Plant Metabolomics

Vicki Maloney

The publishing of complete genome sequences has given genetics a huge boost and we are fortunate enough to have an excellent understanding of the gene to protein link. Nevertheless, our understanding of plant systems often stumbles in the gap between proteome and phenotype. We can look at what genes are being expressed, and what proteins are present, but what, for example, causes certain trees to produce more cellulose than others? Or why does a particular mutant plant only grow half as tall? Metabolomics attempts to answer these questions.

In the same way as the genome is all the genetic information in a plant, and the proteome is all the proteins, the "metabolome" is all the metabolites: metabolomics is the study of all the chemicals in a plant that have a low molecular weight. Metabolites are the end products of cellular functions, and their levels can be viewed as the response of biological systems to environmental or genetic manipulation[1]. These chemicals are integral to a better understanding of plant functions, both the everyday life of a "normal" plant, and behavior of a special plant obtained through classical breeding, mutation, or other manipulation.

Plants react to any change in their surroundings. If a bird perches on a leaf, the leaf needs to adjust its photosynthetic pathway to cope with darkness (and subsequent local nitrogen overload if the bird remains for any length of time). If a fungus attacks a leaf, the leaf will resist its attack. In both cases, everything the plant does can be followed by looking at the changes in the low molecular weight chemicals. Metabolites represent the catabolic and anabolic activities being performed by proteins at any given time. Within species or organisms, single nucleotide polymorphisms can result in small changes in the activity and level of expression of encoded proteins. The cascading effect begins with a modified genome, leading to modified proteins, and consequently, a change in the pattern and/or concentration of metabolites. Therefore, changes in the genotype will be manifested through an observed change in metabolome[2]. Quantitative and qualitative measurements of large numbers of cellular metabolites provide a broad view of the biochemical status of an organism, which can then be used to monitor and assess gene function[3]. If one has a plant that is genetically different — perhaps a mutation or the product of a breeding program- then its basal levels of metabolites and how it reacts to the environment will be different. Studies of metabolomics are aimed at linking these differences to the genetic differences that caused them, however indirectly. Linking the genome and the proteome to the metabolome is one of the major interests of modem plant science. The challenges are first how to measure all these chemicals simultaneously, and second, how to make sense of the vast amount of measurements.

The development of metabolomics

Metabolomics developed from metabolic profiling. In the early 1970's Gas Chromatography - Mass Spectrometry (GC-MS) technologies were used to analyze steroids, acids, and neutral and acidic urinary drug metabolites[4,5,6]. Soon afterwards, the concept of using metabolic profiles to screen, diagnose, and access health began to spread[7,8]. Metabolic profiling research remained stable in the 1980's with approximately 10 to 15 publications a year. With this increase in publications came a divergence in the use of the new technology. However, it was not until the early 1990's that metabolic profiling was first used as a diagnostic technique in plant systems[9].

At the turn of the century, many genome sequencing efforts were underway or near completion, and it soon became clear that a large number of the genes that were being sequenced could be assigned a function[10,11]. It then became apparent that a closer

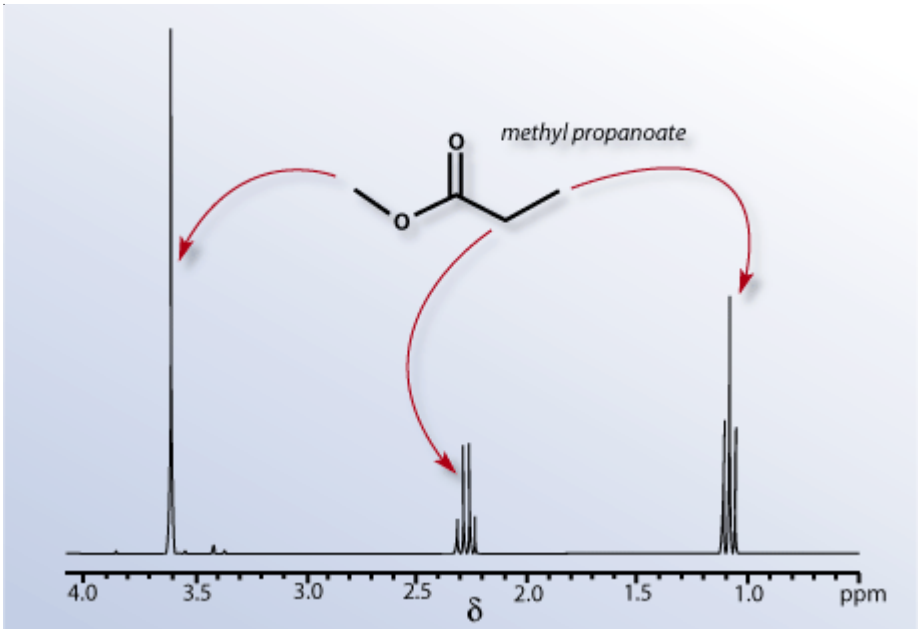

Figure 1. An NMR spectra of methyl propanoate. The structure of each peak is determined by interference from hydrogens on neighbouring carbons.

study of proteins (proteomics) might also be an effective means with which to study gene function. While it did not happen immediately, this thought process eventually trickled down to the consideration of the metabolome[12]. Researchers at the Max Planck Institute in Germany (Trethewey, Willmitzer, Fiehn, Femie) then pioneered this approach for plants based on the methods described by Sauter[9]. Since then, the number of academic and commercial groups using and entering this field has grown exponentially.

Metabolomic technologies

The metabolome is very diverse. It includes lipid soluble chemicals that are normally found in membranes, polar chemicals from aqueous parts of the cell, acidic and basic ions, stable structures and structures that oxidize at the slightest mistreatment. Until a universal measuring machine materializes, anyone working in metabolomics will have to make compromises. An extraction method and machine must be carefully chosen to suit particular interests. The typical equipment used includes nuclear magnetic resonance (NMR)[13,14,15], fourier transform infrared (FT-IR) spectroscopy[16,17] and mass spectrometry, often combined with chromatography[3,18,19].

Chromatography and Mass Spectrometry

The most widely used and powerful methods used to profile low molecular weight chemicals are based on chromatographic separation, followed by detection and validation by mass spectroscopy. These include gas chromatography mass spectroscopy (GC-MS) and more recently, liquid chromatography mass spectroscopy (LC-MS) [20]. GC-MS and LC-MS in combination are able to detect several hundred chemicals, including sugars, sugar alcohols, organic acids, amino acids, fatty acids and a wide range of diverse secondary metabolites[3,20,21]. GC-MS and LC-MS are especially useful in differentiating chemical isomers such as those of common hexoses, which all have the same mass[2]. While GC-MS is a low cost alternative to some other metabolomic technologies and provides high separation efficiencies, it requires that samples be volatile. This requirement is readily accomplished by chemical derivitization, but at the cost of additional time and variance. In LC-MS a derivitization step is not essential, and selection of the compound comes from either the extraction method

or the type of column used. However, LC-MS profiling methods all rely to great extent on comparisons with reference compounds. Often one is able to identify the class of compounds the metabolite belongs to, but not its exact identity.

Nuclear magnetic resonance

NMR methods provide metabolite fingerprints with good chemical specificity for compounds containing elements with non zero magnetic moments, such as ^{1}H, ^{13}C, ^{15}N, and ^{32}P which are commonly found in most biological metabolites[13]. However, ^{1}H NMR spectra of plant extracts are inevitably crowded not only because there is a large number of contributing compounds, but also because of the low overall chemical shift dispersion (chemical shift is caused by the chemistry around the nucleus changing its resonance). Increased specificity and resolution can be achieved by using higher magnetic fields. Thus, NMR spectra of unpurified solvent extracts of plants has the potential to provide a relatively unbiased fingerprint, containing overlapping signals of the majority of the metabolites present in the solution.

Fourier transform infrared spectroscopy

Fourier transform infrared (FT-IR) spectroscopy primarily measures vibrations of functional groups and highly polar bonds. These (bio)chemical 'fingerprints' are made up of the vibrational features of all the sample's components. FT-IR spectrometers record the interaction of IR radiation with experimental samples, measuring the frequencies at which the sample absorbs the radiation and the intensities of the absorptions. Determining these frequencies allows identification of the sample's chemical makeup since chemical functional groups are known to absorb light at specific frequencies[16,17]. However, there is a lack of differentiation between isomers, and the presence of fragment and adduct ions and problems of quantification caused by ion suppression make the use of powerful classification tools essential. The combination of FT-MS data with other chromatographic MS data is potentially powerful[22].

Metabolomic analysis

Currently, there is no method to extract all metabolites and measure them all. Additionally, there is always a risk that some metabolites will be lost along the way[23,24,25]. Nevertheless, advances in technology now allow much larger ranges to be measured than ever before. These advances in technology make it necessary to improve bioinformatics and data handling methods: there is no point in collecting numbers if we can not process them meaningfully. While metabolomics is still in its infancy, there are a number of bioinformatics tools currently being used, and a number in development. These tools are used to align chromatographic data and differentiate components in large datasets.

Correlation optimized warping (COW)

Until recently, most analysis of metabolomic data was performed on reduced data sets using areas of selected peaks detected in the chromatograms. We look at a pathway and guess what should have changed in it when the plant was subject to some sort of manipulation. Of course, the plant often has compensated for the original change in some way, and the changes we see are often very different from what we expect. Significant changes may occur in pathways and processes quite far from where we were looking. This is where a broader approach pays off; if we only measure the metabolites where we expect to see a change, we may end up with uninterpretable data and have no idea what happened, because we missed the true consequences of what was going on. This reduction in dataset size also introduced the problem of extracting specific peak data from each chromatogram. These disadvantages can be overcome by using the entire dataset that is generated from the chromatographic and/or spectroscopic instruments, but, as small unavoidable differences in experimental conditions can cause changes and drift in peaks, it is necessary to align all of the peaks.

There are a number of approaches that have been formulated to tackle this problem.[26,27] Unfortunately, they still failed to use the entire data matrix or still relied on preprocessing and peak detection. COW aims to align two chromatographic profiles by

piecewise linear stretching and compression - also known as warping — of the time axis of one of the profiles. The optimal alignment will be determined by calculation of correlation, which means that no knowledge of the compounds present is required, and individual peaks do not need to be detected. COW uses two input parameters (section length and flexibility), which can be estimated from the observed peak width. COW is useful for almost any kind of chromatographic data[28].

Chrompare

Chrompare is a more recently developed software system which allows the alignment of chromatographic data. This system includes methods for manual and automated correction of retention times and responses by corresponding standards. This comparison is based on a univariate peak-by-peak approach using peak height areas and corresponding retention times[29].

Principle component analysis (PCA)

Even after the datasets have been aligned, there is much effort involved in getting any information from the data. A single GC-MS metabolite profile can yield 300-500 distinct compounds[2]. There is a wealth of information that can be gained from this, but the challenges lie in the data processing. One challenge is determining exactly what those distinct compounds are. Many problems, however, can be solved without knowing what the chemicals are. PCA uses basic vectors that span an n-dimensional space to give the best sample separation. The concept behind PCA is to describe the variance in a set of multivariate data in terms of a set of underlying orthogonal variables (principle components). The original variables (metabolic concentrations) can be expressed as a particular linear combination of linear components (loadings). PCA is an additive model, in the sense that each principal component (PC) accounts for a portion of the total variance Of the data set. Often, a set of PCs (2 or 3) account for over 90% of the total variance, and plotting just those PC's can effectively reduce the dimensions of the dataset and provide a rapid means of visualizing similarities or differences in the data set. The same approach can be used in the characterization of unknown mutants. Most mutants have changes in genes where there are already known mutants that have that same gene change. It is wasted effort to study every gene in detail, when different mutations will often have identical effects. A mutation that turns off the gene responsible for the red color of a flower will always yield a flower that is not red, whether the mutation is at one end of the gene or the other. A particular mutation is likely to give a particular pattern of metabolites, and if we recognize the pattern, we can recognize that mutation; to do so we do not need to identify which chemicals are responsible for each bit of the pattern. However, if there is a need to identify the components of a profile, plots of the loadings themselves may be used to explore which compounds are most responsible for separating samples into groups, as the most important compounds (peaks) tend to correspond to high absolute loading values.

Hierarchical cluster analysis (HCA)

Hierarchical cluster analysis is an agglomerative methodology that finds clusters of observations within a data set. Three of the better known algorithms for clustering are average linkage, complete linkage, and single linkage. Generally, each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster, which replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster remains. The different algorithms differ in how the distance between two clusters is computed[2].

Metabolite identification

The identification of unknown metabolites from chromatographic data can be quite difficult. There are many databases developed for mass spectra, but few of the precursor molecules in plant biosynthesis pathways have been previously identified, and therefore most are not present in these databases. Also, the need to derivitize samples before they are injected into the GC causes the resulting fragments to be dominated by the derivitized group and identification of the original intact chemical becomes difficult. Furthermore, GC's

with electron impact ionization use high energies, which disrupt chemical bonds and leave fragments that are characteristic of the chemical structures, but without some of the ions which represent the intact structure. This hampers any new identification of unknown metabolites. One way to overcome these problems is to artificially synthesize the precursor molecules in the pathway that one studies and test for matches within the data. While this approach is time consuming and not always effective, coupled with PCA or HCA, it can give one a good idea of where changes in the pathway are taking place. Chemical ionization, which is a much softer technique, can also be used in combination with electron ionization in order to obtain the correct mass of the structure in question.

Applications in plant science

Plant metabolomics is still a field in its infancy, but the opportunities are almost endless. Metabolomics offers the unbiased ability to characterize and differentiate genotypes and phenotypes based on metabolite levels. The following is just a subset of the possible applications.

Characterization of metabolism

Metabolites are frequently measured to provide insight into the responses of plants to physiological or environmental changes. This is illustrated by a group of studies with a focus on measuring changes in the level of amino acids in leaves during diurnal rhythms and in response to changes in the rate of photosynthesis and photorespiration[29,30,31]. A line of tobacco was bred to have 40-45 of the wildtype nitrate reductase (NT). It was believed that the decrease in NT would decreases the nitrate assimilation and therefore decrease the growth of the plants, but the mutants grew just as fast. These studies were able to show that plants use sophisticated mechanisms to maintain relatively constant levels of central metabolites. It was also shown that a remarkable consistency in the ratio between the individual minor amino acids existed even though the overall levels varied dramatically.

Another study by Roessener et al. used GC-MS based metabolic profiling to study for independent potato genotypes characterized by modifications in sucrose metabolism. Using PCA and HCA, they were able to identify clusters for each individual plant system. Extraction analysis allowed for identification of the most important components of the clusters. This data confirmed that the reduction in starch accumulation resulted from the partitioning of carbon flux into glycolysis. In a second approach, they used wildtype potato tissue that was subjected to environmental manipulations. These plants also separated in the PCA. These data demonstrate the use of metabolic profiling in conjunction with data mining tools as a technique for the comprehensive characterization of a plant genotype[21].Integration with transcript and protein profiling Integrating metabolic profiling data with transcript and protein profiles creates a multilayer characterization of the system response. Some recent reports on carbon-nitrate interactions have provided datasets that covered transcripts, enzyme activation and metabolites[24,30,31]. These studies, while being limited to a segment of metabolism, illustrate the need for a more integrated approach. These multilayer characterizations will enable us to reveal when important changes in metabolism occur independently of changes in transcription.

Identification of regulated key sites in networks

One of the first steps in elucidating the important regulatory mechanisms within a network is to carry out a systematic investigation of all metabolites in a metabolic segment. This provides an unbiased and powerful means to identify the site(s) at which metabolic flux is regulated. In a study by Tiessen et al., researchers were able identify the site at which starch synthesis is inhibited after detaching a growing potato tuber from the mother plant[32]. After investigating the subcellular level of every metabolite in the pathway between sucrose and starch, they identified ADP-glucose pyrophosphorylase (AGPase) as the unique site of regulation. Further biochemical studies showed that AGPase activity is inhibited by a novel redox-dependent post-transcriptional mechanism involving the formation of an inter-molecular cysteine bridge between the two catalytic subunits of AGPase.

Investigation of gene function

The results of a metabolic analysis can be used to define the role of a gene and to determine if expression level controls pathway activity. A challenge in functional genomics is assigning function to genes of poorly defined or unknown function. Studies of antisense plants with progressively decreasing expression of a target enzyme have shown that diagnostic changes in metabolites can often be detected in lines where the alteration of enzyme activity is too small to produce any visual phenotype[24,31,33,34]. A move towards using metabolic analyses will enable us to perform an unbiased determination between different lines and cultivars. Compared to conventional methods of phenotype determination, which are time consuming and not always accurate at low expression levels, metabolic profiling allows for subtle but potentially important differences to be detected.

Where do we go from here?

While much has been accomplished in the field of metabolomics, there is always room for improvement. New analytical techniques need to be developed that can increase the amount of the metabolome that can be sampled in each step. With an increase in the amount of metabolites sampled, it will become more important to have instruments that are integrated with peak matching software. One of the most time consuming steps in the present procedure is trying to compensate for changes that occur in the chromatogram because of minor changes in the extraction protocol and instruments over time. Therefore, it will also be beneficial to create automated procedures that can identify unknown peaks by combining the information from MS fragmentation patterns, isotope ratios, exact masses, structure generators, and biochemical databases. Once the analytical and extraction procedures have been perfected and the bioinformatics tools developed, there are endless metabolic linkage networks that need to be defined.

Conclusions

Metabolomics has the potential to make a large impact on areas of biology that extend far beyond the scope of this paper. The ultimate goal is to understand and to predict the behaviour of complex systems (such as plants) by using the results obtained from data mining tools for subsequent modeling and simulation. Metabolomics has developed to the point where it can be applied alone and in combination with other technologies of functional genomics.

References

1. Fiehn, 0. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* **48**, 155-171 (2002).

2. Sumner, L.W., Mendes, P., Dixon, R. A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817-836 (2003).

3. Fiehn, 0., Kopka, J., Dormann, P., Altmann,T., Trethewey, R.N., Willmitzer, L. Metabolic Profiling for plant functional genomics. *Nature Biotechnology* **18**, 1157-1161 (2000).

4. Devaux, P.G., Homing, M.G., Homing, B.C. Benyzl-oxime derivative of steroids; a new metabolic profile procedure for human urinary steroids. *Analytical Letters* **4**, 151 (1971).

5. Horning, B.C., and M.G. Homing. *Metabolic Profiles: chromatographic methods for isolation and characterization of a variety of metabolites in man.* In Methods in Medical Research (Olsen, R.E. Ed.) 369 (Yearbook Medical Publishers, Chicago, 1970).

6. Horning, E.C., and Homing, M.G. Human metabolic profiles obtained by GC and GC/MS. *Journal of Chromatographic Science* **9**, 129-140 (1971).

7. Cunnick, W.R., Cromie, J.B., Cortell, R., Wright, B., Beach, E., Seltzer, F., Miller, S. Value of biochemical profiling in a periodic health examination program: analyses of 1,000 cases. *Buletin of New York Academy of Medicine* **18**, 5-22 (1972).

8. Mroczek, W.J. Biochemical profiling and the natural history of hypertensive diseases. *Circulation* **45**, 1332-1333 (1972).

9. Sauter, H., Lauer, M., Fitsch, H. *Metabolic profiling of plants a new diagnostic technique*. In American Chemical Society Symposium Series (Baker, D.R., Fenyes, J.G., Moberg, W.K Eds.) (1991).

10. The Arabidopsis Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **40 8**, 796-815 (2000).

11. Yu, J., et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indiea*). *Science* **296**, 79-92 (2002).

12. Oliver, S. Yeast as a navigational aid in genome analysis. *Microbiology* **143**, 1483-1487 (1997).

13. Blingly, R. and R. Douce. NMR and plant metabolism. *Current Opinion in Plant Biology* **4**, 191-196 (2001).

14. Roberts, J.K.M. 2000. NMR adventures in the metabolic labrynth within plants. *Trends in Plant Science* **5**, 30-34 (2000).

15. Ward, J.L., C. Harris, J. Lewis, M.H. Beale. 2003. Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsi thaliana*. *Phytochemistry* **62**, 949-957 (2003).

16. Chen, L., Carpita, N.C., Reiter, W.-D., Wilson, R.H., Jeffries, C., and McCann, M.C. A rapid method to screen for cell-wall mutants using discriminant analysis of Fourier transform infrared spectra. *Plant Journal* **16**, 385-392 (1998).

17. Coimbra, M.A., Barros, A., Barros, M., Rutledge, D.N., & Delgadillo, I. Multivariate analysis of uronic acid and neutral sugars in whole pectic samples by FT-IR spectroscopy. *Carbohydr. Polym.* **37**, 241-248 (1998).

18. Fiehn, O., Kopka, J., Trethewaey, R.N. & Willmitzer, L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry* **72**, 3573-3580 (2000).

19. Stitt, M. & Fernie, A.R. From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon-nitrogen interactions. *Current Opinion in Biotechnology* **14**, 136-144 (2003).

20. Tolstikov, R.N. & Fiehn, O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectroscopy. *Annals of Biochemistry*, **301**: 298-307. (2002).

21. Roessner, U., Luedemann, A., Burst, D., Fiehn, O., Linke, T., Willmitzer, L. & Femie, A.R. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *The Plant Cell* **13**, 11-29 (2001).

22. Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L., & Bino, R. Plant Metabolomics: The Missing Link in Functional Genomics Strategies. *Plant Cell* **14**, 1437-1440 (2002).

23. Lilley R.M., Stitt, M., Mader, G. & Heldt, H.W. Rapid fractionation of wheat leaf protoplasts using membrane filtration - the determination of metabolite levels in the chloroplasts, cytosol, and mitochondria. *Plant Physiology*, **70**, 965-970 (1982).

24. Scheible W.R., Krapp, A. & Stitt, M. Reciprocal diurnal changes of phosphoenolpyruvate carboxylase expression and cytosolic pyruvate kinase, citrate synthase and NADP-isocitrate dehydrogenase expression regulate organic acid metabolism during nitrate assimilation in tobacco leaves. *Plant Cell Environment* **23**, 1155-1167 (2000).

25. Stitt M & Grosse, H. Interactions between sucrose synthesis and C02 fixation. 1. Secondary kinetics during photosynthetic induction are related to a delayed activation of sucrose synthesis. *Journal of Plant Physiology* **133**, 129-137 (1988).

26. Malmquist, G. & Danielsson, R. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *Journal of Chromatography A* **687**, 71-88 (1994).

27. Round, A. J., Aguilar, M.I. & Hearn, T.W. High-performance liquid chromatography of amino acids, peptides and proteins: CXXXIII. Peak tracking of peptides in reversed-phase high-performance liquid chromatography. *Journal of Chromatography A* **661**, 61-75 (1994).

28. Nielson, N.P.V., Carstensen, J.M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping. *Journal of Chromatography A* **805**, 17-35 (1998).

29. Masclaux-Daubresse, C., Valadier, M.H., Carrayol, E., Reisdorf-Cren, M. & Hirel, B. Dirunal changes in the expression ofglutamate dehydrogenase and nitrate reductase are involved in the C/N balance of tobacco source leaves. *Plant Cell Environment* **35**, 1451-1462 (2002).

30. Matt, P., Geiger, M., Walch-Liu, P., Engels, C., Krapp, A. & Stitt, M. The immediate cause of the diurnal changes of nitrogen metabolism in leaves of nitrate-replete tobacco: a major imbalance between the rate of nitrate reduction and the rates of nitrate uptake and ammonium metabolism during the first part of the light period. *Plant Cell Environment* **24**, 177-190 (2001).

31. Scheible, W.R., Gonzalez-Fontes, A., Morcuende, R., Lauerer, M., Geiger, M., Glaab, J., Goon, A., Schuize, E.D. & Stitt, M.Tobacco mutants with a decreased number of functional nia genes compensate by modifying the diurnal regulation of transcription, post-translation modification and turnover of nitrate reductase. *Planta* **203**, 304-319 (1997).

32. Tiessen, A., Hendriks, J.H.M., Stitt, M., Branscheid, A., Gibon, Y., Farre, E.M. & Geigenberger, P. Starch Synthesis in Potato Tubers Is Regulated by Post-Translational Redox Modification of ADP-Glucose Pyrophosphorylase: A Novel Regulatory Mechanism Linking Starch Synthesis to the Sucrose Supply. *Plant Cell* **14**, 2191-2213 (2002)

33. Stitt, M. Nitrate regulation of metabolism and growth. *Current Opinion in Plant Biology* **2**, 178-186 (1999).

34. Stitt, M. & Sonnewald, U. Regulation of metabolism intransgenic plants. *Anna Rev. Plant Phys*. **46** 341-368 (1995).

35. Frenzel, T., Miller, A. & Engel, K.H. A methodology for automated comparative analysis of metabolite profiling data. *Eur. Food. Res. Technol*. **216**, 335-342 (2003)