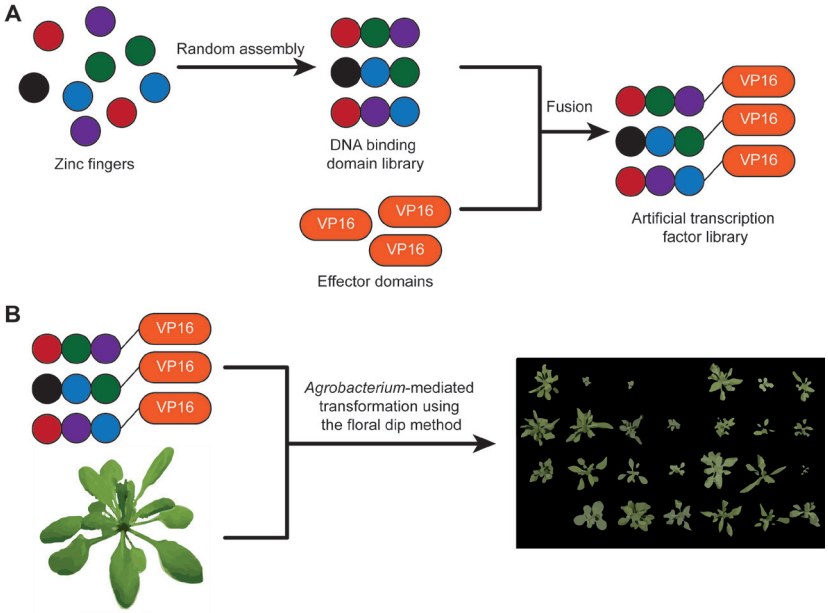


Figure 1.1: A) Development of the artificial transcription factor library by binding three zinc fingers together to prepare the DNA binding domain library and fuse this with the VP16 effector domain. B) Transformation of Arabidopsis thaliana plants with the artificial transcription factor library using the floral dip method to obtain Arabidopsis thaliana mutants with different phenotypes caused by the genome interrogation of the artificial transcription factors.

The first step in the floral dip method is to prepare a cell suspension of *Agrobacterium* strains including a binary vector containing the genes of the zinc-fingers. Then young flowering *Arabidopsis thaliana* plants are dipped into the *Agrobacterium* suspension to inoculate the plants with the *Agrobacterium*. The plants are grown in their normal growing conditions till the plants produce seeds. The seeds are collected and used to grow the first generation of transgenic plants (T_1).[11,22] These transgenic plants are screened for interesting novel phenotypes, such as growth enhancement[23], enhanced photochemistry[24] and increased tolerance to salinity.[25]

The purpose of this PhD thesis is to understand the novel mutant phenotypes from a systems biology perspective where omics technologies, like metabolomics and transcriptomics, are integrated with bio-imaging methods performed in a partner project.[26] The ultimate goal is to understand which pathways are leading to a specific phenotype so that it can be subsequently implanted to interesting crops for biomass production, such as plants of other *Brassicaceae* family that are used in bioenergy production[27] e.g. rapeseed (*Brassica napus*), Ethiopian mustard (*Brassica carinata*) and camelina (*Camelina sativa*).[28] And also to other plant families which are interesting for bioenergy production, such as maize (*Zea mays*), sugar beet (*Beta vulgaris*), sugarcane (*Saccharum officinarum*) and soybean (*Glycine max*).[29]

1.2 The role of metabolomics in systems biology

To understand the biological pathways underlying the novel mutant phenotypes, a systems biology approach can be used.[30-32] In systems biology, the information and interaction of the functional physical structure and the genetic information are integrated to provide a comprehensive model of the organism (Figure 1.2A). Different high-throughput technologies are used to study the genetic program of the various -omics fields: genomics, transcriptomics, proteomics and metabolomics (Figure 1.2B). This is complemented with information from

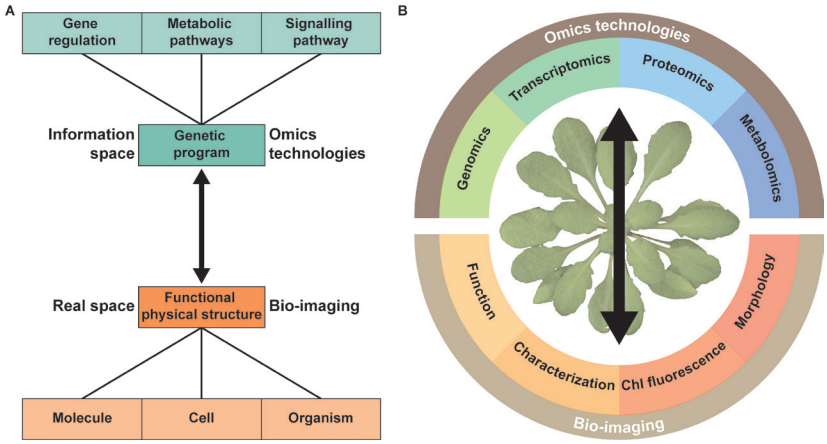


Figure 1.2: A) In systems biology, the information from the genetic program is integrated with information from functional physical structures to provide a comprehensive model of Arabidopsis thaliana. B) The genetic program can be studied with omics technologies and integrated with information from the functional physical structure using bio-imaging methods.

the functional physical structure using bio-imaging methods where e.g. the morphology, chlorophyll fluorescence, characterisation of pigments and functional studies are performed to better understand an organism (Figure 1.2B).

The link between the gene regulatory network and the functional physical structure (the double arrow in Figure 1.2A) is generally considered highly complex, with many pathways and pathway nodes interacting in what are often considered multifactorial processes. While it is undoubtedly flexible and adaptable to environmental constraints, the underlying links for a specific phenotype may turn out to be monofactorial, in particular in plants that can be grown under highly controlled conditions. The specific purpose of my thesis is to demonstrate that this can indeed be the case, and that reduction of complexity in systems biology may turn out to be the rule, rather than the exception, provided unbiased whole genome screens are performed to allow for unexpected factors. There are intriguing examples from other scientific disciplines as well, that indicate monofactorial connections. For instance, in host-pathogen interactions in human health, pathogens are likely to operate with very limited intervention interactions.[33] Thus, once the toolbox in Figure 1.2B for in depth analysis of the connections between the gene regulatory network and the functional physical structure is established, it can be applied in many areas of the life science, including human health.

In this PhD thesis, the ultimate goal is to understand the route for reducing the complexity of the different *Arabidopsis thaliana* mutants using metabolomics and to understand the underlying pathways of the phenotype of the mutants in a general framework.[34] Metabolomics is the study of all metabolites in an organism both qualitatively and quantitatively to get a clear metabolic picture of a living organism under specific conditions.[35-37] The metabolome is most closely related to the phenotype of a plant since metabolites are the end products of cellular processes.[35] Metabolomics is used to study development under normal and abiotic conditions (temperature, light, salt)[38] and biotic stress conditions (fungal, insects)[39,40], safety assessment of genetically modified crops[41], speed up crop improvements[42], effect of fruit storage[43] and the detection of food fraud.[44,45]

1.3 Analytical techniques in metabolomics

To study the metabolic profile of a plant, mass spectrometry (MS) or liquid-state nuclear magnetic resonance (NMR) spectrometry are the most common techniques in metabolomics. Both techniques have their own advantages and limitations as shown in Table 1.1. NMR spectrometry is a method which is non-destructive, with a high reproducibility and allows to quantify metabolites. On the other hand, while MS is more sensitive, which allows to detect more metabolites in a sample, it needs different chromatography techniques such as gas chromatography (GC) or liquid chromatography (LC) for different classes of metabolites.[42,46]

For both techniques, extraction of metabolites from the sample is necessary to obtain the metabolic profile. The drawback of this extraction is that it is not only time-consuming, but also that metabolites might be lost or degraded during extraction.[51] One way to eliminate the extraction procedure is to use high-resolution magic angle spinning (HR-MAS) NMR which allows using of intact tissue samples.[46,52,53]

Table 1.1: The advantages and limitations of NMR spectrometry and mass spectrometry for metabolic profiling.[47-50]

	NMR spectrometry	Mass spectrometry
Sensitivity	Low sensitivity, but can be improved with higher field strength and cryo- or microprobes.	High sensitivity, can reach the detection limit of attomolar (10^{-18}) concentrations.
Sample measurement	In one measurement with a detectable concentration can be detected.	Need chromatography techniques for different classes of metabolites.
Sample recovery	Non-destructive technique Several analyses can be performed on the same extracted sample.	Destructive technique.
Reproducibility	Very high.	Moderate.
Quantification	Absolute quantitation of metabolites possible by adding one standard with known concentration.	Quantification is possible with authentic standards, which are not available for newly identified compounds. Also, ionization efficiencies, ion suppression and matrix effects have influences on the concentration.
Targeted or untargeted approach	Untargeted approach.	Untargeted and targeted approach, but mainly used for targeted analysis.

1.4 Theoretical background of HR-MAS NMR

An NMR experiment can be described with a nuclear spin Hamiltonian:

$$H = H_{CS} + H_D^{IS} + H_D^{II} \tag{1.1}$$

Here,

$$H_{CS} = \{\sigma_{iso}\gamma B_0 + \tfrac{1}{2}\delta[3\cos^2(\theta) - 1 - \eta\sin^2(\theta)\cos(2\phi)]\}I_z \tag{1.2}$$

represents the chemical shift anisotropy interaction of the nuclei with the electronic environment,

$$H_D^{IS} = -\frac{\mu_0}{4\pi}\hbar\sum_i\sum_j\frac{\gamma^I\gamma^S}{r_{ij}^3}\frac{1}{2}(3\cos^2(\theta_{ij}) - 1)2I_z^iS_z^j \tag{1.3}$$

is the heteronuclear dipolar coupling between two different nuclear species I and S, and

$$H_D^{II} = -\frac{\mu_0}{4\pi}\hbar\sum_i\sum_j\frac{\gamma^2}{r_{ij}^3}\frac{1}{2}(3\cos^2(\theta_{ij}) - 1)(3I_z^iI_z^j - \mathbf{I}^i\cdot\mathbf{I}^j) \tag{1.4}$$

is the homonuclear dipolar coupling.[53-55]

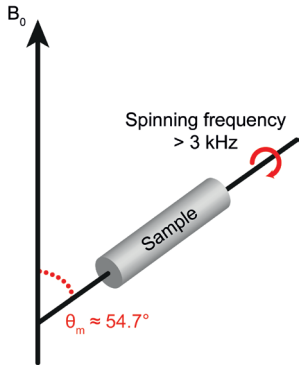


Figure 1.3: HR-MAS setup where the sample is rotated with high frequency (> 3 kHz) tiled by the magic angle θ_m with respect to the magnetic field B_0.

Here, σ_{iso} is the isotropic value, γ the gyromagnetic ratio and η is the asymmetry parameter. For the heteronuclear and homonuclear dipolar coupling, r_{ij} is the distance between the nuclei i and j and θ_{ij} is the angle between r_{ij} and the z-axis. The I spin is the abundant spin and S is the rare spins.

All three interaction terms depend on $\frac{1}{2}(3\cos^2(\theta)-1)$, where θ is the polar angle that describes the orientation of the magnetic field B_0 in the principle axis frame of the chemical shift tensor or dipolar interaction tensor. With HR-MAS NMR, the solid sample is rapidly rotated at the magic angle $\theta_m = 54.7°$. The angular dependences of the spin Hamiltonian will be averaged to zero over the sample and the broadening will be effectively removed (Figure 1.3). Although the anisotropic interactions produce spinning sidebands, these are suppressed when spinning at high frequencies (> 3 kHz), and the spectra will have narrow signals.

HR-MAS NMR is a combination of solid- and liquid-state NMR techniques, which can obtain spectra with similar resolution as spectra from liquid-state NMR experiments but make use of semi-solid samples with restricted molecular mobility.[56] Semi-solid samples, like biological tissues, can be used without extraction steps using this technique. In HR-MAS NMR, the effect of hetero- and homonuclear dipolar coupling is minimized at a frequency of a few kHz, while rigid solid samples need spinning frequencies of 20 – 50 kHz.

1.5 HR-MAS NMR-based workflow

In my PhD thesis, I propose an HR-MAS NMR-based workflow and apply the workflow to *Arabidopsis thaliana* mutants. The HR-MAS NMR-based workflow is shown in Figure 1.4. The workflow starts with the harvesting of the leaves from the *Arabidopsis thaliana* plant for preparation of a sample in the rotor, followed by performing the HR-MAS NMR experiments. The pulse sequences which can be used in metabolomics are described in section 1.6. The data are pre-processed and reduced by bucketing (section 1.7). Multivariate analysis is executed to determine biomarkers for the mutants in three steps: detection of outliers, investigation of the variation between different samples and selection of potential biomarker candidates (section 1.8). Finally, the biomarkers will be quantified and biological

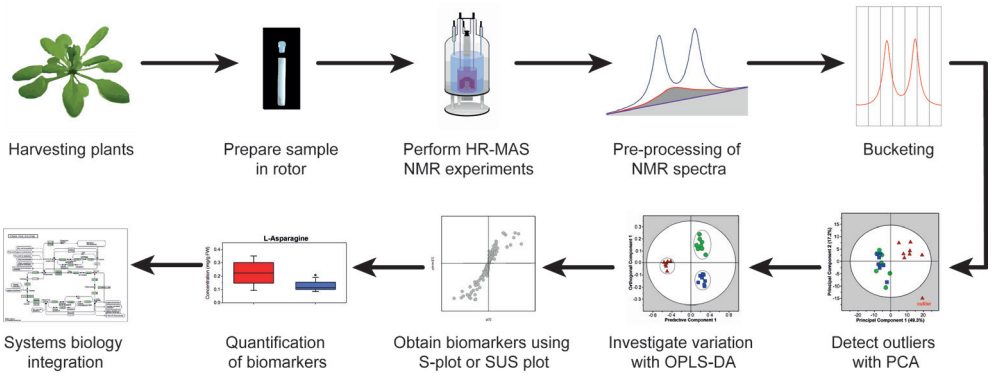


Figure 1.4: A typical high-resolution magic angle spinning (HR-MAS) NMR-based workflow

interpretation will be performed in a comprehensive systems biology approach by using available information from the literature. This workflow is based on a recently established liquid-state NMR approach[57], and the advantage of using HR-MAS NMR on leaves is that experiments can be performed genuinely *in vivo*, which will be illustrated with selected plant metabolomics applications (section 1.9).

1.6 Pulse sequences used in metabolomics

A complementary set of pulse sequences is used in NMR-based metabolomics to identify and quantify metabolites. One-dimensional spectra are used mostly to quantify metabolites. The mostly used pulse sequences are the one-dimensional ^{1}H NOESY (nuclear overhauser effect spectroscopy) with water presaturation and the ^{1}H CPMG (Carr-Purcell-Meiboom-Gill) sequence. NOESY spectra provide a complete and quantitative profile of the observed metabolites with the suppression of the water peak without an effect on the intensity of the other peaks.[58-60] CPMG is a pulse sequence which removes the broad signals from macromolecules, like proteins and lipids.[58,61]

In one-dimensional NMR spectra, signals from different metabolites strongly overlap. A way to solve this is to use two-dimensional NMR experiments. ^{1}H homonuclear correlation experiments are commonly used for identification. COSY (correlation spectroscopy) identifies spin-spin coupling of protons[58,61] and TOCSY (total correlation spectroscopy) provides information about the correlation between all protons in metabolites.[59,61] Another experiment is the ^{1}H *J*-resolved where the effect of chemical shift and *J*-coupling is separated into two independent directions.[62]

With the identification of new metabolites, it is sometimes helpful to make use of ^{1}H-^{13}C heteronuclear correlation experiments. These experiments provide information about the coupling between a proton and a carbon.[59,61] HSQC (heteronuclear single-quantum correlation) gives input about the correlation between a proton and a carbon which are separated by one bond. in addition, HMBC (heteronuclear multiple-bond correlation) gives information about the correlation over multiple bonds.[63]

Nearly all pulse sequences used for liquid samples in conventional high-resolution NMR

are also applicable for HR-MAS NMR. Out of the many pulse sequences that are available, three pulse sequences have been used for this work as shown in Figure 1.5. A CPMG pulse sequence with water suppression is used to obtain one-dimensional spectra (Figure 1.5A). The pulse sequence starts with a relaxation delay (d1) followed by a 90° pulse. Then there is looped n times over a spin echo delay (d20) followed by a 180° pulse and the FID signal is reordered. The CPMG pulse sequence is chosen because it minimizes the signal from lipids.[53] To confirm assignment, a magnitude-mode gradient-selected two-dimensional ^{1}H-^{1}H COSY and ^{1}H *J*-resolved pulse sequences (Figure 1.5B-C) can be used. The COSY pulse sequences start with water presaturation and a relaxation delay (d1) followed by a 90° pulse. After an increment delay (d0), a gradient pulse is executed, then a 90° pulse is applied and a gradient pulse at the same time followed by data acquisition. COSY spectra are very useful to resolve overlapping signals in the NMR spectra, especially in the aromatic region (6.0 – 8.0 ppm).[61] The long acquisition time of the COSY pulse sequence is a disadvantage. On the other hand, *J*-resolved NMR can acquire a spectrum within 20 minutes. The pulse sequence of the *J*-resolved experiment is shown in Figure 1.5C. In this sequence, after a relaxation delay d1, a 90° pulsed is applied followed by an increment delay d0 and a 180° pulse followed by a second increment delay and the FID signal is recorded.[61,62]

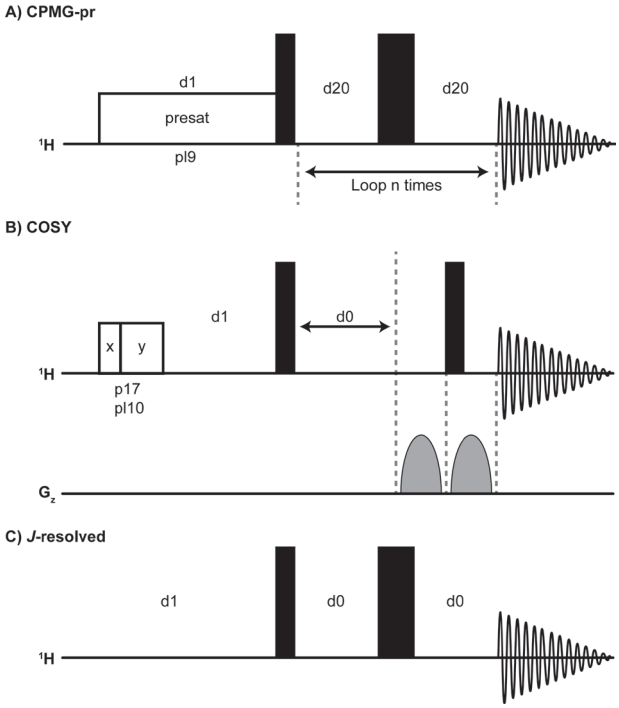


Figure 1.5: Pulse sequences for (A) the CPMG ("cpmgpr1d"), COSY ("cosygpppqf") (B) and J-resolved ("jresqf") (C) experiments. Narrow and wide black bars indicate 90° and 180° pulses, respectively. d0 is the increment delay, d1 is the relaxation delay and d20 is the fixed echo time to allow elimination of J modulation effects. pl9 is the power level for presaturation. pl10 is the power level of TOCSY-spinlock and p17 is the trim pulse at pl10.

1.7 Pre-processing of one-dimensional HR-MAS NMR spectra

Prior to multivariate analysis and quantification, raw spectra need to be pre-processed. Incorrect pre-processing can lead to spurious results.[64,65] For one-dimensional ^{1}H NMR spectra, pre-processing involves alignment, baseline correction, bucketing, normalization and scaling.

Spectral alignment

NMR resonances can be shifted due to several factors such as changes in pH, temperature, salt concentration and inhomogeneous magnetic fields. This can give rise to variations between spectra collected from the same sample species. To solve this problem, standard chemical shifts can be used for the metabolites, and the spectra can be aligned to the standard to construct a dataset for multivariate analysis.[61,65] A more elegant, unbiased protocol to align the spectra is by using an internal shift reference, since this leaves the relative shifts unaffected. This is done by adding a reference compound with a known chemical shift with the sample. Most often 3-(trimethylsilyl)-2,2',3,3'-tetradeuteropropionic acid (TSP) or 4-4-dimethyl-4-silapentane-1-sulfonic acid (DDS) is used as a reference compound. Both compounds have a methyl resonance with 0 ppm chemical shift relative to tetramethylsilane (TMS), the standard reference across the entire field of ^{1}H NMR spectroscopy.[64,65]

Baseline correction

The NMR responses of metabolites are superimposed on a broad background that does not contribute any signal of interest but affects the multivariate analysis and impedes quantification of metabolites. Polynomial-fitting of the regions in between the NMR signals is used to perform automated baseline correction.[65] After baseline correction, the spectra are truncated to have only signals from metabolites. The region between 0.1 and 8 ppm is used for further analysis. Although water suppression is employed during acquisition, a weak remaining water signal can interfere with the multivariate data analysis and the region of the water peak around 4.8 ppm is also removed.[64-66]

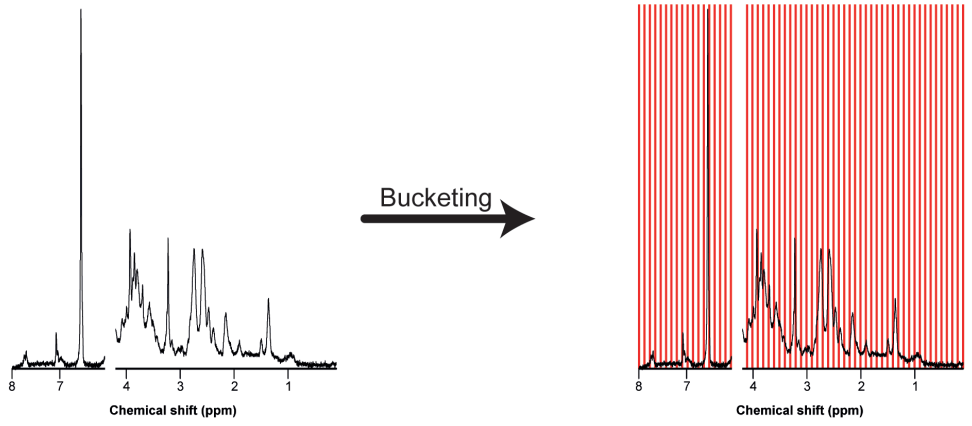


Figure 1.6: Truncated NMR spectrum before and after bucketing into equally spaced buckets of 0.04 ppm width. Bucketing allows for moderate shift averaging at the expense of resolution and provides a matrix for further processing (Figure 1.7).

Bucketing

The truncated NMR spectra typically have around 22.000 data points. It is common to reduce the resolution of the data by bucketing, also known as binning.[64-66] The spectral intensities are summed over equally spaced buckets of 0.04 ppm width (Figure 1.6). This procedure averages minor variations in chemical shift and reduces the amount of data for the multivariate analysis.[64-66] After bucketing, an $i \times j$ data matrix X is obtained with on the rows the different samples, while the columns represent the chemical shifts. The elements of the matrix contain the intensity of the bins, i.e. the signal at the different shifts for each sample.

Normalization

Biological differences between preparations, for instance different weight or dilution, result in different concentrations of specific metabolites. Normalization methods aim to remove such systematic errors.[64,65] A standard method is to normalize the individual samples (i.e. rows) of the bucket matrix X according to

$$x_{ij} = \frac{x_{ij}}{\sum_1^j x_i} \tag{1.5}$$

This is illustrated in Figure 1.7 for a simple case of three samples.

Scaling

Since higher concentration metabolites generally also exhibit the strongest variation, scaling of the columns is necessary to avoid only the selection of the most abundant metabolites in the multivariate analysis.[65]

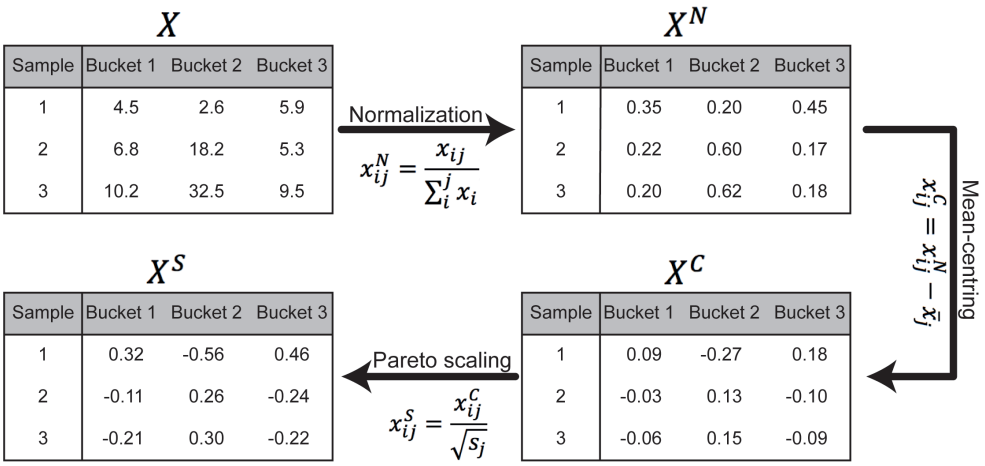


Figure 1.7: Every data point in the bucket matrix X ($i \times j$) is normalized. Every column of the normalized data matrix is mean centred and scaled using the Pareto method. The obtained data matrix X^S is used for multivariate analysis. x_{ij} is an element located in the ith row and the jth column. $\bar{x}_j$ and s_j are respectively the mean and standard deviation of the values of the jth column.

In this thesis, the Pareto scaling method has been used (Figure 1.7).[65] The first step in Pareto scaling is mean-centring of the samples, where the high-concentration and low-concentration metabolites are converted to values which vary around zero by subtracting the mean values from the columns. Pareto scaling uses the square root of the standard deviation from the columns as a scaling factor. This provides data X^S which is closely related to the real data to study the covariance of the data matrix in multivariate analysis.[65-67]

1.8 Multivariate analysis

Multivariate analysis considers multiple variables simultaneously to identify patterns in the data corresponding to signal patterns from metabolites.[65-68] These generally contain more than one proton, and their signals are therefore spread over several buckets. First, unsupervised methods, methods with no assumption of any prior knowledge, are used to explore the data, find outliers and group the data.[65-68] In this thesis, unsupervised principal component analysis (PCA) is performed, where an orthogonal transformation is used to convert the set of correlated intensities (Bucket 1, Bucket 2, ..., Bucket n) with coordinates x_{ij}^S for the samples into a set of linearly uncorrelated intensities called principal components $(\mathbf{PC_1}, \mathbf{PC_2}, ..., \mathbf{PC_n})$. PCA operates with two mathematical constraints, largest possible variance and orthogonality. The first principal component $\mathbf{PC_1}$ has the largest possible variance under the linear transformation. The subsequent vectors $\mathbf{PC}_i$ are orthogonal to the preceding components and each has the highest possible variance in their coordinates under the constraints of the prior vectors $(\mathbf{PC_1}, ..., \mathbf{PC_{i-1}})$.[65,67,69] PCA is converting the correlated X^S into an uncorrelated orthogonal basis set of vector components $(\mathbf{PC_1}, \mathbf{PC_2}, ..., \mathbf{PC_n})$, containing the scores, the new coordinates of the samples. Scores are represented in a two-dimensional score plot where each point represents a single sample on two principal component coordinates (Figure 1.8A). The transformation matrix that provides the connection with the data after preprocessing is named the loadings and describe how the old bucket intensities are linearly combined to the principal components and indicate which buckets have the most influences on the principal component and can

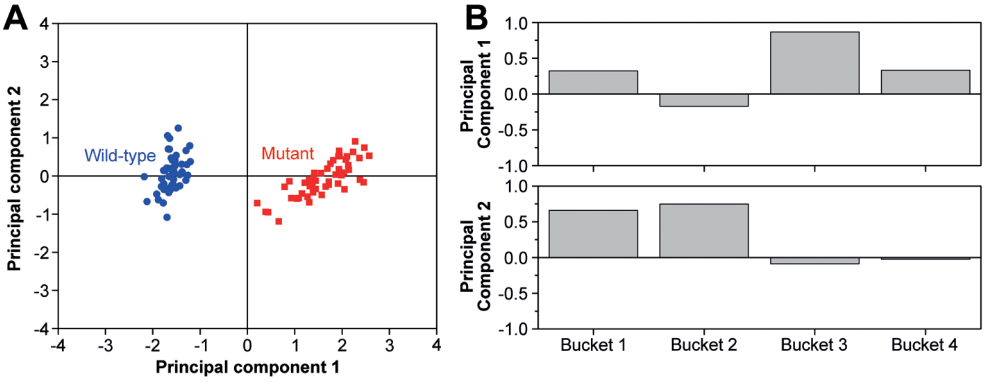


Figure 1.8: PCA score (A) and loading plot (B) of a data set including 50 wild-type samples and 50 mutant samples and 4 buckets for every sample. The score plot shows a clear separation between wild-type and mutant. The PCA loading plot shows that bucket 3 has the most influences on the first principal component and bucket 1 and 2 has the most influences on the second principal component.

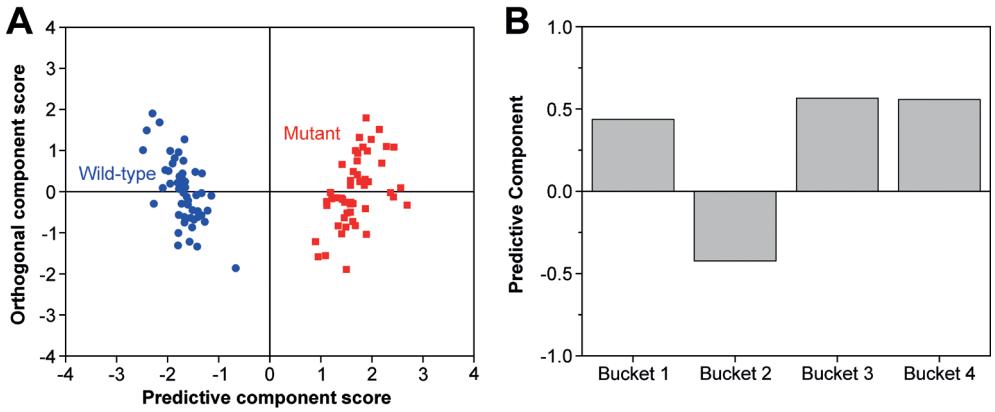


Figure 1.9: OPLS-DA score (A) and loading plot (B) for the same data set as described in Figure 1.8. In the OPLS-DA model, there is also clear separation between wild-type and mutant in the score plots. The loading plot shows that bucket 3 and bucket 4 have the most influence on the between-class variation.

be represented into a loading plot (Figure 1.8B).[67,70-72] The next step is to use databases, like the biological magnetic resonance bank (BMRB) and the human metabolome database (HMDB), to identify the metabolites corresponding to these buckets and to perform further downstream systems biology analyses.[73]

Supervised methods is used to cluster the data and to determine biomarkers by following how clusters of buckets representing a specific metabolite change between e.g. wild-type and a specific mutant. The model is applied with a priori knowledge of sample classes. Supervised methods can, therefore, be used to mark the separation between two or more sample classes at the level of individual metabolites.[65,67,69] In this thesis, orthogonal partial least squares discriminant analysis (OPLS-DA) has been utilized. OPLS-DA is a multiple regression method which uses the pre-processed data matrix X^S and a newly defined vector $\mathbf{y}$ with the value 0 for the wild-type samples and 1 for the mutant.[74] The data matrix X^S is split into a part correlated to $\mathbf{y}$, also named the predictive component (X_p^S), and another part that is uncorrelated to $\mathbf{y}$, also called the orthogonal component (X_o^S)[66,69,71-74]:

$$X^S = X_p^S + X_o^S = T_p P_p^T + T_o P_o^T \qquad (1.6)$$

In the formula above, T represents the score matrix and P the loading matrix which can be represented respectively in a score plot and a loading plot (Figure 1.9). The loading plot of the predictive component represents the between-class variation, *i.e.* wild-type vs mutants, and indicates which buckets have the strongest impact on the variation. The metabolites corresponding to these buckets are identified using metabolome databases.[73] Description of the technical details of the OPLS-DA procedure falls outside the scope of this thesis, and the reader is referred to Trygg and Wold.[75]

1.9 Applications of HR-MAS in plant metabolomics

HR-MAS NMR combined with multivariate analysis can be a powerful tool to study plant metabolomics. However, HR-MAS NMR is not used very often in plant biology. Over the last decades, ±30 publications report on HR-MAS NMR-based metabolomics studies in plants. These publications are summarized in Table 1.2.

The HR-MAS NMR-based metabolomics studies in plants have been used for wide range of applications. The influences of biotic and abiotic stress on the metabolic profile has been predominantly studied by in HR-MAS NMR. Examples are the influences of drought on plants[76-79] and the effect of fungicides or pesticides on the plants.[80-85] Metabolomics can help in understanding developmental processes, like fruit ripening. The metabolic profile throughout the ripening process is studied in mango[86] and tomato[87]. The impact of storage time on the metabolic profile is studied on Golden Delicious apples.[88] HR-MAS NMR-based metabolomics can also be used to study the metabolic profile of specific cell type to understand the plant better. Mucci *et al*. studied different tissues of lemons and citrons to understand the similarities and the differences between these two fruits.[89] In addition, it is possible to use metabolic profiling to characterize newly discovered plants[90] or mutants of plants.[91-93] The geographical origin of sweet peppers[94], garlic[95] and cocoa beans[96] has been investigated with HR-MAS NMR. The original geographical origin of some food products has been certified, for example with Protected Geographical Indications (PGIs). HR-MAS NMR-based metabolomics is a useful tool for these certified products to avoid fraud.[44] Examples are the cherry tomatoes of Pachino[97,98], Interdonato lemon of Messina[98,99] and tomatoes from Almería.[100] HR-MAS NMR can also be used to determine different classes or cultivars of plants. This is helpful when only one class has a medical application as in the case of *Trichilia catigua*[101] or *Withania somnifera*.[102] It also can help to distinguish different cultivars of apples[103], melons[104], rice[105] and persimmons.[106]

Table 1.2: Summary of the publications of metabolomics studying using high-resolution magic angle spinning NMR

Plant	Research objective	Magnetic field strength (MHz)	Pulse sequences	Multivariate models
Influences of biotic or abiotic stress				
Winter wheat (*Triticum aestivum*)[76]	Evaluate the influences of different drought treatments	400	1D	PCA
Soybean[77]	Determine the influences of water deficiency	600	CPMG, NOESY	PLS-DA
Jatropha curcas[78]	Determine the impacts of pruning procedures and water management	400	Zg	-

Ribes nigrum[79]	Determine the effect of seasonal asymmetric warming	600	CPMG, HSQC	PCA
Jatropha curcas[80]	Studying the effect of *Jatropha* mosaic virus	400	NOESY, CPMG, COSY	-
Pear (*Pyrus communis*) and Quince (*Cydonia oblonga*)[81]	Study the effect of humic acid on the morphogenesis	400	CPMG, COSY, TOCSY, HSQC	PCA
Lettuce (*Lactuca sativa*)[82]	Influences of the fungicide mancozeb on the leaves at different growth stages	800	NOESY, TOCSY, HSQC	PCA, PLS-DA
Tomato (*Solanum lycopersicum*)[83]	Study the influences of 6-pentyl-2H-pyran-2-one and harzianic acid on the leaves	400	CPMG, COSY, TOCSY, *J*-res, HSQC, HMBC	PCA
Maize (*Zea mays*)[84]	Determine the toxic effects on maize root tips of organo-chlorine pesticides	600	CPMG	OPLS-DA
Maize (*Zea mays*)[85]	Determine the effect of mineral or compost fertilization and inoculation with arbuscular mycorrhizal fungi	400	CPMG, COSY, TOCSY, *J*-res, HSQC, HMBC	PCA
Study the ripening and storage of fruits				
Mango fruit (*Mangifera indica*)[86]	Studying the metabolic profile of mango pulp during ripening	400	^{1}H 1D, TOCSY, *J*-res	-
Tomato (*Solanum lycopersicum*)[87]	Studying different tissues of the tomato during fruit ripening	500	NOESY, TOCSY, HMQC	PCA
Golden Delicious apples[88]	Determine the impact of storage time and production systems	500	NOESY, COSY, TOCSY	PCA, PLS-DA
Studying different cell types of plants				
Lemon (*Citrus limon*) and Citron (*Citrus medica*)[89]	The metabolic profile of different parts of the lemon and citron are studied	400	^{1}H, CPMG, COSY, TOCSY, HSQC	-
Characterizing of plant				
Crocus sativus[90]	Establish the main metabolites present in *C. sativus* petals	400	^{1}H, COSY, TOCSY, HSQC, HMBC	-
Understanding transgenic plants				
Poplar tree (*Populus tremula*)[91]	Studying the time- and growth-related metabolic profile of PttMYB76 and wild-type poplar tree	500	CPMG	PCA, PLS-DA

Common bean (*Phaseolus vulgaris*)[92]	Distinction between wild-type and transgenic common beans	500	CPMG	PCA
"Swingle" citrumelo[93]	Evaluate the metabolic profile of non- and transgenic citrumelo	500	^{1}H, HSQC, TOCSY	PCA, PLS-DA
Geographical origin of plants				
Sweet peppers (*Capsicum annum*)[94]	Discriminate peppers according to their geographical origin	400	NOESY, 1D ^{13}C, TOCSY	PLS-DA
Garlic (*Allium sativum*)[95]	Characterisation of two varieties cropped in different regions	400	NOESY, ^{13}C, TOCSY, HMQC	PLS-DA
Cocoa beans[96]	Assess the geographical origins of fermented and dried cocoa beans	400	^{1}H	PCA, PLS-DA, OPLS-DA
Cherry tomatoes of Pachino[97]	Determine the major metabolites present in cherry tomatoes of Pachino	700	^{1}H	PCA
PGI Cherry Tomato of Pachino, PGI Inter-donato Lemon of Messina, Red Garlic of Nubia[98]	Identifing and quantifying metabolites from 3 typical food products of the Mediterranean diet	700	^{1}H	PCA
PGI Interdonato Lemon of Messina[99]	Determine metabolites unique for PGI Interdonato Lemon of Messina	700	^{1}H, COSY, TOCSY, HSQC	-
Tomato (*Lycopersicon esculentum*)[100]	Establish differences between commercially available varieties	500	NOESY, HSQC	PCA
Distinguish between different cultivars				
Trichilia catigua[101]	Classification of commercial samples of Catuaba	400	CPMG	PCA, HCA
Withania somnifera[102]	Evaluate metabolic profile of different chemotypes	800	CPMG, COSY, HSQC	PCA
Apples[103]	Discriminate 3 different apple cultivars by their metabolic profile	500	NOESY, COSY, TOCSY	PCA, PLS-DA
Melon (*Cucumis melo*)[104]	Quantification of sugars and compare two varieties	400	^{1}H	-
Rice (*Oryza sativa*)[105]	Determine the metabolic variation of diverse rice cultivars	700	CPMG, TOCSY, HSQC, STOCSY	PCA, OPLS-DA
Persimmon (*Diospyros kaki*)[106]	Follow the metabolic changes during development of different cultivars	400	NOESY	PCA

1.10 Outline of this thesis

The purpose of this thesis is to understand the mechanisms of enhanced growth characteristics the phenotypically engineered *Arabidopsis* mutants by establishing a non-invasive metabolic approach using HR-MAS NMR combined with multivariate analysis.

In **Chapter 2**, the *in vivo* metabolic profile has been established directly from the *Arabidopsis thaliana* leaves using HR-MAS NMR throughout the circadian cycle to reveal primary metabolites and their functional periodicity of the circadian rhythm. In **Chapter 3**, HR-MAS NMR is used to obtain the metabolic profile from *Arabidopsis thaliana* mutants with enhanced growth characteristics at the middle of the light period. Combined with multivariate analysis, this approach suggests that both mutants have a diminished defence response leading to an altered growth-defence trade-off. Physiological functions like growth and defence are regulated by the circadian rhythm. In **Chapter 4**, the circadian rhythm of the identified metabolites is studied in both *Arabidopsis* mutant with enhanced growth characteristics. This shows that the circadian rhythm of the metabolites was not affected, only the levels of the metabolites differ throughout the circadian cycle. **Chapter 5** provides a general discussion of the work presented in this thesis and future prospects.

1.11 References

[1] UNFCCC, Report No. FCCC/CP/L.9/Rev.1 2015.
[2] UNGA Resolution ARES 2015.
[3] United Nations, Department of Economic and Social Affairs, Population Division, World Population Prospects: the 2017 Revision, Key Findings and Advance Tables, 2017.
[4] International Energy Agency, Tracking Clean Energy Progress, 2017.
[5] International Energy Agency, Technology Roadmap; Delivering Sustainable Bioenergy, 2017.
[6] C. Mba et al., Agriculture & Food Security 2012, 1, 7–17.
[7] A. Kumar et al., OMICS 2015, 19, 581–601.
[8] A. S. Edison et al., Metabolites 2016, 6, 8.
[9] M. Koornneef, D. Meinke, Plant J. 2010, 61, 909–921.
[10] U. Krämer, eLife Sciences 2015, 4, e06100.
[11] X. Zhang et al., Nat. Protocols 2006, 1, 641–646.
[12] J. M. Van Norman, P. N. Benfey, Wiley Interdiscip Rev Syst Biol Med 2009, 1, 372–379.
[13] F. Breseghello, A. S. G. Coelho, J. Agric. Food Chem. 2013, 61, 8277–8286.
[14] N. van Tol, B. J. van der Zaal, Plant Science 2014, 225, 58–67.
[15] T. K. Mohanta et al., Genes (Basel) 2017, 8, 399.
[16] B. I. Laufer, S. M. Singh, Epigenetics & Chromatin 2015 8:1 2015, 8, 34.
[17] T. Gaj et al., Trends in Biotechnology 2013, 31, 397–405.
[18] A. Beltran et al., Assay Drug Dev Technol 2006, 4, 317–331.
[19] T. Sera, Advanced Drug Delivery Reviews 2009, 61, 513–526.
[20] D. J. Segal et al., Proc. Natl. Acad. Sci. U.S.A. 1999, 96, 2758–2763.
[21] B. I. Lindhout et al., The Plant Journal 2006, 48, 475–483.
[22] A. Bent, in Agrobacterium Protocols (Ed.: K. Wang), Humana Press, Totowa, 2006, pp. 87–104.
[23] N. van Tol et al., PLoS One 2017, 12, e0174236.
[24] N. van Tol et al., Sci. Rep. 2017, 7, 3314.
[25] N. van Tol et al., Plant Cell Environ 2016, DOI 10.1111/pce.12805.
[26] N. van Tol, Phenotypic Engineering of Photosynthesis Related Traits in *Arabidopsis Thaliana* Using Genome Interrogation, 2016.
[27] C. Eynck et al., in Biofuel Crop Sustainability, John Wiley & Sons, Ltd, 2013, pp. 165–204.

[28] H. I. Gomes, Environmental Technology Reviews 2012, 1, 59–66.
[29] J. S. Yuan et al., Trends in Plant Science 2008, 13, 165–171.
[30] T. Ogura, W. Busch, Annu. Rev. Cell Dev. Biol. 2016, 32, 103–126.
[31] B. P. Sheth, V. S. Thaker, Planta 2014, 240, 33–54.
[32] G. W. Bassel et al., Plant Cell 2012, 24, 3859–3875.
[33] A. Dix et al., Clin Microbiol Infect 2016, 22, 600–606.
[34] E. K. F. Chan et al., PLoS Genet 2010, 6, e1001198.
[35] O. Fiehn, Plant Mol Biol 2002, 48, 155–171.
[36] H. K. Kim et al., Nat Protoc 2010, 5, 536–549.
[37] N. Schauer, A. Fernie, Trends in Plant Science 2006, 11, 508–516.
[38] I. Ahuja et al., Trends in Plant Science 2010, 15, 664–674.
[39] J. W. Allwood et al., Physiologia Plantarum 2008, 132, 117–135.
[40] J. J. Jansen et al., Metabolomics 2008, 5, 150.
[41] C. Simó et al., Int. J. Mol. Sci. 2014, 15, 18941–18966.
[42] R. Kumar et al., Front. Plant Sci. 2017, 8, 443.
[43] S. Brizzolara et al., Postharvest Biology and Technology 2017, 127, 76–87.
[44] E. Cubero-Leon et al., Food Chemistry 2018, 239, 760–770.
[45] D. I. Ellis et al., Innovation in food science - Foodomics technologies 2016, 10, 7–15.
[46] A.-H. M. Emwas, in Metabonomics, Humana Press, New York, NY, 2015, pp. 161–193.
[47] A.-H. M. Emwas et al., Metabolomics 2013, 9, 1048–1072.
[48] A.-H. M. Emwas, in Metabonomics: Methods and Protocols (Ed.: J.T. Bjerrum), Springer New York, New York, NY, 2015, pp. 161–193.
[49] J. L. Markley et al., Current Opinion in Biotechnology 2017, 43, 34–40.
[50] F. Matsuda, Mass Spectrometry 2016, 5, S0052.
[51] T. S. Maier et al., Plant Methods 2010, 6, 6–6.
[52] J. Z. Hu, Metabolomics : open access 2016, 6, e147.
[53] O. Beckonert et al., Nat Protoc 2010, 5, 1019–1032.
[54] A. Alia et al., Photosynth Res 2009, 102, 415–425.
[55] P. Mazzei, A. Piccolo, Chemical and Biological Technologies in Agriculture 2017, 4, 11.
[56] M. Vermathen et al., Chimia (Aarau) 2012, 66, 747–751.
[57] C. Deborde et al., Progress in Nuclear Magnetic Resonance Spectroscopy 2017, 102-103, 61–97.
[58] J. Kruk et al., Appl Magn Reson 2017, 48, 1–21.
[59] B. Elena-Herrmann, in NMR-Based Metabolomics, Royal Society of Chemistry, Cambridge, 2018, pp. 22–38.
[60] A. Le Guennec et al., Anal. Chem. 2017, 89, 8582–8588.
[61] A. C. Dona et al., Comput Struct Biotechnol J 2016, 14, 135–153.
[62] C. Ludwig, M. R. Viant, Phytochem. Anal. 2009, 21, 22–32.
[63] T. De Meyer et al., Anal Bioanal Chem 2010, 398, 1781–1790.
[64] L. R. Euceda et al., Scand J Clin Lab Invest 2015, 75, 193–203.
[65] A. Smolinska et al., Analytica Chimica Acta 2012, 750, 82–97.
[66] K. H. Liland, TrAC Trends in Analytical Chemistry 2011, 30, 827–841.
[67] B. Worley, R. Powers, CMB 2012, 1, 92–107.
[68] X. Qi et al., Eds., Plant Metabolomics, Springer Netherlands, Dordrecht, 2015.
[69] I. T. Jolliffe, J. Cadima, Phil. Trans. R. Soc. A 2016, 374, 20150202.
[70] J. Trygg et al., J. Proteome Res. 2007, 6, 469–479.
[71] J. Trygg, S. Wold, J. Chemometrics 2002, 16, 119–128.
[72] M. Bylesjö et al., J. Chemometrics 2006, 20, 341–351.
[73] J. J. Ellinger et al., CMB 2013, 1, 10.2174/2213235X11301010028.
[74] S. Wiklund et al., Anal. Chem. 2008, 80, 115–122.

[75] J. Trygg, S. Wold, J. Chemometrics 2002, 16, 119–128.
[76] H. Winning et al., Journal of Experimental Botany 2008, 60, 291–300.
[77] I. D. Coutinho et al., Phytochem. Anal. 2017, 28, 529–540.
[78] O. N. A. Santos et al., Industrial Crops and Products 2017, 109, 918–922.
[79] M. Pagter et al., J. Agric. Food Chem. 2017, 65, 10123–10130.
[80] O. P. Sidhu et al., Planta 2010, 232, 85–93.
[81] G. Marino et al., J. Agric. Food Chem. 2013, 61, 4979–4987.
[82] S. I. Pereira et al., Food Chemistry 2014, 154, 291–298.
[83] P. Mazzei et al., J. Agric. Food Chem. 2016, 64, 3538–3545.
[84] C. Blondel et al., Environ Pollut 2016, 214, 539–548.
[85] P. Mazzei et al., J. Agric. Food Chem. 2018, acs.jafc.7b04340.
[86] A. M. Gil et al., J. Agric. Food Chem. 2000, 48, 1524–1536.
[87] E. M. S. Pérez et al., Food Chemistry 2010, 122, 877–887.
[88] M. Vermathen et al., Food Chemistry 2017, 233, 391–400.
[89] A. Mucci et al., Food Chemistry 2013, 141, 3167–3176.
[90] V. Righi et al., J. Agric. Food Chem. 2015, 63, 8439–8444.
[91] S. Wiklund et al., Plant Biotechnology Journal 2005, 3, 353–362.
[92] R. Choze et al., Food Chemistry 2013, 141, 2841–2847.
[93] C. S. de Oliveira et al., Magn. Reson. Chem. 2014, 52, 422–429.
[94] M. Ritota et al., J. Agric. Food Chem. 2010, 58, 9675–9684.
[95] M. Ritota et al., Food Chemistry 2012, 135, 684–693.
[96] A. Marseglia et al., Food Research International 2016, 85, 273–281.
[97] D. Mallamace et al., Physica A 2014, 401, 112–117.
[98] C. Corsaro et al., J Anal Methods Chem 2015, 2015, 175696.
[99] N. Cicero et al., Nat Prod Res 2015, 29, 1894–1902.
[100] E. M. S. Pérez et al., FRIN 2011, 44, 3212–3221.
[101] C. Daolio et al., Phytochem. Anal. 2008, 19, 218–228.
[102] S. K. Bharti et al., Magn. Reson. Chem. 2011, 49, 659–667.
[103] M. Vermathen et al., J. Agric. Food Chem. 2011, 59, 12784–12793.
[104] T. Delgado-Goñi et al., Planta 2013, 238, 397–413.
[105] E.-H. Song et al., J. Agric. Food Chem. 2016, 64, 3009–3016.
[106] A. D. D. C. Santos et al., Food Chemistry 2018, 239, 511–519.