



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Integrated metabolome and proteome analysis for the molecular characterization of natural product formation during developmental processes of cacao seed, lotus seed and grape berry“

verfasst von / submitted by

Lei Wang

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

A 794 685 437

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Biologie

Betreut von / Supervisor:

Univ. Prof. Dr. Wolfram Weckwerth

Acknowledgements

First and foremost, I would like to thank my supervisor **Prof. Dr. Wolfram Weckwerth** for giving me such a great opportunity to study in his lab. Thank him for all his supports, guidance and especially those encouragements during my whole PhD study process. His profound knowledge in biochemistry always brings attractive ideas during discussions and his positive attitude always inspires me to present the result more confidently. His broad knowledge in music, history, literature and many other aspects makes him more than an academic tutor.

Many thanks to **Mrs. Doris Dell'mour** for helping me find a flat before I came and for helping me with all the documentary works related with my study in University of Vienna. Many thanks to **Dr. Alexander Seidel** for his technical supports in computer maintenance and troubleshooting.

I would like to acknowledge **Dr. Doris Engelmeier** and **Dr. Hannes Dörfler** for their kind help and patient introduction at the very beginning of my study in MoSys. Thank them for leading me into the metabolomics field.

I would also like to express warm appreciation to **Mag. Lena Fragner** not only for her advises and supports during my sample preparation and measurement with GC-MS but also for the wonderful trip in California we shared together.

Many greetings to **Dr. Thomas Nägele** for his help in reverse Jacobian calculation, insightful discussion about the results and great contribution to the cacao manuscript. His meticulous, diligent and efficient working style and unbreakable faith always inspire us to try our best.

With my deep gratitude to **MSc. Ella Nukarinen** for helping me “communicate” with “Mr. Orbitrap” when he does not want to work. Many thanks for her guidance with my proteomic work. It is always helpful and enlightening to discuss with her. So lucky to have her as a great, encouraging and inspiring friend in life. Thanks for those hiking trips with her which made us braver and more confident when facing challenges.

Thank **Dr. Palak Chaturvedi** for showing me her procedure of proteomic data analysis and also for sharing her experience in proteomic data exploration. Thank Mag. **Reinhard Turetschek** for his suggestion in protein sequence blast.

Many appreciations to **MSc. Jakob Weismann** for his supports in statistical analysis. Thanks him for teaching me those German words that I will never learn from a real German course. I did not learn so much aber besser als ein Stein am Schädel. Thanks him for helping me translate the abstract to German. Great appreciation to all those warm helps he offered.

Many gratefulness to all the members of **MoSys Running Team**. Thank all of them for sticking to our running schedules. Their encouragement brought me to a distance I never thought I could reach. It is a great experience to join the Vienna city half marathon with all of them.

Thank all of the supports and helps from our lovely **MoSys members**.

Many thanks to **the staff of University of Vienna**, especially those who are working in PhD administration and StudienServiceCenter for providing helpful information regarding study process, thesis writing and defense preparation.

Thanks for the financial support from **China Scholarship Council (CSC)** and **my supervisor Prof. Dr. Wolfram Weckwerth**. Thank **University of Vienna** providing me financial assistance for joining a conference in America.

最后，要感谢我的家人这么多年来给予我的关爱与鼓励。特别是我的父母，感谢他们用勤劳的双手为我和家人创造的平凡但幸福的生活。感谢爸妈用最质朴的方式呵护着我和弟弟。(Last but not least, I would like to express my deepest gratitude towards my parents and my brother for their continuous backup. Thank my parents for offering us a normal, happy life.)

Table of Contents

I. Introduction	5
1. Biological and biochemical background of plant materials	5
1.1 <i>Theobroma cacao</i>	5
1.1.1 History, distribution and cultivars of cacao	5
1.1.2 Health benefits of cacao	7
1.1.3 Chemical composition of cacao beans	8
1.2 <i>Nelumbo nucifera</i>	8
1.2.1 Botany, phytogeography, economic and cultural importance of lotus	9
1.2.2 Phytochemical constitutes and bioactivities of lotus	9
1.3 <i>Vitis vinifera</i>	10
1.3.1 History, distribution and cultivars of grape	10
1.3.2 Phytochemical constitutes and bioactivities of grape	10
2. System biology in plant research	12
2.1 Omics technologies present systemic and genome-scale overviews of plant metabolism	12
2.2 Mass spectrometry based proteomics and metabolomics	13
2.2.1 Mass spectrometry based platform	13
2.2.2 Mass spectrometry based proteomics	14
2.2.3 Mass spectrometry based metabolomics	15
2.2.4 Experimental setup for integrated metabolomic and proteomic analysis	16
2.2.5 Combinatorial omics analysis in understanding interactions and correlations of primary and secondary metabolism	17
2.3 Systematic approaches in understanding seed and fruit development	18
2.3.1 Seed development	18
2.3.2 Grape berry development	20
3. References	22
II. Publications	31

1. Metabolomic and proteomic profiles of developing cacao seed	32
1.1 Preamble.....	32
1.2 Declaration of authorship	32
1.3 Published manuscript	32
2. Metabolome and proteome of developing lotus seed	48
2.1 Preamble.....	48
2.2 Declaration of authorship	48
2.3 Published manuscript	48
3. mzGroupAnalyzer	61
3.1 Preamble.....	61
3.2 Declaration of authorship	61
3.3 Published manuscript	61
4. Metabolism dynamics of developing grape berry	73
4.1 Preamble.....	73
4.2 Declaration of authorship	73
4.3 Manuscript.....	73
III. Summary.....	113
IV. Zusammenfassung	116

I. Introduction*

Plants are precious resources for all life on earth. They fix carbon, produce oxygen, reserve water as well as provide habitats, food and medicine for organisms. Spermatophytes (also known as seed plants) including all the gymnosperms and angiosperms are widespread on earth. Seeds play an important role in the life cycle of spermatophytes. The plant life cycle starts with the germination of seeds and complete with the formation of new seeds. Seeds that formed from the fertilized ovules contain genetic information as well as storage compounds for reproduction. Not only essential to the life cycle of plants, seeds are also crucial food sources for human and other animals as they contain carbohydrates, protein, fatty acids, dietary fiber and minerals. Seeds of wheat, rice, and maize are the dominate source of staple foods worldwide. Some seeds are the primary source of vegetable oil either used in cooking or burned as fuel. Fleshy fruits covering seeds protect the seeds and promote seed dispersal via the contribution of herbivores. They are also important to human diet.

1. Biological and biochemical background of plant materials

Theobroma cacao, *Nelumbo nucifera* and *Vitis vinifera* that belong to Malvaceae, Nelumbonaceae and Vitaceae family (Table 1) respectively are nutritionally and economically important crops. Cacao seed, which are usually called cacao beans, are the source for chocolate production. Lotus seed is a traditional Chinese medicine and also a popular snack and cusine ingredient in Asia countries. Grape berry is a popular fruit worldwide and either eaten freshly or made into raison, juice and wine. These three crops are the main study objects in this thesis. All of these three crops have long cultivation histories and possess great economic importance. The genomes of grape, cacao and lotus were sequenced in 2007, 2011 and 2013, respectively (Jaillon et al., 2007; Velasco et al., 2007; Argout et al., 2011; Ming et al., 2013) which have profound impact on the understanding of the evolution and metabolism of these plants on molecular levels. Their taxonomy information is shown in Table 1.

1.1 *Theobroma cacao*

1.1.1 History, distribution and cultivars of cacao

Theobroma cacao ($2n = 20$) is one of the 22 species of *Theobroma* in the family of Malvaceae (Table 1). It is native to tropical regions of Central and South America. Its seeds, usually called cacao beans, are the material for chocolate production. Cacao has a long using history in Mesoamerica. The domestication, cultivation, and usage of cacao could be traced back to

* Citations of own work are in bold.

1900 BC (Henderson et al., 2007; Powis et al., 2011; Thomas et al., 2012). The main cacao production areas include some countries in Latin America and West Africa as well as Indonesia. The climate conditions (e.g. high temperate and humidity, adequate and well distributed rainfall, sunlight and shading conditions) in these tropical regions are appropriate for cacao growth.

Table 1. Taxonomy and genome information of cacao, lotus and grape

	<i>Theobroma cacao</i>	<i>Nelumbo nucifera</i>	<i>Vitis vinifera</i>
Order	Malvales	Proteales	Vitales
Family	Malvaceae	Nelumbonaceae	Vitaceae
Genus	<i>Theobroma</i>	<i>Nelumbo</i>	<i>Vitis</i>
Species	<i>T. cacao</i>	<i>N. nucifera</i>	<i>V. vinifera</i>
Genome sequencing	2011	2013	2007
Genome size	430 Mbp	929 Mbp	504.6 Mbp
Chromosome number	2n = 20	2n = 16	2n = 38

There are three main cultivars of cacao named as Criollo, Forastero and Trinitario. They show differences in appearance of pods and seeds, disease resistant ability, yield and flavor (Elwers et al., 2009; Pohlen and Pérez, 2010) (Table 2). The pods of Criollo cacao usually are red or yellow with plump white seeds embedded in pulp. Criollo cacao trees are susceptible to diseases and hard to grow thus have a low yield and contribute only 5% to 10% of the world chocolate production. However, Criollo beans are considered to have a delicate flavor which is less bitter and more aromatic. Consequently, Criollo cacao beans are most rare and prized cacao beans in market. Different from Criollo, Forastero cacao trees are resistant to diseases thus have a high yield and contribute to over 85% of chocolate production in the world. The pods of Forastero are yellow and smooth with flat and usually purple beans inside. The beans of Forastero with bitter taste are normally used as background flavor. Trinitario is a hybrid cultivar of Criollo and Forastero. Trinitario cacao trees are more resistant to diseases and have a higher yield than Criollo. Trinitario together with Criollo cacao beans are considered to be “flavor beans” and Forastero beans are “bulk beans” (Pohlen and Pérez, 2010). The seeds used in this study are from Forastero cacao trees.

The flavor of the cacao beans are not only determined by their genetic background, but also affected by culture condition and post-harvesting processes (fermentation, drying, roasting and conching). During fermentation, yeasts, lactic-acid and acetic acid bacteria interact with cacao beans (Schwan and Wheals, 2004; Ho et al., 2014) causing a release of volatile compounds (i.e. organic acids, alcohols, esters, and aldehydes) (Rodriguez-Campos et al., 2011) and formation of aroma precursors (i.e. amino acids, sugars) (Caligiani et al., 2007; Adeyeye et al.,

2010). In contrast, polyphenols and purine alkaloids decrease in abundance during fermentation due to oxidation and condensation of polyphenols and diffusion out of cacao beans (Sotelo and Alvaez, 1991; Caligiani et al., 2007; Camu et al., 2008; Trognitz et al., 2013). The drop in purine alkaloids and polyphenols causes the decline in astringency. The oxidation and condensation of polyphenols decrease their solubility and endow cacao seed brownish color. Moisture of the cacao beans is reduced in the following drying step to prevent the formation of molds. Roasting is another crucial step for flavor formation. Aroma precursors are converted to odorant compounds via Maillard reactions (Frauendorfer and Schieberle, 2008). Roasting under a relatively high temperature may lead to a significant degradation of tocopherols and phytosterols (Oracz et al., 2014) whereas alkylpyrazines and aldehydes are increased (Huang and Barringer, 2011).

Table 2. Features of the three main cacao varieties.

	Criollo	Trinitario	Forastero
Pods	Red or yellow	Yellow or red	Yellow
Seeds	Round & white	Round or flat	Flat & purple
Diseases	Susceptible	Resistant	Resistant
Yield	Low	Middle	High
Production	5% - 10%	10% - 15%	85%
Flavour	Sweet aroma	Fruity & acid	Fruity & bitter

1.1.2 Health benefits of cacao

In the past few years, cacao products attracted more and more attentions due to their health benefits. Polyphenols were the most frequently reported active compounds. Miller et al (Miller et al., 2006) compared the levels of total polyphenols and procyanidins in natural cacao and cacao products such as dark chocolate and milk chocolate. Their result indicated that natural cacao contains higher levels of polyphenols and procyanidins and showed stronger antioxidant activity. Antioxidant activity also contributed to counteract cytotoxicity caused by chemicals like acrylamide (Rodriguez-Ramiro et al., 2011) and dioxins (Mukai et al., 2008), neuroprotective (Ramiro-Puig et al., 2009), and antiinflammatory activities (Arteel et al., 2000). Concurrent with the rise of flavanol and procyanidin levels, an increase in antioxidant capacity and decrease in erythrocyte hemolysis was observed in plasma of rats that feed with cacao extract (Zhu et al., 2002). Several studies showed that consumption of cacao products could modestly inhibit platelet function indicating potential benefits of cacao products to patients with cardiovascular diseases (Murphy et al., 2003; Park and Schoene, 2006). Cacao fiber consumption also

showed benefits in reducing cardiovascular risks by reducing lipid peroxidation and improving the lipidemic profile in rats fed with hypercholesterolemic diet (Lecumberri et al., 2007). Additionally, consumption of flavanol-rich dark chocolate showed effects in decreasing blood pressure and LDL cholesterol which is beneficial in hypertensives (Grassi et al., 2005). The oligomer flavanol fraction showed potential antiobesity and antidiabetic bioactivities in mice that fed with high-fat diet with the evidence of preventing weight gain and ameliorating insulin sensitivity (Dorenkott et al., 2014). Another bioactivity of cacao is antimutation. *In vitro*, *ex vivo* and *in vivo* experiments indicated that polyphenol extracts of cacao inhibited carcinogenesis of pancreas (Yamagishi et al., 2000), lung (Yamagishi et al., 2003) and liver (Lee et al., 2010) but not mammary (Yamagishi et al., 2000). Phenolic-rich extracts of cacao beans inhibited drug-triggered liver cytotoxicity indicating its potential ability in hepatopathy prevention and therapy (Arlorio et al., 2009).

1.1.3 Chemical composition of cacao beans

Chemical compositions of cacao beans and cacao products are widely studied as they are direct contributors to the flavor and bioactivities. The main component of cacao seed is fat which is usually called cacao butter and widely used in chocolate and cosmetic industries. The most characteristic metabolites of mature cacao beans are purine alkaloids especially theobromine and caffeine which are responsible for the bitter taste. They are in higher content in mature seed than in cacao flowers and leaves (Sotelo and Alvarez, 1991). During cacao seed development, theobromine and caffeine are out of detection at early developmental stages and significantly accumulated during maturation (Pereira-Caro et al., 2013; Wang et al., 2016b). Alkylpyrazines are another family of nitrogen containing metabolites. They are volatile and present sweet, nutty, chocolate, caramel as well as defective aroma such as rancid and smoky sensory (Misnawi and Ariya, 2011). They are formed from sugars, amino acids and peptides mainly during roasting via Maillard reactions and their concentration varies depending on the time and temperature during roast thus they could be used as index composition for the roasting process (Magi et al., 2012). Other volatile flavor components formed during roasting process include alcohols, esters, carboxylic acids, aldehydes, ketons, amines (Misnawi and Ariya, 2011).

Polyphenols are another family of metabolites rich in cacao seed and cacao products. They are widely studied due to their health benefits. The main polyphenols in cacao seed and its products include flavan-3-ols (epicatechin, catechin), procyanidins, *N*-phenylpropenoyl amino acids. Polyphenol compounds endow cacao products bitter and astringent taste (Stark et al., 2005; Stark and Hofmann, 2005; Stark et al., 2006).

1.2 *Nelumbo nucifera*

1.2.1 Botany, phytogeography, economic and cultural importance of lotus

Nelumbo nucifera ($2n = 16$), also known as sacred lotus, is an aquatic perennial plant widely cultivated in Asia countries. It is in the family of Nelumbonaceae and belongs to the genus of *Nelumbo* which contains only two species. The other species is *Nelumbo lutea*, also known as American lotus which is mainly distributed in the eastern areas of North America and Central America.

Sacred lotus (*N. nucifera*) is named for its religious significance in both Hindus and Buddhists. It is the national flower of India and Vietnam. In China, it is not only used for pond decoration but also consumed as food and medicine. All the parts of lotus are valuable. The rhizomes are consumed as vegetable. The leaves and flower petals can be used as tea. The lotus seeds are consumed as snacks or used in some cuisines. They are also a traditional Chinese medicine and used as sedative, antipyretic and antidiarrheal agents. Besides the economical values, lotus is also an important symbol in literature due to its cultural connotation. As 'the Gentleman in Flowers', it is admired by Chinese for its noble and unstained character as though rising from mud, lotus always keeps purity. The self-cleaning property of lotus which is known as 'lotus effect' (Barthlott and Ehler, 1977) attracted attention of scientists. Inspired by the microstructure of lotus leaf surface which is hydrophobic with nano-scale hair-like structures, scientists developed coatings, paints, fabrics with self-cleaning property.

1.2.2 Phytochemical constitutes and bioactivities of lotus

Almost all parts of lotus are used in dietary or/and medical aspects. The phytochemical profiles of lotus leaves, rhizomes, seeds, flowers and even stems (2014 Joyce) and pollens (2011 Xu) have been studied. The traditional usages of lotus have obtained more and more attention.

Patel *et al* investigated the haematopoietic activity of ethanolic extract of aerial parts of lotus on cyclophosphamide induced aplastic anaemia and haloperidol induced iron deficiency anaemia in rats (Patel *et al.*, 2012). The result indicated that the extract stimulated erythropoiesis process and such anti-anaemic activity was thought due to the richness in iron and other constituents (Patel *et al.*, 2012).

Lotus seed is a traditional Chinese medicine used as sedative and antidiarrheal agent. A study indicated that the main alkaloid of lotus seed, neferine, dose-dependently inhibited locomotor activity in mice (Sugimoto *et al.*, 2008). Neferine also induced hypothermia and showed anti-anxiety activity without affecting muscle coordination (Sugimoto *et al.*, 2008). In another study, lotus seed oil (hexane extraction) showed inhibition on several enteric bacteria strains (Anitha and Arunkumar, 2012) indicating its antidiarrheal activity. Such lotus seed oil also effectively controlled the spread of several dermatophytes (Anitha and Arunkumar, 2012) which

suggested its potential usage in preparation of ointment to against bacterial and fungal infections. The antioxidant activity of ethanol extraction of lotus seeds was extensively studied (Ling et al., 2005; Rai et al., 2006; Anitha and Arunkumar, 2012) and suggested to relate with its hepatoprotective effects (Sohn et al., 2003). The inorganic elements in lotus seed ash showed hypoglycemic properties in diabetic rats (Mani et al., 2010).

The chemical composition of lotus leaves was also widely studied. With high abundance of flavonoids, the extracts of lotus leaves showed antioxidant properties (Huang et al., 2010; Zhu et al., 2015). Besides flavonoids, lotus leaf extracts also contain terpenes (Kim et al., 2009) and alkaloids (Kashiwada et al., 2005). Some of the alkaloids strongly inhibited pancreatic lipase activity, adipocyte differentiation and absorption of lipids and carbohydrates but accelerated lipid metabolism and energy expenditure indicating their potential benefit in suppression of obesity (Ono et al., 2006; Ahn et al., 2013).

Rhizomes of lotus are normally consumed as vegetables in Asia countries. The extraction of rhizomes are rich in polyphenolic compounds endowing them antioxidant property (Shad et al., 2012). The ethanolic extract of lotus rhizome showed potent hypoglycemic effect in normal and diabetic rats by improving glucose tolerance and enhancing the action of insulin indicating its potential benefits for diabetic patients (Mukherjee et al., 1997).

The outstanding and beautiful lotus flowers are popular pond decorations. Some benzyloisoquinoline alkaloids isolated from lotus petals and leaves showed substantial inhibition of melanogenesis (Nakamura et al., 2013) indicating their potential usage in cosmetic products.

1.3 *Vitis vinifera*

1.3.1 History, distribution and cultivars of grape

Vitis vinifera is a species of grape in the family of Vitaceae. This species is mainly distributed in Mediterranean region and central Europe. It could trace back to 6000 B.C. when grape was domesticated (Myles et al., 2011). As one of the most economic fruit crops, most of the harvested grape berries are produced to wine and others are either consumed freshly as table grapes or processed into juice, vinegar and raisins. Currently, there are thousands of grape cultivars. The grape berries used in this study are from Early Campbell which is a hybrid of *Vitis vinifera* and *Vitis labrusca* by crossing Moore Early with Belvidere × Muscat of Hamburg. The fruit taste and disease resistance of Early Campbell are similar to Concord (Robinson et al., 2013) probably due to the Concord genetic background of Moore Early and Belvidere.

1.3.2 Phytochemical constitutes and bioactivities of grape

Grape berries are rich in sugars, organic acids, amino acids (Lamikanra and Kassa, 1999), terpenes (Marais, 1983; Kalua and Boss, 2009) and polyphenolic compounds. Sugars such as fructose, glucose and sucrose endow grape berries the sweet taste. The main organic acids in grape are tartaric acid, malic acid and citric acid. They are responsible for the titratable acidity and sour taste in wine. In red wine, they also play roles as stabilizing agents for anthocyanins which are essential for the wine color stabilization.

Polyphenolic compounds include benzoic acids and those derived from phenylpropanoid pathway such as cinnamic acids, stilbenoids and flavonoids. These polyphenolic metabolites are extensively studied due to their antioxidant activity and related health benefits. The cinnamic acids in grapes and its products include caffeic acid, *p*-coumaric acid, ferulic acid and their derivatives (usually are conjugated esters with tartaric acid). Stilbenoids in grape include resveratrol, resveratrol glycosides and resveratrol oligomers. The content of stilbenoids in different varieties can be vary considerably (Sun et al., 2006). Resveratrol, as the most attractive composition in grape and wine has been extensively studied. The bioactivities of stilbenoids such as antioxidant, antiproliferation, inhibition of lipid oxidation, inhibition of platelet aggregation, anti-inflammation, phytoestrogenic effect were summarized in the reviews (King et al., 2006; Yang and Xiao, 2013). The flavonoids in grape are mainly in the subgroups of flavonols, flavan-3-ols, and anthocyanins. Consumption of grape seed extract, which is rich in polyphenolic compounds, especially proanthocyanins, could improve lipid profiles of hyperlipidemic individuals (Tebib et al., 1994; Razavi et al., 2013) indicating their potential benefits in cardiovascular system protection (Woodward et al., 2004), anti-obesity (Kim et al., 2013), against diabetic nephropathy (Zhang et al., 2013). Grape seed oil was said to be good dietary oil for its positive alteration in lipid profile (Kim et al., 2010). Grape seed and peel extract inhibited lipid oxidation (Shaker, 2006) which might be correlated with their neuroprotective function in rats with brain injury that caused by hypoxia-ischemia (Feng et al., 2007).

The phytochemical constituents of grape berries are affected by genetic backgrounds, environmental conditions and developmental stages of grape berries. The most noticeable difference between white and red cultivar is their anthocyanin content (Degu et al., 2015). The anthocyanin profiles of red cultivars also distinct from each other (Liang et al., 2011; Negri et al., 2015; Silva and Queiroz, 2016). Some terpenoids and aroma differences are responsible for the unique flavor of various grape cultivars. The environmental conditions such as light, temperature, water condition and infections showed strong affection on the metabolite accumulation. Some polyphenolic compounds and terpenoids with antioxidative activities were greatly accumulated under biotic and abiotic stresses indicating their potential protective property for plants to against these stresses (Gil et al., 2012; Fernandes de Oliveira and Nieddu,

2016). The understanding of the defense strategy of the plants can also be helpful in enriching targeting metabolites.

2. System biology in plant research

2.1 Omics technologies present systemic and genome-scale overviews of plant metabolism

Humans have a long history of cultivating and selecting crops to make them meet our demands and preferences. Attempting to increase in yield and quality traits require an improving understanding of seed and fruit development. Morphological as well as biochemical changes happen during seed and fruit developmental processes at various biological levels. For a better systemic understanding of these processes, omics technologies with high throughput platforms are required for large scale data acquisition and analysis. Systems biology integrates the information from genome-scale metabolic reconstruction, transcriptomic, proteomic and metabolomic levels to present a holistic view of the metabolism or status of the study objects (Figure 1) (Weckwerth, 2003;2011a;b). Different from traditional approaches which are usually driven by hypothesis and aim at single gene, protein or metabolite, omics experiments generate hypothesis based on the acquired data and their associations (Horgan and Kenny, 2011;Weckwerth, 2011a). The application of omics strategies in plant biology expand our understanding of fruit and seed development and accelerate the discovery of biomarkers that associate with development and stresses response (Chaturvedi et al., 2013;Bellaire et al., 2014;Ischebeck et al., 2014;Chaturvedi et al., 2015;Wang et al., 2016a;Wang et al., 2016b). This was recently demonstrated in cacao seed development. The integrated metabolomic and proteomic analysis of different stages of development revealed the defense preparation of the seeds for abiotic and biotic stresses (Wang et al., 2016b) (see below sections 2.2.4; 2.2.5; 2.3.1)

The development of sequencing technologies in the past two to three decades triggered genome sequencing of not only model plants but also economically important crops. These reference genome sequences provide access to interpret metabolism regulation and plant evolution at genome-scale (Weckwerth, 2011a;b). Transcriptomics aims at analysis of the total mRNA in a cell or organism which reflects gene expression levels under a specific condition at a given moment. Proteomics and metabolomics aim at characterizing the whole protein and metabolite profiles of a biological system. Proteins and metabolites are essential components of metabolic pathways thus are vital in living organisms not only to maintain physiological status but also to react to environmental stresses. They are the downstream products of gene transcription and closer to the phenotype of research organisms. Accordingly, proteomics and

metabolomics approaches are applied in this thesis for understanding seed and fruit development (Wang et al., 2016a; Wang et al., 2016b; Wang et al., 2016c).

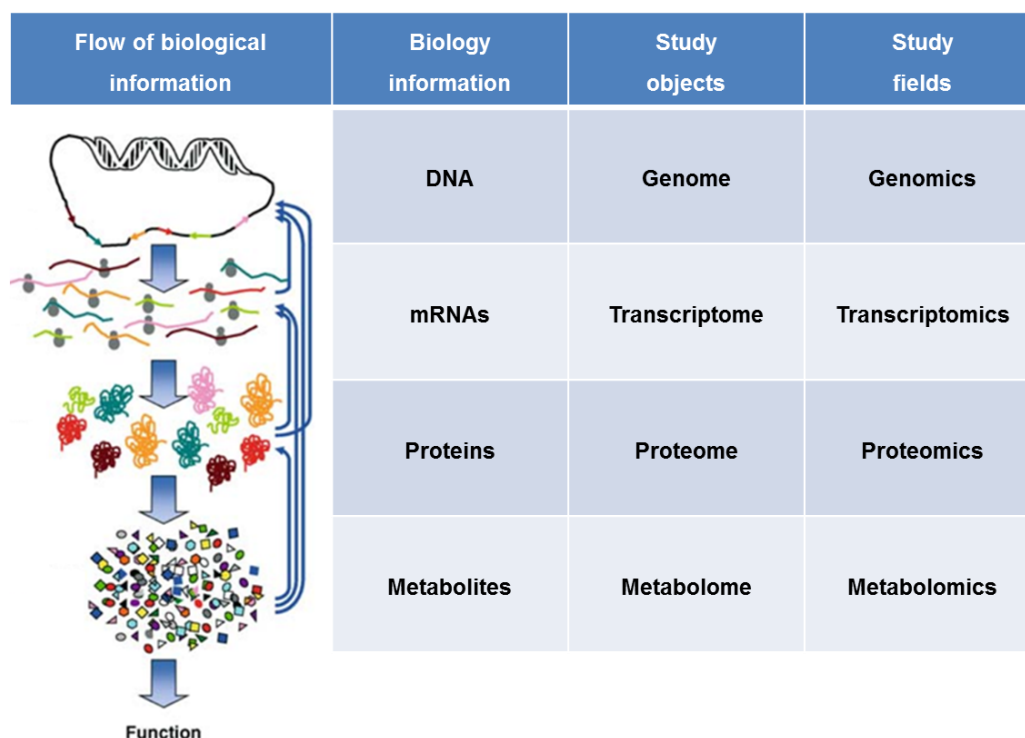


Figure 1. General flow of biological information and the schematic of the omics organizations. Adapted from (Goodacre, 2005)

2.2 Mass spectrometry based proteomics and metabolomics

2.2.1 Mass spectrometry based platform

Mass spectrometry (MS) is a commonly used analytical technique in metabolomics and proteomics. A general MS platform consists of a sample injection system, mass spectrometer and signal processing system (Figure 2).

Samples could be loaded directly or after a chromatographic pre-separation. Gas chromatography (GC), liquid chromatography (LC) and capillary electrophoresis (CE) are the common chromatographic techniques for the pre-separation which is helpful for data acquisition and following data analysis. Mass spectrometer is the core part of the platform. The injected metabolites and peptides are firstly ionized. The ionization methods include electron ionization (EI), electrospray ionization (ESI), chemical ionization, fast atom bombardment (FAB), atmospheric pressure chemical ionization (APCI), matrix-assisted laser desorption ionization (MALDI) and so on. EI is a hard ionization method using energetic electrons to attack chemical molecules and leading to not only ionization but also a high degree of fragmentation. EI is commonly applied in GC-MS system. ESI, on the contrary, is a soft ionization method

commonly used in LC-MS system and hardly cause any fragmentation. Next, the charged ions are sorted in mass analyzer according to their mass-to-charge ratio (m/z). The mass analyzers that used nowadays include magnetic sector, time of flight (TOF), quadrupole, fourier transform ion cyclotron resonance (FT-ICR), ion trap and orbitrap. They separate ions in different ways thus offer different abilities in terms of resolution, mass accuracy, mass range, dynamic range and cost (Glish and Vachet, 2003). Finally, a detector record the electric signal when the ions pass by. These signals are then transformed to m/z and corresponding intensity information which is used for mass spectrum formation.

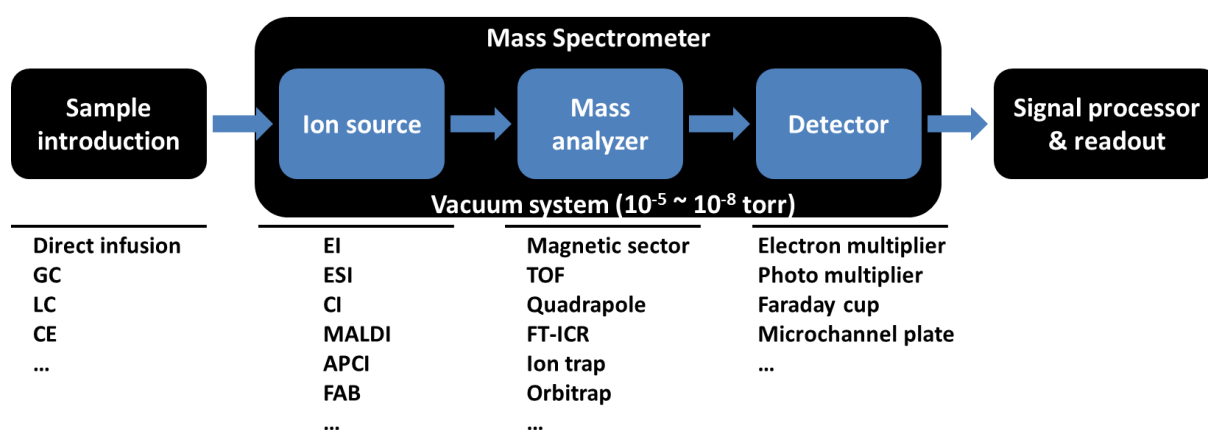


Figure 2. Mass spectrometry based platform.

Tandem mass spectrometry (MS/MS) is a crucial technology in the contribution to metabolomics and proteomic studies. It involves multiple steps of ion selection, fragmentation and detection. In the first stage of MS (MS1), the ions formed in the ion source are separated according to their m/z . Subsequently, an ion with a specific m/z is selected and fragmented by collision-induced dissociation or other processes. These fragments/product ions are then separated and recorded in MS2. Repeating this process leads to MS^n with n indicating the MS levels of the measurement which can be performed e.g. in ion trap mass spectrometry. The fragment information obtained from tandem mass spectrometry is crucial for metabolite structure and protein sequence elucidation. The instrument applied in this study is LTQ Orbitrap XL system. It hybrids a linear ion trap with an Orbitrap featuring fast scan rate, high sensitivity, high ion storage capacities as well as high mass accuracy, high resolution and wide dynamic range .

2.2.2 Mass spectrometry based proteomics

Application of MS in proteomic studies robustly improved the efficiency in studies of protein identification, quantification and modification. MS based proteomics has been widely used in clinical studies to search for disease related biomarkers and to unravel the mechanisms of

drug actions. In plant science, proteomic studies of model plants and economic crops have revealed their developmental processes and reactions to biotic and abiotic stresses. Generally, proteins are analyzed either intact (top-down strategy) or after enzymatic digestion (bottom-up strategy, shotgun proteomics). Top-down approach measures the masses of intact proteins and their fragments. This strategy promised a higher sequence coverage thus ensured a more confident identification. By measuring the protein abundance directly, it offers a more reliable quantification result. However, due to technological limitations, it is still limited in analysis of single protein or simple protein mixture. On the contrary, bottom up strategy fulfill a high through-put proteome analysis. Instead of intact protein analysis, this strategy firstly digests proteins with enzymes and then measures the obtained peptides. The most commonly used enzyme is trypsin. Recently, the properties of some C terminal proteases (chymotrypsin, LysC, ArgC, GluC) and N terminal proteases (LysN and AspN) were systematically evaluated (Giansanti et al., 2016). Tandem MS is usually used for data acquisition. Subsequently, peptide sequence interpretation and protein assembly are driven by database search. Some open source as well as commercial softwares are available for sequence interpretation and protein assemble by applying different algorithms (Gonzalez-Galarza et al., 2012). For protein quantification, some labelling strategies and label-free approaches have been developed. Label-free methods with simple procedure and lower cost are widely used in high throughput quantitative proteomic studies. In label-free approaches, proteins are normally quantified with precursor intensity, spectral counting or peak integration (Weckwerth et al., 2004;Hoehenwarter et al., 2008;Zhu et al., 2010). Spectral counting takes the total spectra number of all the peptides that belonging to a protein as indicator of relative protein abundance (Liu et al., 2004;Paoletti et al., 2006;Wienkoop et al., 2006).

2.2.3 Mass spectrometry based metabolomics

Metabolites are end products of biological flow with essential functions for organisms to survive, grow and reproduce. Metabolomics aims at systematically study the metabolite profile of an organism, an organ or a cell under a given condition. Currently, nuclear magnetic resonance (NMR) and MS based methods are popular for metabolite detection. NMR provides a non-destructive, reproducible and robust measurement of metabolites. The abundant structural information in NMR spectra promises an unambiguous identification of metabolites. However, the sensitivity of NMR is much lower than MS (Tsedilin et al., 2015). Besides high sensitivity, the conjugation of MS with other separation systems such as HPLC, GC and CE is easier and more developed than hyphenated NMR platforms. The hyphenated separation processes are helpful in complex sample simplification and provide retention time as an additional identity information. Comparing with proteins, metabolites are small molecules (< 2000 Da) but with huge diversities in structure and property thus require different separation, ionization, data

deconvolution and structure elucidation methods for sample analysis and data interpretation. GC-MS and LC-MS are the most commonly applied MS based platforms for metabolomics studies with complementary application scope (Weckwerth, 2003;2010).

GC-MS is a well established method for analysis of volatile and thermally stable compounds. For those polar and/or thermal sensitive metabolites, derivatization is required for improving volatility and stability. The ionization method in GC-MS is EI which is a hard ionization method that not only ionizes but also fragments molecules. The obtained specific and reproducible spectra are included in databases or libraries for metabolite identification (Stein, 1995;Kopka et al., 2005). In the last decades GC-MS has developed into one of the gold standards of metabolomics technology (Weckwerth, 2003;Weckwerth et al., 2004;Weckwerth, 2010).

The compounds that are not suitable for GC-MS measurement can be analyzed by LC-MS/MS. Different from GC-MS using EI for ionization, tandem LC-MS ionizes metabolites by ESI which is a soft ionization method. MS² and MSⁿ spectra containing fragmentation pattern information are helpful for structure determination. Softwares have been developed for structure interpretation basing on compound fragmentation rules (Ma et al., 2014) and biological synthetic pathway information (**Doerfler et al., 2014**). Besides annotating metabolites according to their spectra features, mzGroupAnalyzer additionally maps the annotated metabolites to their biosynthetic pathways according to structure transforming information (Doerfler et al., 2013).

2.2.4 Experimental setup for integrated metabolomic and proteomic analysis

Metabolic processes are composed of biological molecules as well as interactions between them. A systematic view of biological processes should not be restricted to gene expression, protein and metabolite production but rather connections, correlations and interactions of these biological molecules (Weckwerth, 2003). Consequently, integrated analysis of transcriptomic, proteomic, and metabolomics data is of great importance for genome-wide understanding of organism status. Combined extraction protocols for DNA, RNA, protein and metabolite extraction were developed (Weckwerth et al., 2004;Valledor et al., 2014) to overcome the biases caused by differences among sample divisions. In the present thesis, metabolites and proteins were consecutively extracted from samples basing on the aforementioned protocols with some modifications (Wang et al., 2016b). The experimental setup is showing in Figure 3.

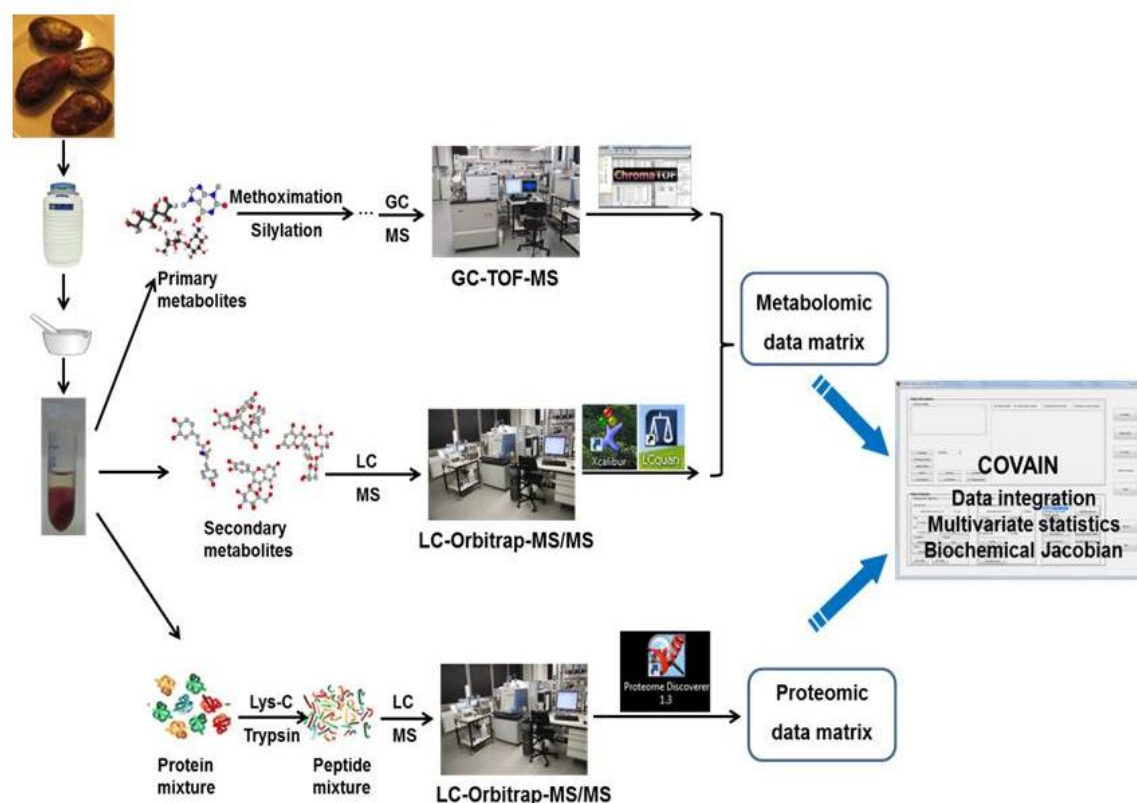


Figure 3. Experimental setup (Wang et al., 2016b)

The metabolites were measured with GC-TOF-MS after derivatization and LC-Orbitrap-MS/MS for both primary and secondary metabolite identification and quantification. Proteins were digested with Lys-C and trypsin and then measured with LC-Orbitrap-MS/MS. The obtained metabolomic and proteomic data matrices were integrated followed by multivariate statistical analysis.

2.2.5 Combinatorial omics analysis in understanding interactions and correlations of primary and secondary metabolism

High through-put omics technologies have been speeding up data acquisition with considerable coverage of metabolic actions. Primary metabolites involved in universal and central primary metabolic pathways are essential compounds involved in fundamental living processes in organisms. Secondary metabolites, on the contrary, are not critical for organisms to survive in short term but rather crucial for environmental interactions of organisms responding to biotic and abiotic stresses, e.g. the “innate immune system” of plants. Secondary metabolites are also precious resources of medicines, flavors and pigments making them of great interest. The regulation of primary and secondary metabolite production is genetically controlled and affected by environmental factors. Understanding the interactions and correlations of primary and secondary metabolite are of great importance. Primary metabolites

serve as substrates for secondary metabolite synthesis. On the one hand, sufficient supply of primary metabolites is a precondition for the production of secondary metabolites. On the other hand, secondary metabolite accumulation could send feedback signals to influence transportation, anabolism and catabolism of primary metabolites. The regulation and reactions of metabolic processes involve gene expressions, protein production and modification as well as metabolite composition changes. The interactions between primary and secondary metabolic actions could be direct or indirect, immediate or time lagged. Classic methods in understanding metabolic relations include metabolic engineering and isotope labelling experiments. Recently, the application of mathematical modeling and correlation analysis of omics dataset combining with biological pathway information provides many interesting correlation targets (Nägele and Weckwerth, 2012;Doerfler et al., 2013;Nägele, 2014). In the study of cacao seed development, comparison the Jacobian entries of developing cacao seed metabolomics data showed interactions between aromatic amino acids metabolism with flavonoids synthesis and reprogramming of metabolic flow to *N*-phenylpropenoyl amino acid and flavonoids (Wang et al., 2016b). In the study of developing grape berries, clustering analysis following with Granger causality analysis was applied for searching directed and nonlinear correlations among metabolites and proteins involved in primary and secondary metabolism pathways (see in the last Manuscript (Wang et al., 2016c)). Those candidates with high correlation scores are interesting targets for understanding metabolism interactions during seed and fruit development.

2.3 Systematic approaches in understanding seed and fruit development

2.3.1 Seed development

Seed, as a sink organ, contains genetic information and nutrition for production of next generation. Seeds form from fertilization of ovules with pollens then undergo morphological, physiological and biochemical changes to finally become mature seeds. Systematic approaches have been launched for understanding seed developmental processes of model plant *Arabidopsis thaliana* (Fait et al., 2006;Hajduch et al., 2010;Le et al., 2010) as well as essential oil crops such as soybean (Agrawal et al., 2008;Li et al., 2015), oilseed rape (Schwender et al., 2015;Xu et al., 2015), and cereals such as wheat (Guo et al., 2012;Yu et al., 2016;Zhen et al., 2016), rice (Xu et al., 2008;Xu et al., 2012;Deng et al., 2013;Hu et al., 2016), maize (Alonso et al., 2010;Lu et al., 2013;Chen et al., 2014;Li et al., 2014) and barley (Finnie et al., 2002;Sreenivasulu et al., 2008).

In general, seed developmental process contains embryogenesis, reserve accumulation, and maturation phases. Embryogenesis phase is characterized by fast cell division and histodifferentiation. Reserve accumulation phase marks the initiation of reserve production.

The main reserves in seeds are carbohydrates, proteins and lipids which are accumulated as energy storage reserves to support germination and early growth of seedlings. The composition of these reserves in seeds varies among different species. For instance, cereals mainly accumulate carbohydrates while legume seeds accumulate higher amount of proteins and oilseeds store higher content of lipids. Mature cacao beans contain high level of lipids (Martini et al., 2008) whereas sacred lotus seeds strongly accumulate starch and proteins during ripening (Wang et al., 2016a). Competitions in starch-lipid and protein-lipid accumulation were observed in seeds of oilseed rape (Schwender et al., 2015) and *Arabidopsis* (Kanai et al., 2016). The importance and mechanisms of using reserves for seed germination and young seedling growth were studied in model plant *Arabidopsis thaliana* (Gomez et al., 2009; Eastmond et al., 2015). Reserves continue to accumulate during ripening phase and finally endow the mature seed desiccation tolerance as well as germination power.

Metabolic dynamics during seed development are intricate. In this thesis, metabolism profiles of developing cacao seeds, sacred lotus seeds and grape berry development are presented with respect to metabolome and proteome dynamics. Integration of metabolome and proteome datasets of developing cacao seeds shows metabolism adjustment during seed developmental process (Wang et al., 2016b) which is presented in Figure 4.

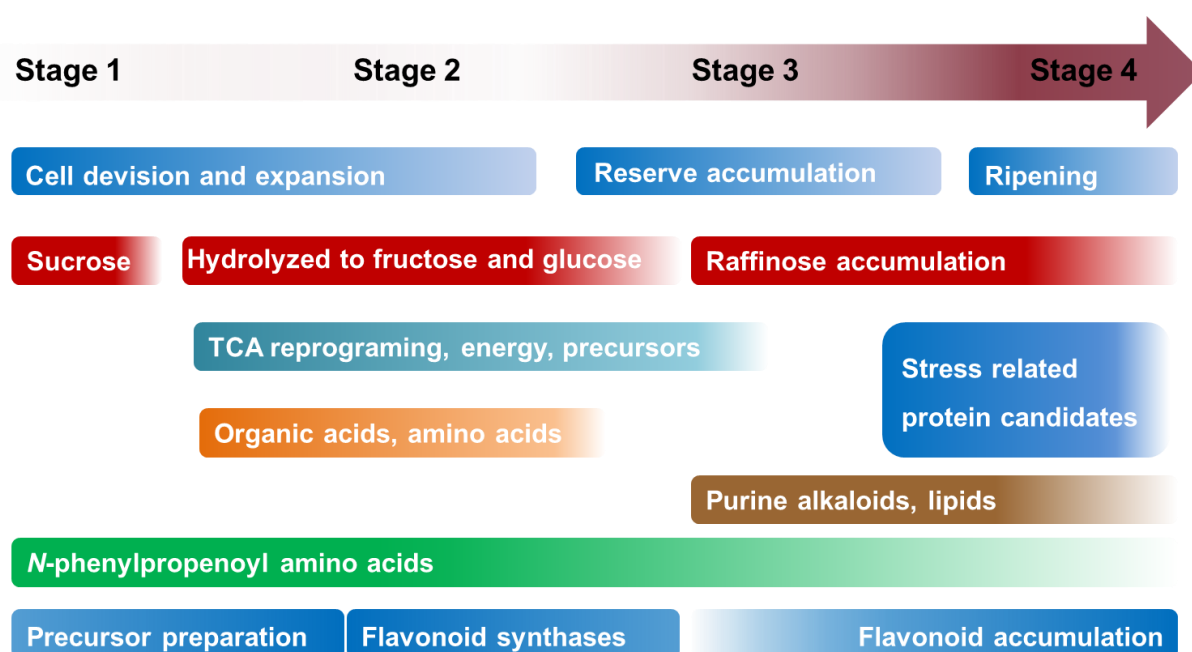


Figure 4. Metabolism characters of developing cacao seed.

Transported sucrose mainly hydrolyzed to fructose and glucose at early developmental phase. These hexoses further incorporate into active glycolysis and tricarboxylic acid (TCA) cycle for energy and carbon skeleton production which provide essential fundamentals for synthesis and transportation of organic acids and amino acids. Metabolic switches happened at the

reserve accumulation phase. Sucrose metabolism directed to raffinose synthesis instead of hexoses production. The metabolism flux along phenylpropanoid pathway altered from *N*-phenylpropenoyl amino acid synthesis to flavonoids production. Amino acids and organic acids were involved in protein and lipid accumulation. Proteins against biotic and abiotic stresses accumulated in ripening beans assisting them cooperate with environmental conditions. Finally, in the mature seeds, the accumulated raffinose family oligosaccharides, flavonoids and stress responsive proteins synergistically protect the cacao seed from environmental stresses.

2.3.2 Grape berry development

Grape berry developmental process contains three phases with two sigmoidal growth phases separated by a lag phase (**Error! Reference source not found.**). The first growth phase starts from fruit setting. Berries at this phase undergo fast cell division leading to rapid volume enlargement. Phase II is defined as a lag phase and characterized by a pause in volume expansion. The last period, phase III, starts from veraison indicating the transition of grape berry from growth to ripening. At this phase, grape berries go through a second volume expansion mainly due to the enlargement of mesocarp cells. During ripening, grape berries become soft and the peels of those red cultivars become reddish.

The three growth phases of grape berry are also characterized by their distinctive metabolism profiles which have close relation with their flavor and taste formation. Young berries at early developmental stages are rich in volatile compounds especially terpenes and some esters (Kalua and Boss, 2009). During phase I, the level of terpenes and esters are declined but with more varieties of volatile compounds are formed (Kalua and Boss, 2009). Since veraison, the variety and quantity of volatile compounds decreased but with an increase in levels of alcohols and benzene derivatives in mature berries (Kalua and Boss, 2009). Acids and sugars are responsible for the titratable acidity and sweetness of grape berries. The predominant organic acids in grape berries are malic acid and tartaric acid. Their levels steadily increase until veraison and then decline afterwards. On the contrary, the total sugar level significantly increases during ripening. Grape berries are also rich in phenolic compounds, especially flavonoids. Hydroxycinnamic acids, flavan-3-ols, flavonols and their derivatives (Ali et al., 2011) are in high abundance in young grape berries and decrease during phase I. After veraison, grape berries of red cultivars strongly accumulate anthocyanins which endow grape berries the reddish color.

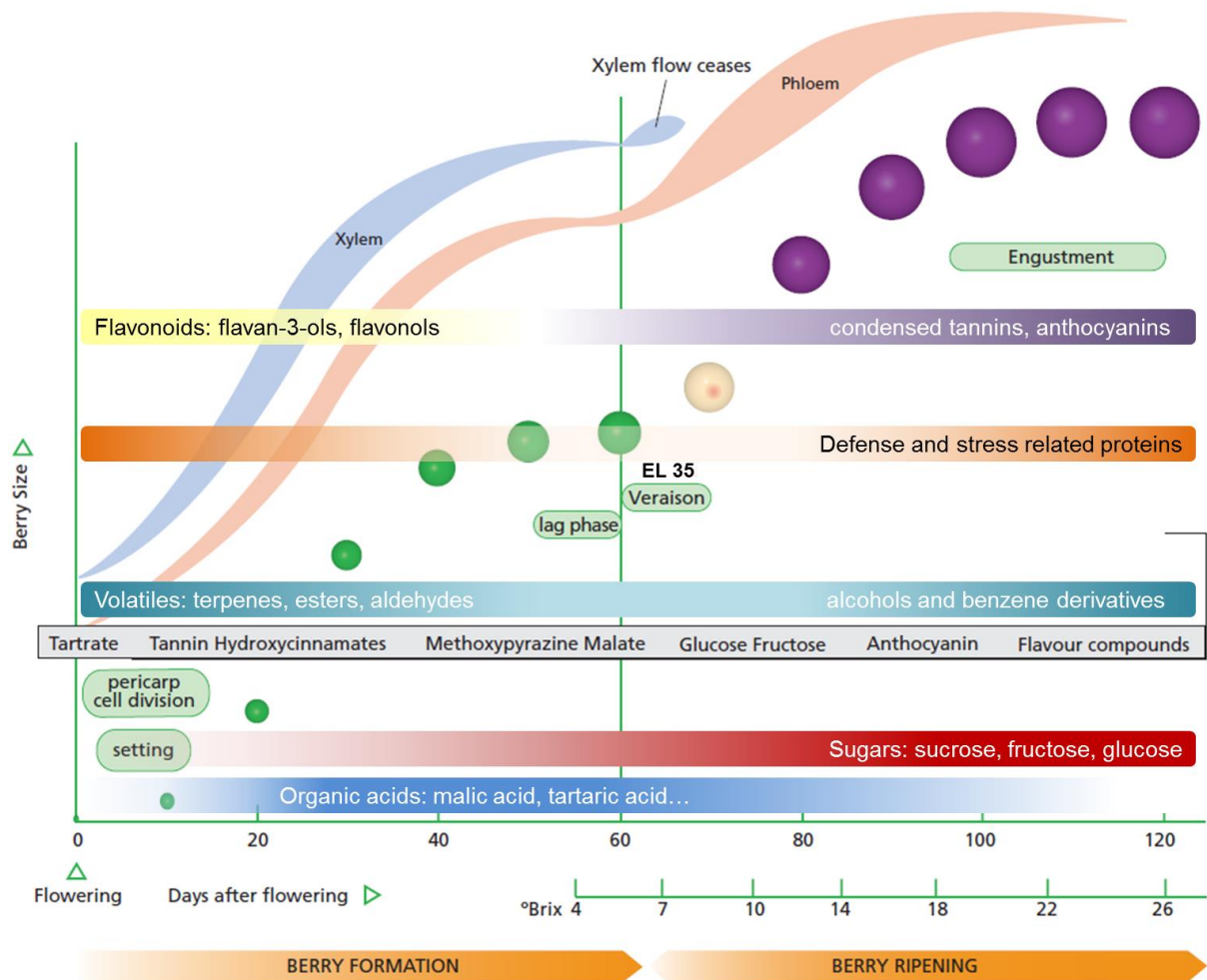


Figure 5. Developmental process of grape berry. Figure adapted from (Kennedy, 2002)

The protein content was reported to decrease during the developmental process of grape berries but with more diversity (Fraige et al., 2015). The corresponding proteins involved in metabolisms of carbohydrates, organic acids, flavonoids, as well as those related with defense and stress responses were reported in several studies (Sarry et al., 2004; Giribaldi et al., 2007; Martinez-Esteso et al., 2011; Fraige et al., 2015). In the presented manuscript of this thesis we have analysed metabolite and protein dynamics during this developmental trajectory and analysed their interactions with granger causality analysis (Wang et al., 2016c). Granger causality is a technique in multivariate statistical analysis which identifies causal correlations based on time-shifted correlation analysis, e.g. precursor-product relationships (Sun and Weckwerth, 2012; Doerfler et al., 2013). A complete granger causation network of the metabolome and proteome of different stages during grape berry development is found in supplementary figure S4. This analysis identified highly connected biochemical hubs and cluster in the interaction network. Those hubs and cluster are discussed in the manuscript and serve as potential physiological markers for various processes during development, e.g. the acidity of grapes during developmental processes. Future analysis will exploit those reference

points in multivariate statistical analysis to define predictor-response networks which are able to predict specific properties of grape berries during the ripening process.

3. References

- Adeyeye, E.I., Akinyeye, R.O., Ogunlade, I., Olaofe, O., and Boluwade, J.O. (2010). Effect of farm and industrial processing on the amino acid profile of cocoa beans. *Food Chemistry* 118, 357-363.
- Agrawal, G.K., Hajduch, M., Graham, K., and Thelen, J.J. (2008). In-depth investigation of the soybean seed-filling proteome and comparison with a parallel study of rapeseed. *Plant Physiol* 148, 504-518.
- Ahn, J.H., Kim, E.S., Lee, C., Kim, S., Cho, S.H., Hwang, B.Y., and Lee, M.K. (2013). Chemical constituents from *Nelumbo nucifera* leaves and their anti-obesity effects. *Bioorg Med Chem Lett* 23, 3604-3608.
- Ali, K., Maltese, F., Fortes, A.M., Pais, M.S., Choi, Y.H., and Verpoorte, R. (2011). Monitoring biochemical changes during grape berry development in Portuguese cultivars by NMR spectroscopy. *Food Chemistry* 124, 1760-1769.
- Alonso, A.P., Dale, V.L., and Shachar-Hill, Y. (2010). Understanding fatty acid synthesis in developing maize embryos using metabolic flux analysis. *Metab Eng* 12, 488-497.
- Anitha, A., and Arunkumar, D. (2012). Fatty acid composition and antidermatophytic and antidiarrheal activity of *Nelumbo nucifera* seed oil. *International Journal of Pharmacy and Pharmaceutical Sciences* 4, 264-270.
- Argout, X., Salse, J., Aury, J.M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J.F., Sabot, F., Kudrna, D., Ammiraju, J.S., Schuster, S.C., Carlson, J.E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Berard, A., Viot, C., Boccara, M., Risterucci, A.M., Guignon, V., Sabau, X., Axtell, M.J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tah, M., Akaza, J.M., Pitollat, B., Gramacho, K., D'hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W.R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., and Lanaud, C. (2011). The genome of *Theobroma cacao*. *Nat Genet* 43, 101-108.
- Arlorio, M., Bottini, C., Travaglia, F., Locatelli, M., Bordiga, M., Coisson, J.D., Martelli, A., and Tessitore, L. (2009). Protective activity of *Theobroma cacao* L. phenolic extract on AML12 and MLP29 liver cells by preventing apoptosis and inducing autophagy. *J Agric Food Chem* 57, 10612-10618.
- Arteel, G.E., Schroeder, P., and Sies, H. (2000). Reactions of peroxynitrite with cocoa procyanidin oligomers. *J. Nutr.* 130, 2100S-2104S.
- Barthlott, W., and Ehler, N. (1977). Raster-Elektronenmikroskopie der Epidermis-Oberflächen von Spermatophyten. *Tropische und subtropische Pflanzenvelt (Akad. Wiss. Lit. Mainz)* 19, 110.
- Bellaire, A., Ischebeck, T., Staedler, Y., Weinhaeuser, I., Mair, A., Parameswaran, S., Ito, T., Schonenberger, J., and Weckwerth, W. (2014). Metabolism and development - integration of micro computed tomography data and metabolite profiling reveals metabolic reprogramming from floral initiation to silique development. *New Phytol* 202, 322-335.
- Caligiani, A., Cirilini, M., Palla, G., Ravaglia, R., and Arlorio, M. (2007). GC-MS detection of chiral markers in cocoa beans of different quality and geographic origin. *Chirality* 19, 329-334.
- Camu, N., De Winter, T., Addo, S.K., Takrama, J.S., Bernaert, H., and De Vuyst, L. (2008). Fermentation of cocoa beans: influence of microbial activities and polyphenol concentrations on the flavour of chocolate. *J. Sci. Food Agric.* 88, 2288-2297.
- Chaturvedi, P., Doerfler, H., Jegadeesan, S., Ghatak, A., Pressman, E., Castillejo, M.A., Wienkoop, S., Egelhofer, V., Firon, N., and Weckwerth, W. (2015). Heat-treatment-responsive proteins in

- different developmental stages of tomato pollen detected by targeted mass accuracy precursor alignment (tMAPA). *J Proteome Res* 14, 4463-4471.
- Chaturvedi, P., Ischebeck, T., Egelhofer, V., Lichtscheidl, I., and Weckwerth, W. (2013). Cell-specific analysis of the tomato pollen proteome from pollen mother cell to mature pollen provides evidence for developmental priming. *J Proteome Res* 12, 4892-4903.
- Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., and Lai, J. (2014). Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol* 166, 252-264.
- Degu, A., Morcia, C., Tumino, G., Hochberg, U., Toubiana, D., Mattivi, F., Schneider, A., Bosca, P., Cattivelli, L., Terzi, V., and Fait, A. (2015). Metabolite profiling elucidates communalities and differences in the polyphenol biosynthetic pathways of red and white Muscat genotypes. *Plant Physiol Biochem* 86, 24-33.
- Deng, Z.Y., Gong, C.Y., and Wang, T. (2013). Use of proteomics to understand seed development in rice. *Proteomics* 13, 1784-1800.
- Doerfler, H., Lyon, D., Nagele, T., Sun, X., Fagner, L., Hadacek, F., Egelhofer, V., and Weckwerth, W. (2013). Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics* 9, 564-574.
- Doerfler, H., Sun, X., Wang, L., Engelmeier, D., Lyon, D., and Weckwerth, W. (2014). mzGroupAnalyzer - predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. *PLoS One* 9, e96188.
- Dorenkott, M.R., Griffin, L.E., Goodrich, K.M., Thompson-Witrick, K.A., Fundaro, G., Ye, L., Stevens, J.R., Ali, M., O'keefe, S.F., Hulver, M.W., and Neilson, A.P. (2014). Oligomeric cocoa procyanidins possess enhanced bioactivity compared to monomeric and polymeric cocoa procyanidins for preventing the development of obesity, insulin resistance, and impaired glucose tolerance during high-fat feeding. *J Agric Food Chem* 62, 2216-2227.
- Eastmond, P.J., Astley, H.M., Parsley, K., Aubry, S., Williams, B.P., Menard, G.N., Craddock, C.P., Nunes-Nesi, A., Fernie, A.R., and Hibberd, J.M. (2015). Arabidopsis uses two gluconeogenic gateways for organic acids to fuel seedling establishment. *Nat Commun* 6, 6659.
- Elwers, S., Zambrano, A., Rohsius, C., and Lieberei, R. (2009). Differences between the content of phenolic compounds in Criollo, Forastero and Trinitario cocoa seed (*Theobroma cacao* L.). *European Food Research and Technology* 229, 937-948.
- Fait, A., Angelovici, R., Less, H., Ohad, I., Urbanczyk-Wochniak, E., Fernie, A.R., and Galili, G. (2006). Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. *Plant Physiol* 142, 839-854.
- Feng, Y., Liu, Y.M., Leblanc, M.H., Bhatt, A.J., and Rhodes, P.G. (2007). Grape seed extract given three hours after injury suppresses lipid peroxidation and reduces hypoxic-ischemic brain injury in neonatal rats. *Pediatr Res* 61, 295-300.
- Fernandes De Oliveira, A., and Nieddu, G. (2016). Accumulation and partitioning of anthocyanins in two red grape cultivars under natural and reduced UV solar radiation. *Australian Journal of Grape and Wine Research* 22, 96-104.
- Finnie, C., Melchior, S., Roepstorff, P., and Svensson, B. (2002). Proteome analysis of grain filling and seed maturation in barley. *Plant Physiol* 129, 1308-1319.
- Fraige, K., Gonzalez-Fernandez, R., Carrilho, E., and Jorin-Novo, J.V. (2015). Metabolite and proteome changes during the ripening of Syrah and Cabernet Sauvignon grape varieties cultured in a nontraditional wine region in Brazil. *J Proteomics* 113, 206-225.
- Frauendorfer, F., and Schieberle, P. (2008). Changes in key aroma compounds of Criollo cocoa beans during roasting. *J. Agric. Food Chem.* 56, 10244-10251.
- Giansanti, P., Tsiatsiani, L., Low, T.Y., and Heck, A.J. (2016). Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc* 11, 993-1006.
- Gil, M., Pontin, M., Berli, F., Bottini, R., and Piccoli, P. (2012). Metabolism of terpenes in the response of grape (*Vitis vinifera* L.) leaf tissues to UV-B radiation. *Phytochemistry* 77, 89-98.
- Giribaldi, M., Perugini, I., Sauvage, F.X., and Schubert, A. (2007). Analysis of protein changes during grape berry ripening by 2-DE and MALDI-TOF. *Proteomics* 7, 3154-3170.

- Glish, G.L., and Vachet, R.W. (2003). The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov* 2, 140-150.
- Gomez, L.D., Steele-King, C.G., Jones, L., Foster, J.M., Vuttipongchaikij, S., and McQueen-Mason, S.J. (2009). Arabinan metabolism during seed development and germination in Arabidopsis. *Mol Plant* 2, 966-976.
- Gonzalez-Galarza, F.F., Lawless, C., Hubbard, S.J., Fan, J., Bessant, C., Hermjakob, H., and Jones, A.R. (2012). A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. *OMICS* 16, 431-442.
- Goodacre, R. (2005). Metabolomics – the way forward. *Metabolomics* 1, 1-2.
- Grassi, D., Necozione, S., Lippi, C., Croce, G., Valeri, L., Pasqualetti, P., Desideri, G., Blumberg, J.B., and Ferri, C. (2005). Cocoa reduces blood pressure and insulin resistance and improves endothelium-dependent vasodilation in hypertensives. *Hypertension* 46, 398-405.
- Guo, G., Lv, D., Yan, X., Subburaj, S., Ge, P., Li, X., Hu, Y., and Yan, Y. (2012). Proteome characterization of developing grains in bread wheat cultivars (*Triticum aestivum* L.). *BMC Plant Biology* 12, 147.
- Hajdуч, M., Hearne, L.B., Miernyk, J.A., Casteel, J.E., Joshi, T., Agrawal, G.K., Song, Z., Zhou, M., Xu, D., and Thelen, J.J. (2010). Systems analysis of seed filling in Arabidopsis: using general linear modeling to assess concordance of transcript and protein expression. *Plant Physiol* 152, 2078-2087.
- Henderson, J.S., Joyce, R.A., Hall, G.R., Hurst, W.J., and McGovern, P.E. (2007). Chemical and archaeological evidence for the earliest cacao beverages. *Proc Natl Acad Sci U S A* 104, 18937-18940.
- Ho, V.T., Zhao, J., and Fleet, G. (2014). Yeasts are essential for cocoa bean fermentation. *Int J Food Microbiol* 174, 72-87.
- Hoehenwarter, W., Van Dongen, J.T., Wienkoop, S., Steinfath, M., Hummel, J., Erban, A., Sulpice, R., Regierer, B., Kopka, J., Geigenberger, P., and Weckwerth, W. (2008). A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* 8, 4214-4225.
- Horgan, R.P., and Kenny, L.C. (2011). 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist* 13, 189-195.
- Hu, C., Tohge, T., Chan, S.A., Song, Y., Rao, J., Cui, B., Lin, H., Wang, L., Fernie, A.R., Zhang, D., and Shi, J. (2016). Identification of conserved and diverse metabolic shifts during rice grain development. *Sci Rep* 6, 20942.
- Huang, B., Ban, X., He, J., Tong, J., Tian, J., and Wang, Y. (2010). Comparative analysis of essential oil components and antioxidant activity of extracts of *Nelumbo nucifera* from various areas of China. *J Agric Food Chem* 58, 441-448.
- Huang, Y., and Barringer, S.A. (2011). Monitoring of cocoa volatiles produced during roasting by selected ion flow tube-mass spectrometry (SIFT-MS). *J Food Sci* 76, C279-286.
- Ischebeck, T., Valledor, L., Lyon, D., Gingl, S., Nagler, M., Meijon, M., Egelhofer, V., and Weckwerth, W. (2014). Comprehensive cell-specific protein analysis in early and late pollen development from diploid microsporocytes to pollen tube growth. *Mol Cell Proteomics* 13, 295-310.
- Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F., Wincker, P., and French-Italian Public Consortium for Grapevine Genome, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463-467.
- Kalua, C.M., and Boss, P.K. (2009). Evolution of volatile compounds during the development of cabernet sauvignon grapes (*Vitis vinifera* L.). *J Agric Food Chem* 57, 3818-3830.

- Kanai, M., Mano, S., Kondo, M., Hayashi, M., and Nishimura, M. (2016). Extension of oil biosynthesis during the mid-phase of seed development enhances oil content in Arabidopsis seeds. *Plant Biotechnol J* 14, 1241-1250.
- Kashiwada, Y., Aoshima, A., Ikeshiro, Y., Chen, Y.P., Furukawa, H., Itoigawa, M., Fujioka, T., Mihashi, K., Cosentino, L.M., Morris-Natschke, S.L., and Lee, K.H. (2005). Anti-HIV benzyloquinoline alkaloids and flavonoids from the leaves of *Nelumbo nucifera*, and structure-activity correlations with related alkaloids. *Bioorg Med Chem* 13, 443-448.
- Kennedy, J. (2002). Understanding grape berry development. *Practical Winery and Vineyard Journal* 7, 1-5.
- Kim, D.-J., Jeon, G., Sung, J., Oh, S.-K., Hong, H.-C., and Lee, J. (2010). Effect of grape seed oil supplementation on plasma lipid profiles in rats. *Food Science and Biotechnology* 19, 249-252.
- Kim, K.H., Chang, S.W., Ryu, S.Y., Choi, S.U., and Lee, K.R. (2009). Phytochemical constituents of *Nelumbo nucifera*. *Natural Product Sciences* 15, 90-95.
- Kim, Y., Choi, Y., Lee, J., and Park, Y. (2013). Downregulated lipid metabolism in differentiated murine adipocytes by procyanidins from defatted grape seed meal. *Biosci Biotechnol Biochem* 77, 1420-1423.
- King, R.E., Bomser, J.A., and Min, D.B. (2006). Bioactivity of resveratrol. *Comprehensive Reviews in Food Science and Food Safety* 5, 65-70.
- Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R., and Steinhauser, D. (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21, 1635-1638.
- Lamikanra, O., and Kassa, A.K. (1999). Changes in the free amino acid composition with maturity of the noble cultivar of *Vitis rotundifolia* Michx. grape. *J Agric Food Chem* 47, 4837-4841.
- Le, B.H., Cheng, C., Bui, A.Q., Wagmaister, J.A., Henry, K.F., Pelletier, J., Kwong, L., Belmonte, M., Kirkbride, R., Horvath, S., Drews, G.N., Fischer, R.L., Okamuro, J.K., Harada, J.J., and Goldberg, R.B. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci U S A* 107, 8063-8070.
- Lecumberri, E., Goya, L., Mateos, R., Alia, M., Ramos, S., Izquierdo-Pulido, M., and Bravo, L. (2007). A diet rich in dietary fiber from cocoa improves lipid profile and reduces malondialdehyde in hypercholesterolemic rats. *Nutrition* 23, 332-341.
- Lee, D.E., Kang, N.J., Lee, K.M., Lee, B.K., Kim, J.H., Lee, K.W., and Lee, H.J. (2010). Cocoa polyphenols attenuate hydrogen peroxide-induced inhibition of gap-junction intercellular communication by blocking phosphorylation of connexin 43 via the MEK/ERK signaling pathway. *J Nutr Biochem* 21, 680-686.
- Li, G., Wang, D., Yang, R., Logan, K., Chen, H., Zhang, S., Skaggs, M.I., Lloyd, A., Burnett, W.J., Laurie, J.D., Hunter, B.G., Dannenhoffer, J.M., Larkins, B.A., Drews, G.N., Wang, X., and Yadegari, R. (2014). Temporal patterns of gene expression in developing maize endosperm identified through transcriptome sequencing. *Proc Natl Acad Sci U S A* 111, 7582-7587.
- Li, L., Hur, M., Lee, J.Y., Zhou, W.X., Song, Z.H., Ransom, N., Demirkale, C.Y., Nettleton, D., Westgate, M., Arendsee, Z., Lyer, V., J., S., Nikolau, B., and Wurtele, E.S. (2015). A systems biology approach toward understanding seed composition in soybean. *BMC Genomics* 16, S9.
- Liang, Z., Sang, M., Fan, P., Wu, B., Wang, L., Duan, W., and Li, S. (2011). Changes of polyphenols, sugars, and organic acid in 5 *Vitis* genotypes during berry ripening. *J Food Sci* 76, C1231-1238.
- Ling, Z.Q., Xie, B.J., and Yang, E.L. (2005). Isolation, characterization, and determination of antioxidative activity of oligomeric procyanidins from seedpod of *Nelumbo nucifera* Gaertn. *J Agric Food Chem* 53, 2441-2445.
- Liu, H., Sadygov, R.G., and Yates Iii, J.R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem* 76, 4193-4201.
- Lu, X., Chen, D., Shu, D., Zhang, Z., Wang, W., Klukas, C., Chen, L.L., Fan, Y., Chen, M., and Zhang, C. (2013). The differential transcription network between embryo and endosperm in the early developing maize seed. *Plant Physiol* 162, 440-455.
- Ma, Y., Kind, T., Yang, D., Leon, C., and Fiehn, O. (2014). MS2Analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Anal Chem* 86, 10724-10731.

- Magi, E., Bono, L., and Di Carro, M. (2012). Characterization of cocoa liquors by GC-MS and LC-MS/MS: focus on alkylpyrazines and flavanols. *J Mass Spectrom* 47, 1191-1197.
- Mani, S.S., Subramanian, I.P., Pillai, S.S., and Muthusamy, K. (2010). Evaluation of hypoglycemic activity of inorganic constituents in *Nelumbo nucifera* seeds on streptozotocin-induced diabetes in rats. *Biol Trace Elem Res* 138, 226-237.
- Marais, J. (1983). Terpenes in the aroma of grapes and wines: a review. *S. Afr. J. Enol. Vitic.* 4, 49-58.
- Martinez-Esteso, M.J., Selles-Marchart, S., Lijavetzky, D., Pedreno, M.A., and Bru-Martinez, R. (2011). A DIGE-based quantitative proteomic analysis of grape berry flesh development and ripening reveals key events in sugar and organic acid metabolism. *J Exp Bot* 62, 2521-2569.
- Martini, M.H., Lenci, C.G., Figueira, A., and Tavares, D.D.Q. (2008). Localization of the cotyledon reserves of *Theobroma grandiflorum* (Willd. ex Spreng.) K. Schum., *T. subincanum* Mart., *T. bicolor* Bonpl. and their analogies with *T. cacao* L. *Brazilian Journal of Botany* 31, 147-154.
- Miller, K.B., Stuart, D.A., Smith, N.L., Lee, C.Y., Mchale, N.L., Flanagan, J.A., Qu, B.X., and Hurst, W.J. (2006). Antioxidant activity and polyphenol and procyanidin contents of selected commercially available cocoa-containing and chocolate products in the United States. *J. Agric. Food Chem.* 54, 4062-4068.
- Ming, R., Vanburen, R., Liu, Y., Yang, M., Han, Y., Li, L.T., Zhang, Q., Kim, M.J., Schatz, M.C., Campbell, M., Li, J., Bowers, J.E., Tang, H., Lyons, E., Ferguson, A.A., Narzisi, G., Nelson, D.R., Blaby-Haas, C.E., Gschwend, A.R., Jiao, Y., Der, J.P., Zeng, F., Han, J., Min, X.J., Hudson, K.A., Singh, R., Grennan, A.K., Karpowicz, S.J., Watling, J.R., Ito, K., Robinson, S.A., Hudson, M.E., Yu, Q., Mockler, T.C., Carroll, A., Zheng, Y., Sunkar, R., Jia, R., Chen, N., Arro, J., Wai, C.M., Wafula, E., Spence, A., Han, Y., Xu, L., Zhang, J., Peery, R., Haus, M.J., Xiong, W., Walsh, J.A., Wu, J., Wang, M.L., Zhu, Y.J., Paull, R.E., Britt, A.B., Du, C., Downie, S.R., Schuler, M.A., Michael, T.P., Long, S.P., Ort, D.R., Schopf, J.W., Gang, D.R., Jiang, N., Yandell, M., Depamphilis, C.W., Merchant, S.S., Paterson, A.H., Buchanan, B.B., Li, S., and Shen-Miller, J. (2013). Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol* 14, R41.
- Misnawi, and Ariya, B.T.S. (2011). Use of gas chromatography-olfactometry in combination with solid phase micro extraction for cocoa liquor aroma analysis. *International Food Research Journal* 18, 829-835.
- Mukai, R., Fukuda, I., Nishiumi, S., Natsume, M., Osakabe, N., Yoshida, K., and Ashida, H. (2008). Cacao polyphenol extract suppresses transformation of an aryl hydrocarbon receptor in C57BL/6 mice. *J. Agric. Food Chem.* 56, 10399-10405.
- Mukherjee, P.K., Saha, K., Pal, M., and Saha, B.P. (1997). Effect of *Nelumbo nucifera* rhizome extract on blood sugar level in rats. *Journal of Ethnopharmacology* 58, 207-213.
- Murphy, K.J., Chronopoulos, A.K., Singh, I., Francis, M.A., Moriarty, H., Pike, M.J., Turner, A.H., Mann, N.J., and Sinclair, A.J. (2003). Dietary flavanols and procyanidin oligomers from cocoa (*Theobroma cacao*) inhibit platelet function. *Am. J. Clin. Nutr.* 77, 1466-1473.
- Myles, S., Boyko, A.R., Owens, C.L., Brown, P.J., Grassi, F., Aradhya, M.K., Prins, B., Reynolds, A., Chia, J.M., Ware, D., Bustamante, C.D., and Buckler, E.S. (2011). Genetic structure and domestication history of the grape. *P. Natl. Acad. Sci. USA* 108, 3530-3535.
- Nägele, T. (2014). Linking metabolomics data to underlying metabolic regulation. *Front Mol Biosci* 1, 22.
- Nägele, T., and Weckwerth, W. (2012). Mathematical modeling of plant metabolism-from reconstruction to prediction. *Metabolites* 2, 553-566.
- Nakamura, S., Nakashima, S., Tanabe, G., Oda, Y., Yokota, N., Fujimoto, K., Matsumoto, T., Sakuma, R., Ohta, T., Ogawa, K., Nishida, S., Miki, H., Matsuda, H., Muraoka, O., and Yoshikawa, M. (2013). Alkaloid constituents from flower buds and leaves of sacred lotus (*Nelumbo nucifera*, Nymphaeaceae) with melanogenesis inhibitory activity in B16 melanoma cells. *Bioorg Med Chem* 21, 779-787.
- Negri, A.S., Prinsi, B., Failla, O., Scienza, A., and Espen, L. (2015). Proteomic and metabolic traits of grape exocarp to explain different anthocyanin concentrations of the cultivars. *Front Plant Sci* 6, 603.

- Ono, Y., Hattori, E., Fukaya, Y., Imai, S., and Ohizumi, Y. (2006). Anti-obesity effect of *Nelumbo nucifera* leaves extract in mice and rats. *J Ethnopharmacol* 106, 238-244.
- Oracz, J., Nebesny, E., and Żyżelewicz, D. (2014). Effect of roasting conditions on the fat, tocopherol, and phytosterol content and antioxidant capacity of the lipid fraction from cocoa beans of different *Theobroma cacao* L. cultivars. *European Journal of Lipid Science and Technology* 116, 1002-1014.
- Paoletti, A.C., Parmely, T.J., Tomomori-Sato, C., Sato, S., Zhu, D., Conaway, R.C., Conaway, J.W., Florens, L., and Washburn, M.P. (2006). Quantitative proteomic analysis of distinct mammalian mediator complexes using normalized spectral abundance factors. *Proceedings of the National Academy of Sciences* 103, 18928-18933.
- Park, J.B., and Schoene, N. (2006). Clovamide-type phenylpropenoic acid amides, N-coumaroyldopamine and N-caffeoyldopamine, inhibit platelet-leukocyte interactions via suppressing P-selectin expression. *J Pharmacol Exp Ther* 317, 813-819.
- Patel, K.K., Toppo, F.A., Singour, P.K., Chaurasiya, P.K., Rajak, H., and Pawar, R.S. (2012). Phytochemical and pharmacological investigations on aerial parts of *Nelumbo nucifera* Gaertn. for haematopoietic activity. *Indian Journal of Natural Products and Resources* 3, 512-517.
- Pereira-Caro, G., Borges, G., Nagai, C., Jackson, M.C., Yokota, T., Crozier, A., and Ashihara, H. (2013). Profiles of phenolic compounds and purine alkaloids during the development of seeds of *Theobroma cacao* cv. Trinitario. *J Agric Food Chem* 61, 427-434.
- Pohlan, H.A.J., and Pérez, V.D. (2010). Growth and production of cacao. *Soils, plant Growth and crop production* 3, 1-10.
- Powis, T.G., Cyphers, A., Gaikwad, N.W., Grivetti, L., and Cheong, K. (2011). Cacao use and the San Lorenzo Olmec. *Proc Natl Acad Sci U S A* 108, 8595-8600.
- Rai, S., Wahile, A., Mukherjee, K., Saha, B.P., and Mukherjee, P.K. (2006). Antioxidant activity of *Nelumbo nucifera* (sacred lotus) seeds. *J Ethnopharmacol* 104, 322-327.
- Ramiro-Puig, E., Casadesus, G., Lee, H.G., Zhu, X., Mcshea, A., Perry, G., Perez-Cano, F.J., Smith, M.A., and Castell, M. (2009). Neuroprotective effect of cocoa flavonoids on in vitro oxidative stress. *Eur J Nutr* 48, 54-61.
- Razavi, S.M., Gholamin, S., Eskandari, A., Mohsenian, N., Ghorbanihaghjo, A., Delazar, A., Rashtchizadeh, N., Keshtkar-Jahromi, M., and Argani, H. (2013). Red grape seed extract improves lipid profiles and decreases oxidized low-density lipoprotein in patients with mild hyperlipidemia. *J Med Food* 16, 255-258.
- Rodriguez-Campos, J., Escalona-Buendía, H.B., Orozco-Avila, I., Lugo-Cervantes, E., and Jaramillo-Flores, M.E. (2011). Dynamics of volatile and non-volatile compounds in cocoa (*Theobroma cacao* L.) during fermentation and drying processes using principal components analysis. *Food Research International* 44, 250-258.
- Rodriguez-Ramiro, I., Ramos, S., Bravo, L., Goya, L., and Martin, M.A. (2011). Procyanidin B2 and a cocoa polyphenolic extract inhibit acrylamide-induced apoptosis in human Caco-2 cells by preventing oxidative stress and activation of JNK pathway. *J Nutr Biochem* 22, 1186-1194.
- Sarry, J.E., Sommerer, N., Sauvage, F.X., Bergoin, A., Rossignol, M., Albagnac, G., and Romieu, C. (2004). Grape berry biochemistry revisited upon proteomic analysis of the mesocarp. *Proteomics* 4, 201-215.
- Schwan, R.F., and Wheals, A.E. (2004). The microbiology of cocoa fermentation and its role in chocolate quality. *Crit Rev Food Sci Nutr* 44, 205-221.
- Schwender, J., Hebbelmann, I., Heinzl, N., Hildebrandt, T., Rogers, A., Naik, D., Klapperstuck, M., Braun, H.P., Schreiber, F., Denolf, P., Borisjuk, L., and Rolletschek, H. (2015). Quantitative multilevel analysis of central metabolism in developing oilseeds of oilseed rape during *in vitro* culture. *Plant Physiol* 168, 828-848.
- Shad, M.A., Nawaz, H., Yaqoob, M., and Yousuf, B. (2012). Phytochemical composition and antioxidant properties of rhizomes of *Nilumbo nucifera*. *Journal of Medicinal Plants Research* 6, 972-980.
- Shaker, E.S. (2006). Antioxidative effect of extracts from red grape seed and peel on lipid oxidation in oils of sunflower. *LWT - Food Science and Technology* 39, 883-892.

- Silva, L.R., and Queiroz, M. (2016). Bioactive compounds of red grapes from Dão region (Portugal): Evaluation of phenolic and organic profile. *Asian Pacific Journal of Tropical Biomedicine* 6, 315-321.
- Sohn, D.H., Kim, Y.C., Oh, S.H., Park, E.J., Li, X., and Lee, B.H. (2003). Hepatoprotective and free radical scavenging effects of *Nelumbo nucifera*. *Phytomedicine* 10, 165-169.
- Sotelo, A., and Alvaez, R.G. (1991). Chemical composition of wild *Theobroma* species and their comparison to the cacao bean. *J. Agric. Food Chem.* 39, 1940-1943.
- Sreenivasulu, N., Usadel, B., Winter, A., Radchuk, V., Scholz, U., Stein, N., Weschke, W., Strickert, M., Close, T.J., Stitt, M., Graner, A., and Wobus, U. (2008). Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. *Plant Physiol* 146, 1738-1758.
- Stark, T., Bareuther, S., and Hofmann, T. (2005). Sensory-guided decomposition of roasted Cocoa nibs (*Theobroma cacao*) and structure determination of taste-active polyphenols. *J Agric Food Chem* 2005, 5407-5418.
- Stark, T., Bareuther, S., and Hofmann, T. (2006). Molecular definition of the taste of roasted cocoa nibs (*Theobroma cacao*) by means of quantitative studies and sensory experiments. *J Agric Food Chem* 54, 5530-5539.
- Stark, T., and Hofmann, T. (2005). Isolation, structure determination, synthesis, and sensory activity of *N*-phenylpropenoyl-L-amino acids from Cocoa (*Theobroma cacao*). *J. Agric. Food Chem.* 53, 5419-5428.
- Stein, S.E. (1995). Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom* 6, 644-655.
- Sugimoto, Y., Furutani, S., Itoh, A., Tanahashi, T., Nakajima, H., Oshiro, H., Sun, S., and Yamada, J. (2008). Effects of extracts and neferine from the embryo of *Nelumbo nucifera* seeds on the central nervous system. *Phytomedicine* 15, 1117-1124.
- Sun, B., Ribes, A.M., Leandro, M.C., Belchior, A.P., and Spranger, M.I. (2006). Stilbenes: Quantitative extraction from grape skins, contribution of grape solids to wine and variation during wine maturation. *Analytica Chimica Acta* 563, 382-390.
- Sun, X.L., and Weckwerth, W. (2012). COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* 8, S81-S93.
- Tebib, K., Besançon, P., and Rouanet, J. (1994). Dietary grape seed tannins affect lipoproteins, lipoprotein lipases and tissue lipids in rats fed hypercholesterolemic diets. *J Nutr.* 124, 2451-2457.
- Thomas, E., Van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., and Van Etten, J. (2012). Present spatial diversity patterns of *Theobroma cacao* L. in the neotropics reflect genetic differentiation in pleistocene refugia followed by human-influenced dispersal. *PLoS One* 7, e47676.
- Trognitz, B., Cros, E., Assemat, S., Davrieux, F., Forestier-Chiron, N., Ayestas, E., Kuant, A., Scheldeman, X., and Hermann, M. (2013). Diversity of cacao trees in Waslala, Nicaragua: associations between genotype spectra, product quality and yield potential. *PLoS One* 8, e54079.
- Tsedilin, A.M., Fakhrutdinov, A.N., Eremin, D.B., Zalesskiy, S.S., Chizhov, A.O., Kolotyrkina, N.G., and Ananikov, V.P. (2015). How sensitive and accurate are routine NMR and MS measurements? *Mendeleev Communications* 25, 454-456.
- Valledor, L., Escandon, M., Meijon, M., Nukarinen, E., Canal, M.J., and Weckwerth, W. (2014). A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. *Plant J* 79, 173-180.
- Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J.T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J.A., Sterck, L., Vandepoele, K., Grando, S.M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S.K., Fontana, P., Gutin,

- A., Van De Peer, Y., Salamini, F., and Viola, R. (2007). A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. *PLoS ONE* 2, e1326.
- Wang, L., Fu, J., Li, M., Fragner, L., Weckwerth, W., and Yang, P. (2016a). Metabolomic and proteomic profiles reveal the dynamics of primary metabolism during seed development of lotus (*Nelumbo nucifera*). *Front Plant Sci* 7, 750.
- Wang, L., Nägele, T., Doerfler, H., Fragner, L., Chaturvedi, P., Nukarinen, E., Bellaire, A., Huber, W., Weismann, J., Engelmeier, D., Ramsak, Z., Gruden, K., and Weckwerth, W. (2016b). System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance. *The Plant Journal*.
- Wang, L., Sun, X., Weismann, J., and Weckwerth, W. (2016c). Granger causality analysis of metabolite and protein dynamics during Early Campbell grape berry development. *Frontiers submitted*.
- Weckwerth, W. (2003). Metabolomics in systems biology. *Annu Rev Plant Biol* 54, 669-689.
- Weckwerth, W. (2010). Metabolomics: an integral technique in systems biology. *Bioanalysis* 2, 829-836.
- Weckwerth, W. (2011a). Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J Proteomics* 75, 284-305.
- Weckwerth, W. (2011b). Unpredictability of metabolism--the key role of metabolomics science in combination with next-generation genome sequencing. *Anal Bioanal Chem* 400, 1967-1978.
- Weckwerth, W., Wenzel, K., and Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4, 78-83.
- Wienkoop, S., Larrainzar, E., Niemann, M., Gonzalez, E.M., Lehmann, U., and Weckwerth, W. (2006). Stable isotope-free quantitative shotgun proteomics combined with sample pattern recognition for rapid diagnostics. *J Sep Sci* 29, 2793-2801.
- Woodward, D.L., Shanmuganayagam, D., and Folts, J.D. (2004). Grape extracts and risk factors for cardiovascular disease. *Business briefing: US cardiology*, 5.
- Xu, H., Gao, Y., and Wang, J. (2012). Transcriptomic analysis of rice (*Oryza sativa*) developing embryos using the RNA-Seq technique. *PLoS One* 7, e30646.
- Xu, H.M., Kong, X.D., Chen, F., Huang, J.X., Lou, X.Y., and Zhao, J.Y. (2015). Transcriptome analysis of Brassica napus pod using RNA-Seq and identification of lipid-related candidate genes. *BMC Genomics* 16, 858.
- Xu, S.B., Li, T., Deng, Z.Y., Chong, K., Xue, Y., and Wang, T. (2008). Dynamic proteomic analysis reveals a switch between central carbon metabolism and alcoholic fermentation in rice filling grains. *Plant Physiol* 148, 908-925.
- Yamagishi, M., Natsume, M., Nagaki, A., Adachi, T., Osakabe, N., Takizawa, T., Kumon, H., and Osawa, T. (2000). Antimutagenic activity of cacao: inhibitory effect of cacao liquor polyphenols on the mutagenic action of heterocyclic amines. *J. Agric. Food Chem.* 48, 5074-5078.
- Yamagishi, M., Natsume, M., Osakabe, N., Okazaki, K., Furukawa, F., Imazawa, T., Nishikawa, A., and Hirose, M. (2003). Chemoprevention of lung carcinogenesis by cacao liquor proanthocyanidins in a male rat multi-organ carcinogenesis model. *Cancer Letters* 191, 49-57.
- Yang, J., and Xiao, Y.Y. (2013). Grape phytochemicals and associated health benefits. *Crit Rev Food Sci Nutr* 53, 1202-1225.
- Yu, Y., Zhu, D., Ma, C., Cao, H., Wang, Y., Xu, Y., Zhang, W., and Yan, Y. (2016). Transcriptome analysis reveals key differentially expressed genes involved in wheat grain development. *The Crop Journal* 4, 92-106.
- Zhang, Z., Li, B.Y., Li, X.L., Cheng, M., Yu, F., Lu, W.D., Cai, Q., Wang, J.F., Zhou, R.H., Gao, H.Q., and Shen, L. (2013). Proteomic analysis of kidney and protective effects of grape seed procyanidin B2 in db/db mice indicate MFG-E8 as a key molecule in the development of diabetic nephropathy. *Biochim Biophys Acta* 1832, 805-816.
- Zhen, S., Dong, K., Deng, X., Zhou, J., Xu, X., Han, C., Zhang, W., Xu, Y., Wang, Z., and Yan, Y. (2016). Dynamic metabolome profiling reveals significant metabolic changes during grain development of bread wheat (*Triticum aestivum* L.). *J Sci Food Agric* 96, 3731-3740.

- Zhu, M.Z., Wu, W., Jiao, L.L., Yang, P.F., and Guo, M.Q. (2015). Analysis of flavonoids in lotus (*Nelumbo nucifera*) leaves and their antioxidant activity using macroporous resin chromatography coupled with LC-MS/MS and antioxidant biochemical assays. *Molecules* 20, 10553-10565.
- Zhu, Q.Y., Holt, R.R., Lazarus, S.A., Orozco, T.J., and Keen, C.L. (2002). Inhibitory effects of cocoa flavanols and procyanidin oligomers on free radical-induced erythrocyte hemolysis. *Exp Biol Med* 227, 321-329.
- Zhu, W., Smith, J.W., and Huang, C.M. (2010). Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol* 2010, 840518.

II. Publications

This PhD thesis is based on the following first-author publications and one co-author publication, which are presented in their original format:

Wang, L., Nägele, T., Doerfler, H., Fragner, L., Chaturvedi, P., Nukarinen, E., Bellaire, A., Huber, W., Weiszmänn, J., Engelmeier, D., Ramsak, Z., Gruden, K. and Weckwerth, W. System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance. *Plant J*, 2016, 87: 318–332. doi:10.1111/tpj.13201

Wang, L[#], Fu, J.[#], Li, M., Fragner, L., Weckwerth, W., Yang, P. Metabolomic and proteomic profiles reveal the dynamic of primary metabolism during seed development of lotus (*Nelumbo nucifera*). *Front. Plant Sci.*, 2016, 7: 750. doi: 10.3389/fpls.2016.00750 (#Co-first author)

Doerfler, H., Sun, X., **Wang, L.**, Engelmeier, D., Lyon, D., Weckwerth, W. mzGroupAnalyzer- Predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. *PLoS ONE*, 2014, 9, e96188.

Additionally, one manuscript was submitted and added in the thesis:

Wang, L., Weckwerth, W. et al. Granger causality analysis of metabolite and protein dynamics during Early Campbell grape berry development

1. Metabolomic and proteomic profiles of developing cacao seed

Title: System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance

1.1 Preamble

Cacao seeds are the source for chocolate production. Besides economic values, cacao seeds are attracting more and more attention due to their potential health benefits. However, it is still lack of knowledge about the developmental process of cacao seed which is crucial for bioactive metabolite accumulation. Hence, in the following study, mass spectrometry based platforms were applied to analysis the metabolome and proteome of cacao seeds at four developmental stages covering the seed growth period from early embryogenesis to maturation. The metabolites and proteins involved in primary and secondary metabolism pathways were discussed in detail in the following publication. Such data set provides reference information for understanding and further exploration of cacao seed development.

1.2 Declaration of authorship

The results obtained from integrated analysis of metabolome and proteome of developing cacao seeds were published in *Plant Journal* as a resource paper. In this publication, I performed main part of the work including sample preparation, data acquisition and analysis, as well as manuscript writing.

1.3 Published manuscript

RESOURCE

System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance

Lei Wang¹, Thomas Nägele^{1,2}, Hannes Doerfler¹, Lena Fragner¹, Palak Chaturvedi¹, Ella Nukarinen¹, Anke Bellaire^{1,3}, Werner Huber³, Jakob Weizmann¹, Doris Engelmeier¹, Ziva Ramsak⁴, Kristina Gruden⁴ and Wolfram Weckwerth^{1,2,*}

¹Department of Ecogenomics and Systems Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria,

²Vienna Metabolomics Center (VIME); University of Vienna, Althanstrasse 14, 1090 Vienna, Austria,

³Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria, and

⁴Department of Systems Biology and Biotechnology, National Institute of Biology, Vecna pot 111, 1000 Ljubljana, Slovenia

Received 20 October 2015; revised 4 March 2016; accepted 22 April 2016; published online 5 August 2016.

*For correspondence (e-mail wolfram.weckwerth@univie.ac.at).

SUMMARY

Theobroma cacao and its popular product, chocolate, are attracting attention due to potential health benefits including antioxidative effects by polyphenols, anti-depressant effects by high serotonin levels, inhibition of platelet aggregation and prevention of obesity-dependent insulin resistance. The development of cacao seeds during fruit ripening is the most crucial process for the accumulation of these compounds. In this study, we analyzed the primary and the secondary metabolome as well as the proteome during *Theobroma cacao* cv. Forastero seed development by applying an integrative extraction protocol. The combination of multivariate statistics and mathematical modelling revealed a complex consecutive coordination of primary and secondary metabolism and corresponding pathways. Tricarboxylic acid (TCA) cycle and aromatic amino acid metabolism dominated during the early developmental stages (stages 1 and 2; cell division and expansion phase). This was accompanied with a significant shift of proteins from phenylpropanoid metabolism to flavonoid biosynthesis. At stage 3 (reserve accumulation phase), metabolism of sucrose switched from hydrolysis into raffinose synthesis. Lipids as well as proteins involved in lipid metabolism increased whereas amino acids and *N*-phenylpropenoyl amino acids decreased. Purine alkaloids, polyphenols, and raffinose as well as proteins involved in abiotic and biotic stress accumulated at stage 4 (maturation phase) endowing cacao seeds the characteristic astringent taste and resistance to stress. In summary, metabolic key points of cacao seed development comprise the sequential coordination of primary metabolites, phenylpropanoid, *N*-phenylpropenoyl amino acid, serotonin, lipid and polyphenol metabolism thereby covering the major compound classes involved in cacao aroma and health benefits.

Keywords: *Theobroma cacao*, seed development, metabolomics, proteomics, GC-MS, LC-MS, primary metabolism, secondary metabolism, multivariate statistics, plant systems biology.

INTRODUCTION

Theobroma cacao is an economically important crop with a long history of cultivation (Henderson *et al.*, 2007). Its relevance as a crop plant is underlined by the recent publication of the *Theobroma cacao* genome sequence (Argout *et al.*, 2011). Cacao and its popular product, chocolate, are attracting more and more attention because of proposed

potential health benefits including antioxidant effects, inhibition of platelet aggregation and prevention of obesity-dependent insulin resistance (Murphy *et al.*, 2003; Caligiani *et al.*, 2007, 2014; Cooper *et al.*, 2008; Rodriguez-Campos *et al.*, 2011; Dorenkott *et al.*, 2014; Teye *et al.*, 2014). Several studies have focused on the chemical composition of

cacao beans of different varieties (Elwers *et al.*, 2009) or after different processing procedures (Counet *et al.*, 2002, 2004; Magi *et al.*, 2012; Misnawi and Ariya, 2011; Rodriguez-Campos *et al.*, 2011). In contrast, few studies have focused on metabolic compositional variation during cacao seed ripening (Bucheli *et al.*, 2001; Pereira-Caro *et al.*, 2013). None of these studies investigated the interface of primary and secondary metabolism during seed development to reveal key points of regulation for flavonoid and polyphenol accumulation.

As a plant sink organ, seeds not only play a central role in the plant life cycle but are also an essential food source for humans and animals and possess enormous economic and nutritional value. Seed development has been shown to be a highly dynamic process with complex physiological and biochemical interactions (Fait *et al.*, 2006; Mounet *et al.*, 2007; Sreenivasulu *et al.*, 2008). Hence, a comprehensive and systemic understanding of the seed ripening process represents a challenge for current biological research and is necessary due to its essential role in crop quality monitoring and crop yield optimization. To promote our current understanding of cacao seed development and the underlying metabolic regulation, it is crucial to characterize the primary and secondary metabolome. Regulation of primary metabolism determines the C/N-balance on a whole plant level, and thus also directly affects seed sink strength (Götz *et al.*, 2007). Simultaneously, primary metabolism comprises the biosynthetic pathways of the aromatic amino acids (AAAs) tyrosine, tryptophan and phenylalanine being the metabolic precursors of numerous secondary metabolites. All three AAAs are synthesized in the shikimic acid pathway to which over 30% of photosynthetically fixed carbon might be directed in vascular plants (Maeda and Dudareva, 2012). Hence, it can be expected that the regulation of metabolic dynamics at the interface of primary and secondary metabolism significantly affects whole-plant metabolism during various developmental stages, including fruit ripening and seed formation.

To unravel developmental patterns of metabolic regulation between primary and secondary metabolism, it is necessary to record and analyze metabolic dynamics at their interface. In this context, metabolomics has become a central element of plant systems biology research (Weckwerth, 2003; Saito, 2013). It ultimately aims at detecting, identifying, quantifying and interpreting the occurrence and the abundance of metabolites in the context of an organism's biology. The plant kingdom provides the largest range of diverse natural products (Pichersky and Gang, 2000; Fiehn, 2002; Hall *et al.*, 2002; Weckwerth, 2003). It is widely accepted that a single technique cannot be expected to account for the detection of this enormous biochemical diversity (Weckwerth, 2003). Instead, integrated approaches comprising different analytical techniques, e.g. gas

chromatography coupled to mass spectrometry (GC-MS) and liquid chromatography coupled to MS (LC-MS) enable approaches of high-throughput metabolite profiling (Weckwerth, 2010). While GC-MS analysis predominantly focuses on the identification and quantification of small derivatized polar and volatile components, e.g. primary metabolites such as amino acids, sugars and organic acids, LC-MS is applied for the analysis of larger semi- and non-volatile and thermally unstable compounds, such as secondary metabolites (Scherling *et al.*, 2010; Weckwerth, 2011a).

Previously, we applied an integrative metabolomics platform comprising GC- and LC-MS analysis (Scherling *et al.*, 2010) as well as strategies of multivariate statistics and mathematical modelling to unravel the regulation of metabolic dynamics at the interface of primary and secondary metabolism in *Arabidopsis thaliana* (Doerfler *et al.*, 2013). We also applied this platform to characterize the phytochemical composition of medicinal plants (Mari *et al.*, 2013). In the present study, we apply and enlarge this approach to elucidate patterns of metabolic regulation during cacao seed ripening. In a recent study, we proposed the integration of metabolomics and proteomics data by investigating their covariance pattern (Wienkoop *et al.*, 2008). In line with this strategy we measured the proteome of cacao seed development using an integrative extraction protocol for metabolites and proteins from the same samples (Weckwerth *et al.*, 2004; Valledor *et al.*, 2014) and applied a novel strategy for the direct linkage of covariance pattern and metabolic modelling (Weckwerth, 2003, 2011a, b; Sun and Weckwerth, 2012; Doerfler *et al.*, 2013; Nägele *et al.*, 2014). For the shotgun proteomics approach we exploited the available *Theobroma cacao* genome sequence (Argout *et al.*, 2011) and its functional annotation using GoMapMan (Ramsak *et al.*, 2014). The integration of metabolite and protein dynamics and metabolic modelling during cacao seed development identified the relationship between protein regulation and metabolism adjustment with respect to flavonoid accumulation in mature cacao seeds. These results not only reveal a complex interplay between primary and secondary metabolism on both the metabolome and proteome level, but also represent a reference data set of cacao seed development for future studies.

RESULTS

Primary metabolism of developing cacao seeds

Cacao beans were harvested at four developmental stages (Figure 1(b)). The developmental stages are defined by the average seed weight as stage 1 (0.25 g), 2 (0.33 g), 3 (1.38 g) and 4 (1.63 g) (see supporting information Table S9). In the following analysis we always refer to stages 1 to 4. An overview of the extraction procedure and analytical platforms is provided in Figure 2. Analysis of

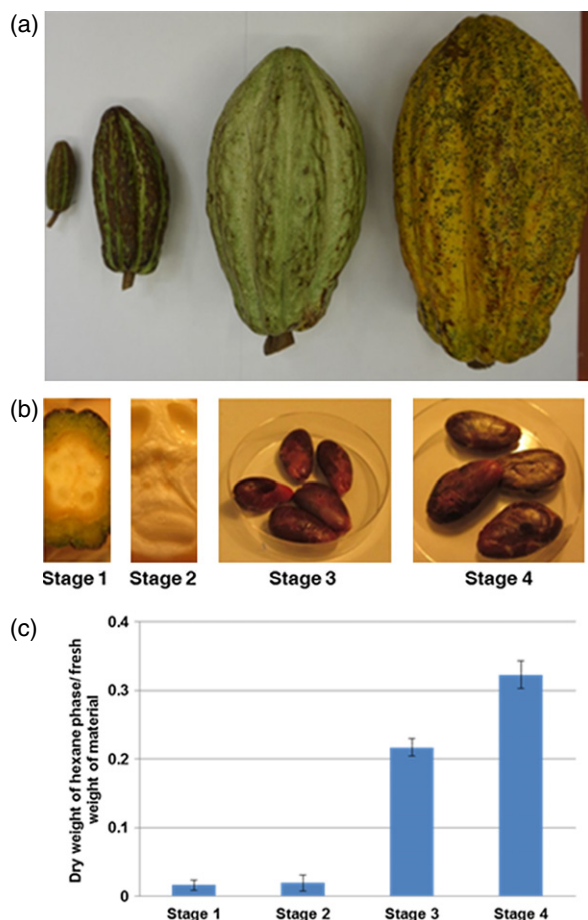


Figure 1. Cacao seed sampling and developmental stages. (a, b) Cacao fruits (a) and beans (b) at four developmental stages (supplementary Table S9). (c) Lipid content of cacao beans at four development stages. Data were normalized to fresh weight and averaged over three biological replicates.

compounds of the primary C/N-metabolism was performed on a GC-TOF-MS platform. Detailed information about retention indices and quantification masses is provided in Tables S1 and S2 as well as in supporting information Figure S1. Components of central primary C/N-metabolism, i.e. sugars, TCA intermediates and proteinogen amino acids, were visualized in a metabolic pathway diagram (Figures 3 and S1).

Of all the amino acids detected (Figure S1), alanine and cysteine were the only two candidates that decreased significantly during cacao seed development (ANOVA results are provided in Table 2). Levels of glutamine, asparagine, aspartic acid, glutamic acid, threonine, proline and lysine peaked in stage 2. In addition, aspartic acid and proline accumulated in the final stage. At developmental stage 3, ornithine was the only amino acid which showed highest level. Levels of histidine, methionine, tyrosine, phenylalanine, leucine, isoleucine, serine and valine significantly increased in stage 4. 5-Hydroxytryptamine (serotonin)

accumulated in the latest mature stage of cacao seed (Figure 3, Figure S1).

Differential dynamics were found for the metabolism of mono-, di- and trisaccharides. While levels of raffinose metabolism-associated sugars and sugar alcohols, i.e. sucrose, galactinol, galactose and raffinose, accumulated until stage 4, the free hexoses glucose and fructose increased to a maximum level in stage 2 (Figure S1). In contrast, the level of *myo*-inositol, which is a substrate for galactinol biosynthesis, decreased during seed ripening (Figure S1).

For organic acids, diverse patterns were detected (Figure S1). About half of these showed an increasing trend. One-third of them, including malic acid, which is one of the most abundant organic acids in cacao seeds, rose at the beginning and peaked at stage 2 but declined afterwards. Citric acid showed a fluctuating pattern during cacao seed development.

Purine alkaloids, theobromine and caffeine, which were below detection limit in stages 1 and 2, were sharply increased at the last stage (Figure S1) and became outstanding components in cacao beans. Most of the other alkaloids also accumulated at stage 4 except *N*-acetylgalactosamine, *N*-acetylneuraminic acid and ethanolamine.

Secondary metabolism of developing cacao seeds

The relative abundances of secondary metabolites were experimentally analyzed by an LC-Orbitrap-MS/MS platform. Here, 20 metabolites, including six *N*-phenylpropenoyl-L-amino acids (NPPAA) and 14 flavonoids, were identified based on exact precursor masses and corresponding sum formulae, associated retention times (RTs) and their fragmentation patterns (for detailed MS² interpretation see Appendix S1). The quality of annotation is indicated by metabolite identification standard levels proposed by the metabolomics standard initiative (MSI) (Sumner *et al.*, 2007) (Table 1). NPPAAs (annotated as 1, 2, 8, 9, 10, 18) were identified based on literature information (Stark and Hofmann, 2005). Associated product ions at *m/z* 147, 163 and 177 represented the *p*-coumaroyl, caffeoyl and feruloyl fragments whereas the ions with *m/z* ratios of 198 and 182 were derived from the fragments of 3-hydroxyl-tyrosyl and tyrosyl residues followed by decarboxylation and resulting ions at *m/z* 152 and 136.

Candidates 4 and 14 (Table 1) were identified as procyanidin B2 and procyanidin A2 based on consistent fragmentation patterns described previously (Appeldoorn *et al.*, 2009; Calderon *et al.*, 2009) and further confirmed with standard compounds. Candidate 5, 11 and 17 were confirmed to be cyanidin-3-*O*-arabinoside, quercetin-3-*O*-glucoside and quercetin by comparing their RTs and MS² spectra with that of authentic compounds.

The main fragments for candidates 3, 11, 12, 13, 19 and 20 (Table 1) were 162 Da less than their precursor ions which indicated the cleavage of a hexose group.

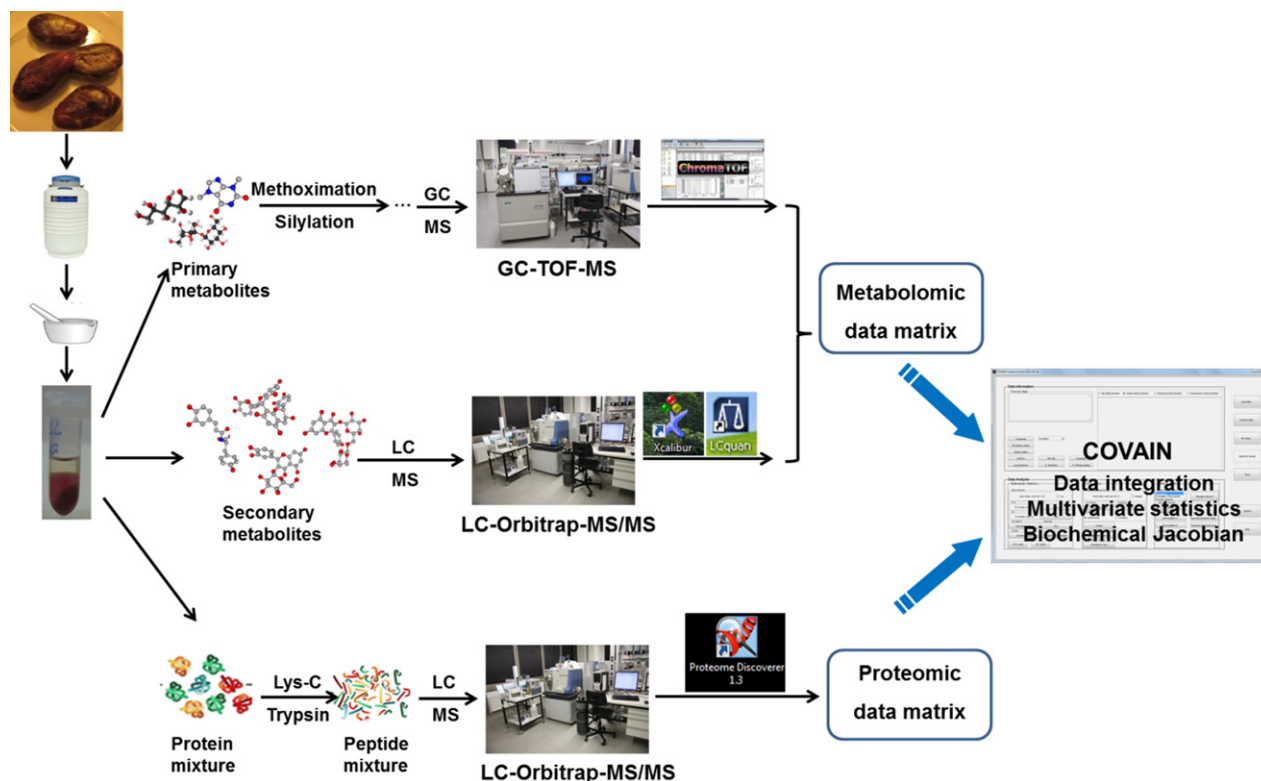


Figure 2. Integrative extraction and measurement of primary and secondary metabolites as well as proteins from the same sample.

Comparing the product ions of 5, 6, 7, 16 with their precursors, losses of 132 Da were found in all these candidates which indicated the presence of pentose moiety. Candidate 15 yielded a fragment at m/z 449 which was $[M + 1 - 146]^+$ followed by a loss of 162 Da pointing to the presence of a rutinoside moiety. The main fragments for all these flavonoids indicated the loss of sugars. The ion at m/z 287 originates from cyanidin (Cy), kaempferol (Ka), luteolin (Lu) or (iso)sakuranetin, m/z 271 from apigenin (Ap) or pelargonidin (Pel), m/z 303 from quercetin (Qu), delphinidin (Del) or tricetin (Tri) and m/z 273 from naringenin. By summing up all this information and former reports (Sanchez-Rabaneda *et al.*, 2003; Pereira-Caro *et al.*, 2013), candidates 3, 7, 13, 15, 16, 19, 20 are proposed to be Cy/Ka/Lu-hexoside, Cy/Ka/Lu-pentoside, Cy/Ka/Lu-hexoside, Ka/Cy/Lu-rutinoside, Qu/Del/Tri-pentoside, Qu/Del/tricetin-hexoside, naringenin-hexoside. Candidate 6 is proposed to be Ap/Pel-pentoside. The fragmentation pattern of Pel-3-O-Glc detected in strawberry reported in 2007 was consistent with that of candidate 6 (da Silva *et al.*, 2007). However, neither Ap-Glc nor Pel-Glc have been detected in cacao so far.

The precursor ion of metabolite candidate 12 (Table 1) was 449.1442, only 0.04 Da different from the precursor ions of candidate 3 and 13 which gave 12 a distinct formula $[C_{22}H_{25}O_{10}]^+$. The product ion at m/z 287 originated from $[M + 1 - 162]^+$ indicating a loss of a hexose unit which provides evidence that the formula of this ion was

$[C_{16}H_{15}O_5]^+$ instead of $[C_{15}H_{11}O_6]^+$ for product ions of candidates 3 and 13. The precursor mass and the fragment of 12 is 14 Da more than those of candidate 20, which was annotated as naringenin-hexoside (Sanchez-Rabaneda *et al.*, 2003); in addition, the formula of candidate 12 has an extra $-CH_2$ compared with that of candidate 20 indicating that 12 is the methylated form of 20. Based on these pieces of information, candidate 12 was deduced as (iso)sakuranetin-hexoside, which has, to our knowledge, not been described in cacao before.

Peak areas of the 20 secondary metabolites were normalized to fresh weight and to the peak area of the internal standard reserpine (Table S3). Flavonoids and NPPAAs displayed highly diverge concentrations during cacao seed development even though both compound classes are derived from the phenylpropanoid pathway. All the identified NPPAAs decreased during cacao seed ripening whereas most of the flavonoids increased (Figure 3). This differential pattern indicated a sequential induction of secondary metabolism during the ripening process of cacao beans with a distinct output in different compound classes.

Linking metabolomic covariance networks with the underlying proposed biochemical network reveals regulatory key steps during development

To promote the analysis of this complex interplay, we applied an approach of multivariate statistics and

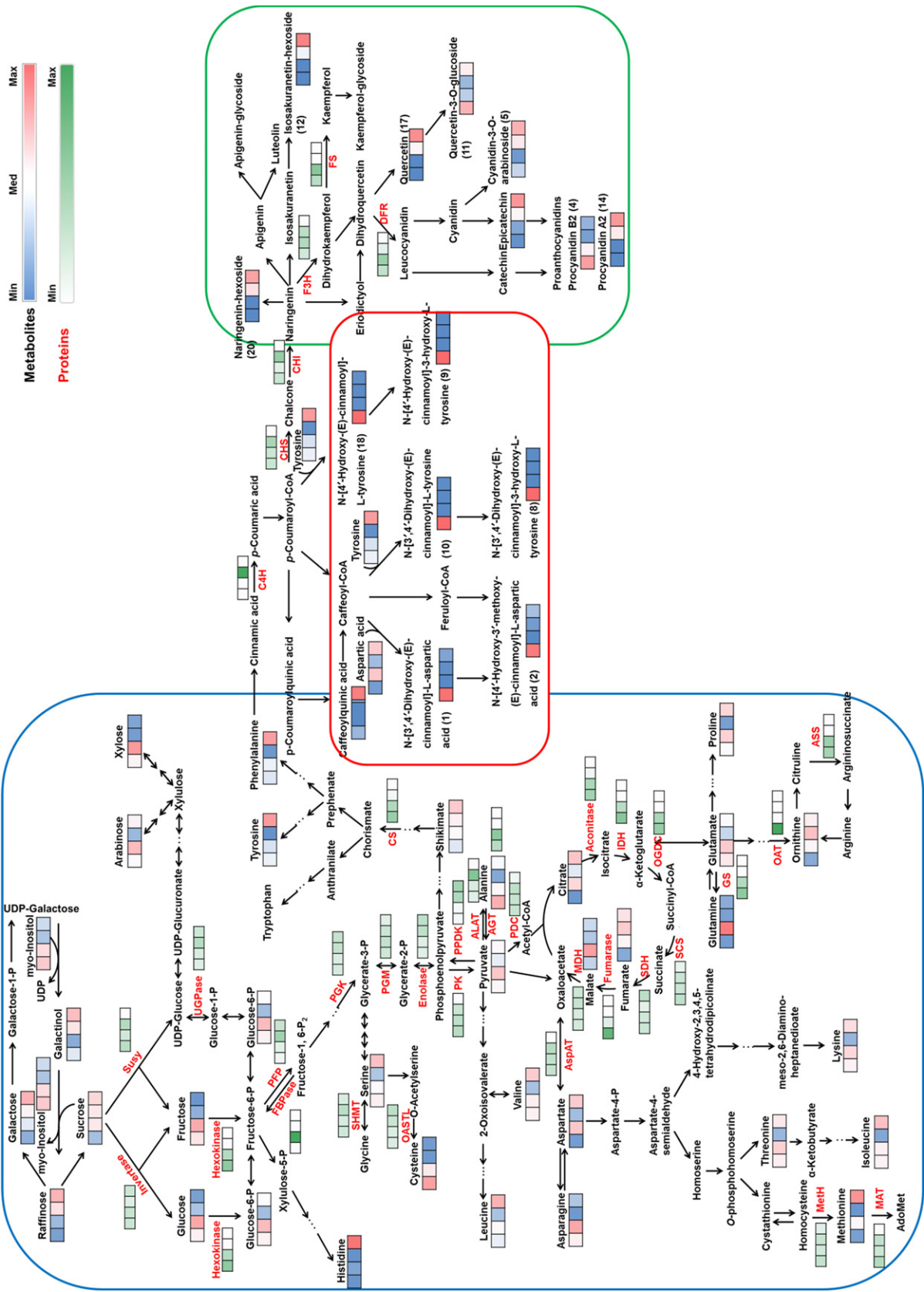


Figure 3. Visualization of metabolite and protein dynamics in a biochemical pathway map.

Primary metabolism pathways are shown within the blue frame. A proposed *N*-phenylpropanoyl-L-amino acids biosynthetic pathway is highlighted in the red frame whereas the flavonoid pathway is shown in a green frame. Metabolites are written in black letters, enzymes are written in red letters. Levels of metabolites and proteins were averaged over three biological replicates after normalization. The proportion of candidates in one stage compared to all stages was visualized in a heat map. The four consecutive squares from the left to the right side indicate stages 1 to 4. Values of metabolite levels from minimal to median to maximal are colored from blue to white to red whereas protein concentrations are colored from white to green.

Abbreviations: Susy, sucrose synthase; UGPase, UDP-glucose pyrophosphorylase; PFP, pyrophosphate-fructose 6-phosphate 1-phosphotransferase; FBPase, fructose 1,6-bisphosphatase; PGK, phosphoglycerate kinase; SHMT, serine hydroxymethyltransferase; OASTL, *O*-acetylserine (thiol)-lyase; PGM, phosphoglycerate mutase; CS, chorismate synthase; PK, pyruvate kinase; PDK, pyruvate orthophosphate dikinase; ALAT, alanine aminotransferase; PDC, pyruvate dehydrogenase complex; IDH, isocitrate dehydrogenase; OGDC, oxoglutarate dehydrogenase complex; SCS, succinyl coenzyme A synthetase; SDH, succinate dehydrogenase; MDH, malate dehydrogenase; AspAT, aspartate aminotransferase; GS, glutamine synthetase; OAT, ornithine-delta-aminotransferase; ASS, argininosuccinate synthase; C4H, cinnamate 4-hydroxylase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; FS, flavonol synthase; DFR, dihydroflavanol 4-reductase.

mathematical modelling by which Jacobian matrices were calculated from experimental metabolomic covariance data (for more details of this approach see (Nägele *et al.*, 2014)). This strategy links the causal relationship of known biochemical pathways and the statistical covariance network of metabolomics data [for details see Weckwerth (2011a,b)]. In the present study, a biochemical interaction matrix (Table S4) was derived from the pathway information of primary and secondary metabolism of a plant sink organ. This interaction matrix was derived from a pathway model (Data S1) which contained qualitative information about the biochemical interconversion of metabolites. Applying an equation combining the metabolomics covariance matrix *C* and the biochemical Jacobian matrix *J* we inversely calculated biochemical regulation directly from metabolomics data according to Nägele *et al.* (2014). In a biochemical context, entries of a Jacobian matrix characterize the dynamic properties of reaction rates with regard to (slight) changes in a (co)substrate or allosteric effector. Hence, a comparison of Jacobian entries between two stages of development provides a characterization of distinct enzymatic activities and their sensitivity to metabolic alterations (Nägele *et al.*, 2014).

Stage-specific comparison of covariance-derived Jacobian entries revealed an induction of AAA metabolism and TCA cycle from developmental stage 1 to stage 2 (Figure 4). During the later developmental stages, this metabolic pattern changed significantly and ratios of Jacobian entries revealed a significant reprogramming of pathways related to flavonoid and NPPAA metabolism (Figure 4(b,c)). At the same time, changes in AAA metabolism were significantly reduced compared to early developmental stages. Another remarkable effect was observed for the metabolism of amino acids belonging to the glutamate family. The most pronounced effect was found to occur between developmental stages 2 and 3, i.e. in parallel to the observed changes in secondary metabolism (Figure 4(b)).

Proteomic profiling of developing cacao beans

Shotgun proteomics analysis enabled the identification of 887 unique protein candidates (Table S5). Their

normalized spectral abundance factors (NSAFs) and ANOVA results are provided in Table S6. Principal components analysis (PCA) showed an overview of proteome variation during cacao seed development (Figure S2(b)). The samples of the first two stages were isolated from the other two stages on PC1 explaining more than 50% of the separation. PC2 revealed compositional variance between the samples of stage 3 and stage 4. The protein loadings and functional categories contributing most to the separation of the different developmental stages are shown in Table S7 (see below).

Protein candidates of glycolysis, TCA, amino acids and flavonoid biosynthesis are mapped on pathways of primary and secondary metabolism (Figure 3). Functional categorization and quantification (Figure 5) was exploited by assigning the protein candidates to the cacao mapping file released by GoMapMan.org (<http://www.gomappman.org/export/current/mapman>). Each functional protein category sums up the dynamics of individual proteins involved in this function. In Figure 5(a), bi-clustering reveals the dynamics of functional protein categories of cacao beans at different developmental stages. In young cacao beans, cell division, ATP synthesis, RNA processing, amino acid synthesis and activation, protein synthesis, sucrose transportation and degradation-associated proteins were highly abundant compared with mature beans. Lipid metabolism-related proteins peaked at stage 3. Protein candidates that protect cacao seed from heat or biotic stress were of high abundance in mature seeds. Simple phenol and phenylpropanoid synthesis was relatively high in young seeds compared with mature ones whereas flavonoid synthesis related protein candidates peaked in later stages (Figure 5(b); see also discussion below).

An integrative view on metabolomic and proteomic regulation during cacao seed development

PCA plots of the combined metabolomic and proteomic data displayed a separation of samples of the four developmental stages (Figure S2(c)). The loadings, i.e. separating metabolic or proteomic variables (Table S7), indicated that theobromine, caffeine, NPPAAs, tyrosine,

Table 1 Secondary metabolites identified in developing cacao beans measured by nanoLC-Orbitrap-MS/MS

Peak no.	Compound	RT (min)	[M + H] ⁺ (m/z)	Mass accuracy (ppm)	Export formula	MS/MS (m/z)	Suggestions in reference	Reference	Annotation levels
1	N-[3',4'-Dihydroxy-(E)-cinnamoyl]-L-aspartic acid	8.1	296.0766	0.32	[C ₁₃ H ₁₄ O ₇ N] ⁺	163	N-[3',4'-Dihydroxy-(E)-cinnamoyl]-L-aspartic acid	Stark and Hofmann (2005)	2
2	N-[4'-Hydroxy-3'-methoxy-(E)-cinnamoyl]-L-aspartic acid	10.5	310.0922	0.22	[C ₁₄ H ₁₆ O ₇ N] ⁺	177, 163, 130, 148	N-[4'-Hydroxy-3'-methoxy-(E)-cinnamoyl]-L-aspartic acid	Stark and Hofmann (2005)	2
3	Cy/Ka/Lu-hexoside	12.1	449.1076	-0.46	[C ₂₁ H ₂₁ O ₁₁] ⁺	287	Cy-3-O-galactoside	Pereira-Caro et al. (2013)	3
4	Procyanidin B ₂ ^a	12.6	579.1493	-0.61	[C ₃₀ H ₂₇ O ₁₂] ⁺	427, 291, 409, 247	Procyanidin B ₂ ^a	Calderon et al. (2009)	1
5	Cy-3-O-arabinoside ^a	13.4	419.0971	-0.17	[C ₂₀ H ₁₉ O ₁₀] ⁺	287	Cy-3-O-arabinoside ^a	Pereira-Caro et al. (2013)	1
6	Ap/Pel-pentoside	16.4	403.1025	0.18	[C ₂₀ H ₁₉ O ₉] ⁺	271	Pel-3-O-arabinoside	da Silva et al. (2007)	3
7	Cy/Ka/Lu-pentoside	18.3	419.0973	0.04	[C ₂₀ H ₁₉ O ₁₀] ⁺	287	Ka/Lu-pentoside	Pereira-Caro et al. (2013)	3
8	N-[3',4'-Dihydroxy-(E)-cinnamoyl]-3-hydroxy-L-tyrosine	18.3	360.1079	0.27	[C ₁₈ H ₁₈ O ₇ N] ⁺	163, 196, 152	N-[3',4'-Dihydroxy-(E)-cinnamoyl]-3-hydroxy-L-tyrosine	Stark and Hofmann (2005)	2
9	N-[4'-Hydroxy-(E)-cinnamoyl]-3-hydroxy-L-tyrosine	22.1	344.1130	0.11	[C ₁₈ H ₁₈ O ₆ N] ⁺	147, 152	N-[4'-Hydroxy-(E)-cinnamoyl]-3-hydroxy-L-tyrosine	Stark and Hofmann (2005)	2
10	N-[3',4'-Dihydroxy-(E)-cinnamoyl]-L-tyrosine	23.0	344.1128	-0.22	[C ₁₈ H ₁₈ O ₆ N] ⁺	163, 182, 147	N-[3',4'-Dihydroxy-(E)-cinnamoyl]-L-tyrosine	Stark and Hofmann (2005)	2
11	Quercetin-3-O-glucoside ^a	23.5	465.1028	0.03	[C ₂₁ H ₂₁ O ₁₂] ⁺	303	Quercetin-3-O-glucoside ^a	Pereira-Caro et al. (2013)	1
12	Isosakuranetin-hexoside	23.6	449.1442	-0.01	[C ₂₂ H ₂₅ O ₁₀] ⁺	287	Isosakuranetin-hexoside	Sanchez-Rabaneda et al. (2003)	3
13	Cy/Ka/Lu-hexoside	24.2	449.1078	-0.02	[C ₂₁ H ₂₁ O ₁₁] ⁺	287	Ka/Lu-glucoside	Sanchez-Rabaneda et al. (2003)	3
14	Procyanidin A ₂ ^a	24.5	577.1341	0.14	[C ₃₀ H ₂₅ O ₁₂] ⁺	425, 437, 287	Procyanidin A ₂ ^a	Appeldoorn et al. (2009)	1
15	Ka/Cy/Lu-rutinoside	24.9	595.1657	-0.01	[C ₂₇ H ₃₁ O ₁₅] ⁺	287, 449	Ka-3-O-rutinoside	Pereira-Caro et al. (2013)	3
16	Qu/Del/Tri-pentoside	25.7	435.0921	-0.22	[C ₂₀ H ₁₉ O ₁₁] ⁺	303	Qu-3-O-arabinoside	Sanchez-Rabaneda et al. (2003)	2
17	Quercetin ^a	25.9	303.0499	-0.17	[C ₁₅ H ₁₁ O ₇] ⁺	257, 229, 285, 165, 247	Quercetin ^a	Sanchez-Rabaneda et al. (2003)	1
18	N-[4'-Hydroxy-(E)-cinnamoyl]-L-tyrosine	26.2	328.1179	-0.03	[C ₁₈ H ₁₈ O ₆ N] ⁺	147, 136	N-[4'-Hydroxy-(E)-cinnamoyl]-L-tyrosine	Pereira-Caro et al. (2013)	2
19	Qu/Del/tricetin-hexoside	30.3	465.1027	-0.02	[C ₂₁ H ₂₁ O ₁₂] ⁺	303	Qu-3-O-glucoside	Sanchez-Rabaneda et al. (2003)	2
20	Naringenin-hexoside	31.7	435.1286	0.14	[C ₂₁ H ₂₃ O ₁₀] ⁺	273, 303	Naringenin-hexoside	Sanchez-Rabaneda et al. (2003)	3

^aConfirmed with corresponding standard.

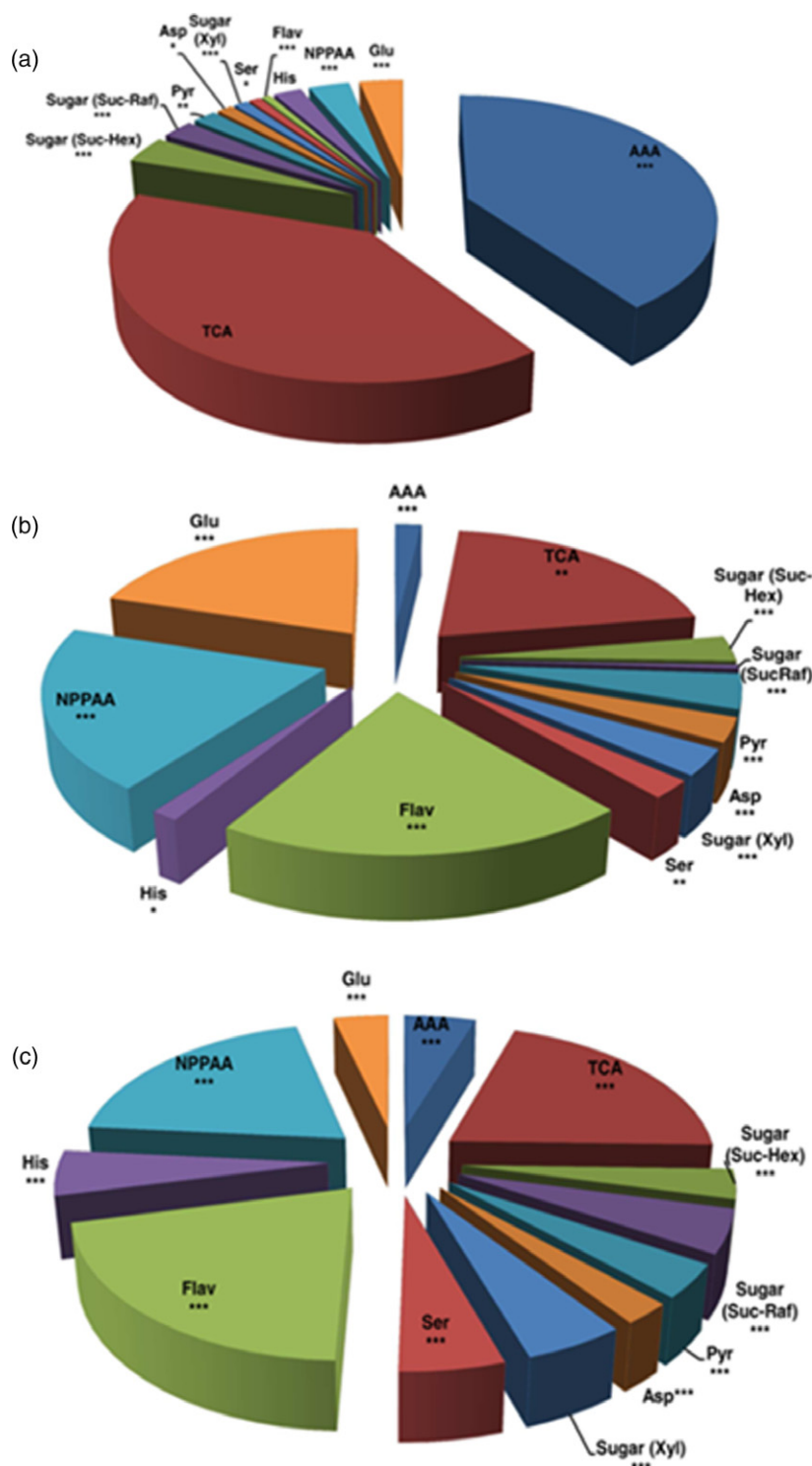
Cy, Ka, Lu, Ap, Pel, Qu, Del and Tri are abbreviations for the flavonoid aglycones cyanidin, kaempferol, luteolin, apigenin, pelargonidin, quercetin, delphinidin and tricetin respectively. Annotation levels were labelled according to the metabolite identification standard published by the metabolomics society (Sumner et al., 2007). Level 1: the compound is identified with at least two independent parameter compared with an authentic compound. Level 2: the candidate was putatively annotated based on spectral similarity with public or commercial spectral libraries. Level 3: the candidate was assigned to a compound class based on spectral similarity to a known compound of a chemical class. Level 4: unknown compounds quantified based on spectral data but without further identification and classification.

Figure 4. Entries of the biochemical Jacobian matrix during cacao seed development.

Diagrams show the relative proportion of changes in Jacobian entries with respect to the associated biochemical pathways and/or substance classes.

(a–c) The ratio of Jacobian entries comparing developmental stage 1 with stage 2 (a), stage 2 with stage 3 (b) and stage 3 with stage 4 (c) respectively. Ratios were built from grouped mean values of Jacobian entries associated to biochemical pathways as indicated by the abbreviations. A detailed list of grouped reactions and corresponding experimental data are provided in Table S4.

TCA: tricarboxylic acid cycle, AAA: aromatic amino acids, Glu: glutamate family amino acids, Ser: serine family amino acids, Asp: aspartate family amino acids, His: histidine metabolism, Pyr: pyruvate metabolism, Suc-Hex: sucrose-hexose part of sugar metabolism, Suc-Raf: sucrose-raffinose part of sugar metabolism, Xyl: xylulose-derived sugar metabolism, Flav: flavonoid metabolism, NPPAA: *N*-phenylpropenoyl-L-amino acid metabolism. Asterisks indicate significant differences of ratios (* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$).



histidine, tartaric acid and protein candidates being related to stress resistance, storage protein synthesis and secondary metabolism contributed most significantly to the sample separation.

An integrated statistical approximation of the developmental trajectory of cacao seed was performed by a

k-means clustering of metabolites and proteins revealing highly stage-specific clusters of proteins and metabolites during development (Figure S3, detailed cluster information see Table S8). The identified metabolites and protein candidates and their dynamics were mapped onto primary and secondary metabolism pathways (Figure 3).

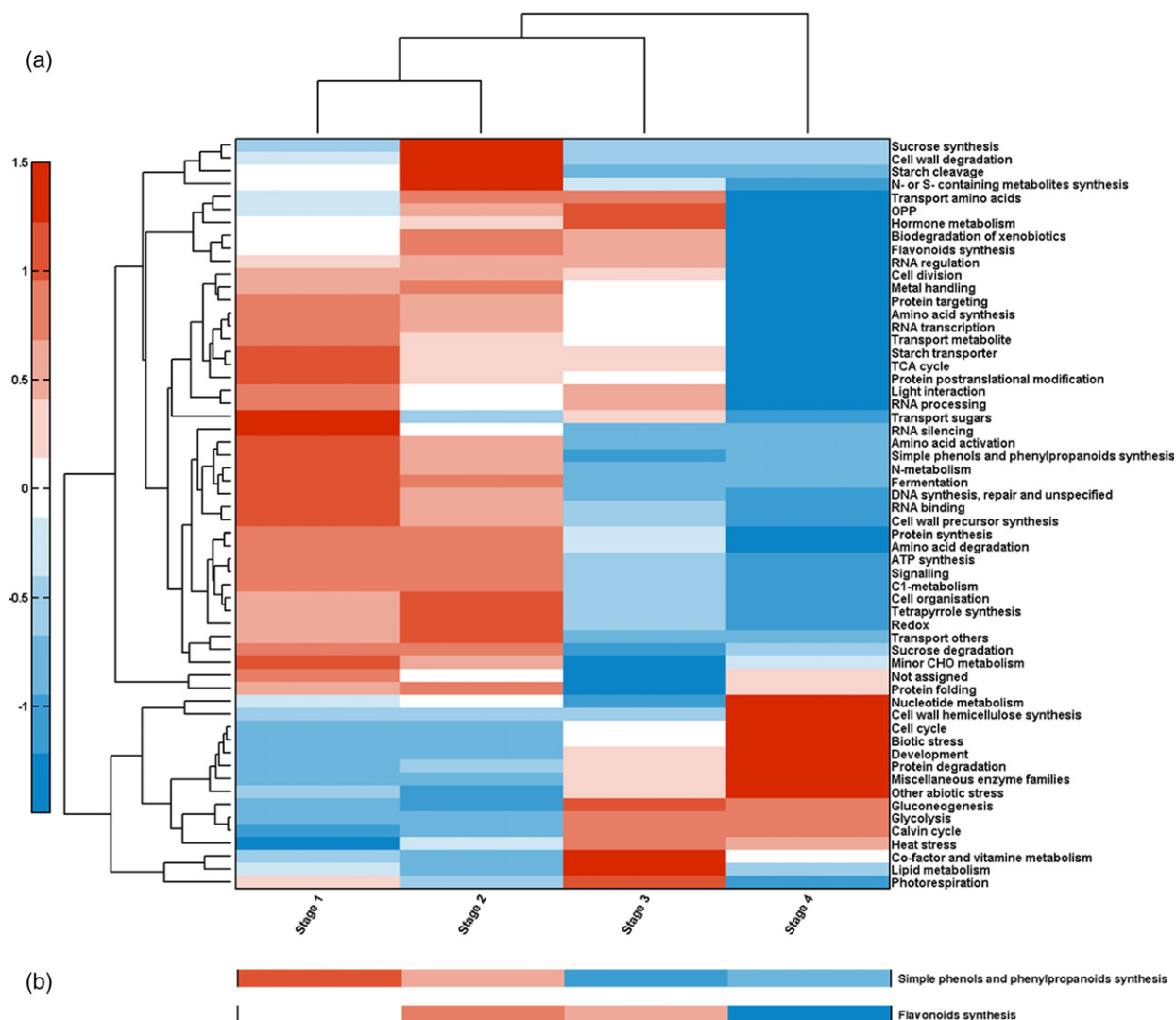


Figure 5. Functional analysis of cacao proteomic data.

(a) Heatmap of protein candidates categorized to functional bins.

(b) Functional category of simple phenols and phenylpropanoids synthesis and flavonoids synthesis. NSAF scores of the protein candidates belonging to the same functional bin were added together and averaged over the three biology replicates.

DISCUSSION

Dynamics of primary metabolism during cacao seed development

Seeds of *Theobroma cacao* represent an economically important nutritional resource. Although the diversity of products containing cacao seeds is huge, comparatively few reports on the metabolic dynamics during cacao seed development are available so far. Bucheli and co-workers studied the sugars, carboxylic acids, purine alkaloids and fatty acid variation during maturation of cacao seeds (Bucheli *et al.*, 2001). In the present study, more than 100 metabolites were quantified on a GC-TOF-MS platform, comprising sugars, amino acids, organic acids and purine alkaloids (Table S1).

Within the primary metabolome, sugars play a central role not only in seeds but on a whole-plant level. They not only provide energy resources and carbon skeletons, but also act as signaling molecules affecting and regulating gene expression (Rolland *et al.*, 2006; Wang and Ruan, 2013). Sucrose, a non-reducing disaccharide, is one of the main transport sugars in plants providing the main source for carbon allocation from source to sink organs. In seeds, sucrose is predominantly hydrolyzed to fructose and glucose during the cell division phase. This is followed by a transition and growth phase in which sucrose represents a storage product rather than a metabolic substrate for hexose synthesis (Fait *et al.*, 2006). The findings of the present study were consistent with the aforementioned pattern. Fructose and glucose increased in young cacao embryos

and peaked at stage 2 followed by a dramatic decrease from stage 3. Raffinose, which is synthesized from sucrose and galactinol (Peterbauer and Richter, 2001) significantly accumulated in mature cacao seeds. The accumulation of raffinose and other raffinose family oligosaccharides (RFOs) is a well described phenomenon which is proposed to associate with their roles as energy storage reserves and osmoprotectants (Peterbauer and Richter, 2001; Karner *et al.*, 2004; Dong *et al.*, 2013). Concomitantly, the increasing level of galactinol during the stages 3 and 4, together with raffinose, might be a potential mechanism to protect the seeds against abiotic stresses such as cold, drought and high salinity (Taji *et al.*, 2002).

Amino acid levels in seeds are significantly affected by source tissue via phloem transport (Tegeder and Rentsch, 2010; Zhang *et al.*, 2010; Galili *et al.*, 2014). They not only serve as building blocks for protein synthesis but also represent precursors for the synthesis of (essential) secondary metabolites (Field *et al.*, 2004; Tzin and Galili, 2010; de Kraker and Gershenzon, 2011; Badrinarayanan and Sperry, 2012). Additionally, amino acids can be catabolized to generate energy by incorporation into the TCA cycle (Angelovici *et al.*, 2009; Galili, 2011; Galili *et al.*, 2014). While the amino acid composition of mature, fermented or roasted cacao beans have been comprehensively described (Patzold and Bruckner, 2006; Adeyeye *et al.*, 2010), the present study additionally provides a description of amino acid dynamics during cacao seed development. Levels of shikimate-derived AAAs, i.e. phenylalanine and tyrosine, which are biochemically linked to the biosynthesis of secondary metabolites (Figure 3), sharply increased in mature beans. There are likely three reasons for the increase in phenylalanine. First, an enhanced rate of biosynthesis might be one reason for the accumulation. Second, the decrease of abundance of flavonoid synthases in mature beans (Figure 5(b)) may lead to a decline of phenylalanine conversion into flavonoid biosynthesis. Third, protein degradation processes (Figure 5(a)) were upregulated in the last stage 4, potentially leading to higher amounts of phenylalanine. Tyrosine, another aromatic amino acid in this pathway, also accumulated during cacao bean ripening which might be explained by a decreased synthesis of NPPAAs.

Most of the amino acids reached their lowest level at stage 3 (Figure 3), except ornithine, glutamine and aspartic acid. This stage might be comparable with the reserve accumulation period of *Arabidopsis* seeds during which amino acids reduced considerably. Fait and colleagues suggested that the decline of amino acid level was due to the incorporation of amino acids into storage protein synthesis rather than the decrease of amino acids synthesis (Fait *et al.*, 2006). This suggestion was based on the finding that the expression of genes related to amino acid biosynthesis increased during this period. Concomitantly,

proteomic data revealed that the transportation of amino acids and lipid metabolism were active during this period (Figure 5(a)) suggesting that the consumption but not the supply was the reason for the low abundance of amino acids at stage 3. Further evidence was obtained by comparing the relative amount of the hexane phase containing the lipid fraction of cacao seeds. Results indeed indicated that the normalized dry weight of the hexane phase increased dramatically from stage 3 on (Figure 1(c)).

Dynamics of secondary metabolism during cacao seed development

Theobromine and caffeine are two typical purine alkaloids in mature cacao seeds. Isotope tracer experiments indicated that most of them were synthesized in cotyledons (Zheng *et al.*, 2004). In PCA, they remarkably contributed to the separation of different developmental stages (Table S7). They were not detected in the first two stages while the abundance significantly increased in mature seed being consistent with previous reports (Bucheli *et al.*, 2001; Pereira-Caro *et al.*, 2013).

Besides the capability of comprehensively analyzing levels of secondary metabolites, the applied LC-Orbitrap-MS/MS platform also provides an efficient and powerful platform for putative elucidation of novel metabolite structures (Doerfler *et al.*, 2013, 2014; Mari *et al.*, 2013). In the present study, 20 secondary metabolites were putatively identified based on the accurate precursor masses, RTs and the product ion information (Table 1). The 20 annotated secondary metabolites represent phenylpropanoid derivatives and can be categorized into two categories: flavonoids and NPPAAs. Flavonoids have previously been shown to belong to the most frequently reported bioactive compounds in cacao being beneficial for human health (Grassi *et al.*, 2005; Urpi-Sarda *et al.*, 2011). NPPAAs have been suggested to be crucial contributors for astringent taste of cacao (Stark and Hofmann, 2005) and as polyphenol derivatives, they are also multifunctional (Hensel *et al.*, 2007). Similar to flavonoids, hydroxycinnamate-amino acid conjugates also showed antioxidative activity (Sanbongi *et al.*, 1998) and were suggested to protect plants during stress exposure (Alemanno *et al.*, 2003). In the present study, levels of the six identified *N*-phenylpropenoyl amino acids decreased during cacao seed development (Figure 3) which is contrasting a previous study (Pereira-Caro *et al.*, 2013) in which NPPAAs were not detected at all in young stages of cacao seeds. However, NPPAAs, such as clovamide (8; Table 1), *N*-*trans*-*p*-coumaroyl-L-tyrosine (18; Table 1) and *N*-*trans*-caffeoyl-L-tyrosine (10; Table 1) were reported to exist in cacao embryo cultured *in vitro* (calli) (Alemanno *et al.*, 2003).

Covariance-derived entries of Jacobian matrices indicated the induction of AAA metabolism in the early

developmental stages being essential for the subsequent induction of secondary metabolite biosynthesis (Figure 4). This is consistent with the variation of flavonoid synthesis related proteins as well as the flavonoid accumulation pattern. At stage 1, high abundance of protein candidates involved in simple phenol and phenylpropanoid synthesis prepared the biosynthesis of flavonoid precursors (Figure 5(b)). Subsequently, flavonoid synthesis related proteins increased (Figure 5(b)) leading to flavonoid accumulation (Figure 3).

Data integration and metabolic modelling at the interface of primary and secondary metabolism identifies the developmentally controlled preparation phases of cacao seeds for the accumulation of distinct flavonoid structures and abiotic and biotic stress protection.

We applied an integrative extraction protocol which allows for the extraction of primary and secondary metabolites together with proteins from the same sample (Weckwerth *et al.*, 2004; Doerfler *et al.*, 2013; Valledor *et al.*, 2014). This method enabled the comprehensive integration of primary and secondary metabolite dynamics derived from mass spectrometry-based metabolite profiles for developing cacao seeds. A dimensionality reduction of the combined data set by PCA (Figure S2) resulted in a developmental trajectory of the cacao seed metabolome. Although all developmental stages could be clearly separated from each other, the separation of the young beans (stages 1 and 2) from the mature beans (stages 3 and 4) was most significant (PC1). This separation was found to be a combinatorial effect of primary and secondary compounds, comprising sugars, amino acids and alkaloids like theobromine and caffeine (Table S7). A similar pattern was also observed when combining the metabolomic data set with the proteomic data (Figure S2(c)).

The overview of dynamic proteomic and metabolomic profiles enabled the monitoring of metabolic switches. During the cell division and proliferation phase (stage 1), sucrose seemed to be primarily hydrolyzed to fructose and glucose followed by incorporation into the TCA cycle to generate sufficient energy. The intermediates produced at this stage might represent metabolic precursors for secondary metabolite synthesis. In the subsequent reserve accumulation phase, a significant reprogramming of sugar metabolism was observed. Instead of being hydrolyzed to hexoses, our findings indicate that sucrose was rather substrate for raffinose biosynthesis. Hence, during maturation, the accumulation of galactinol and raffinose together with the increase of stress-related proteins (Figure 5(a)) might synergistically protect the seed from biotic and/or abiotic stresses.

Recently, we presented a strategy of biochemical modelling and comprehensive experimental data integration for the analysis of the interface of primary and

secondary metabolism in stress-exposed plants of *Arabidopsis thaliana* (Doerfler *et al.*, 2013). In another study, we could also successfully link the prediction of biochemical modelling to the level of enzymatic regulation (Nägele *et al.*, 2014). Based on these previous studies, the metabolic interaction network between primary and secondary metabolism was expanded to cacao seed metabolism. Similar modelling strategies were also recently used by other groups to predict metabolic dynamics in complex biochemical systems (Oksuz *et al.*, 2013). Our modelling results indicated a significant reprogramming of sugar and, particularly, AAA metabolism prior to the induction of secondary metabolism (Figure 4). Particularly, metabolism of glutamate family amino acids was found to be significantly activated as glutamate is a co-substrate for AAA biosynthesis (Maeda *et al.*, 2011; Yoo *et al.*, 2013).

Finally, the integration of metabolite and protein dynamics during cacao seed development identified the relationship between protein regulation and metabolic adjustment with respect to flavonoid accumulation in mature cacao seeds. An early shift in the synthesis of related biosynthetic proteins before the accumulation of flavonoids was observed. Proteins related to the synthesis of simple phenols and phenylpropanoids which are precursors of flavonoid biosynthesis were in high abundance in stage 1 (Figure 5(b)). Compared with stage 1, an induction of AAA metabolism at stage 2 was revealed by stage-specific comparison of covariance-derived Jacobian entries (Figure 4) indicating the synthesis of precursors for flavonoid biosynthesis. Finally, the abundance of enzymes catalyzing flavonoid synthesis increased from stage 2 to stage 3 (Figure 5(b)) followed by an accumulation of the typical cacao seed flavonoids (Figure 3).

In conclusion, the integrative view on experimental high-throughput data remains a challenge in various fields of current biological research. Yet, particularly with respect to developmental patterns and physiological processes like seed ripening, the biochemical interpretation of metabolomic and proteomic data sets is necessary in order to promote the understanding of whole-plant metabolism. In the economically important plant *Theobroma cacao*, seed ripening comprises comprehensive steps of reprogramming of primary metabolism, secondary metabolism and their metabolomic interface. Various proteins and substance classes are affected in their concentration and function, and a highly interrelated and fine-tuned picture of metabolism reprogramming emerges (see Figure 3). In summary, the presented findings indicate a highly interlaced network of metabolic regulation which is the regulatory output and input of a complex sink tissue. Here, we present an integrative metabolomic and proteomic reference data set for cacao

seed development from the very young to the mature stages. In future studies, these dynamics will be analyzed in detail. In this context, the balance between the *N*-phenylpropenoyl-L-amino acid and flavonoid biosynthetic pathways and the accumulation of proteins and metabolites involved in biotic and abiotic stress resistance are of particular interest.

EXPERIMENTAL PROCEDURES

Plant material

Four stages of cacao fruits (Figure 1(a) and Table S9) were harvested from an adult tree of *Theobroma cacao* (Forastero, specimen number: 3776 (WU)) planted at the La Gamba Tropical Field station (8°42'00.9"N, 83°11'35.4"W), a research center of the University of Vienna in Costa Rica (<http://lagamba.univie.ac.at/>) in December 2013. The pulp and testa covering the seeds of stages 3 and 4 were peeled off (Figure 1(b)). In total, 12 samples were collected (three biological replicates for each stage, three to six seeds/replicate). Seeds were frozen in liquid nitrogen and then stored at -80°C until further processing.

Integrative extraction of metabolites and proteins from developing cacao seed

An overview of the extraction procedure and analytical platforms is shown in Figure 2. Metabolites and proteins were extracted from the developing cacao seeds based on a universal isolation protocol (Weckwerth *et al.*, 2004) with some modifications. The seeds were ground to a fine powder with mortar and pestle under continuous cooling with liquid nitrogen. Next, 50 to 100 mg of material were used for metabolite extraction; 750 µl extraction solution (methanol:water:formic acid = 70:28:2) and 250 µl hexane were added followed by vigorous vortexing. The homogenized mixture was incubated 30 min on ice and vortexed every 6 min. Lipophilic and hydrophilic phases were separated by centrifugation for 8 min at 20 000 *g*. The lipophilic and hydrophilic phases were separated into new tubes. The extraction procedure was repeated once and the lipophilic and hydrophilic phases were pooled with the phases from the first extraction. Extracts were then dried under a fume hood and in a vacuum concentrator respectively. Total lipid analysis was performed by drying the lipophilic phase and weighing. The pellets were further used for protein extraction based on the protocol described before (Noah *et al.*, 2013). Protein concentration was determined by the Bradford method with a BSA standard curve. 100 µg of protein was reduced and alkylated with dithiothreitol (DTT, 5 mM, 37°C, 45 min) and iodoacetamide (IAA, 10 mM, 23°C, dark, 60 min) followed by the addition of 5 mM of DTT (23°C, dark, 15 min). Endo-proteinase LysC and trypsin were applied for digestion based on the protocol reported before (Hoehenwarter *et al.*, 2008). After digestion, samples were desalted with a C18-SPEC-96 well plate (15 mg; Agilent, <http://www.agilent.com/home>) according to manufacturer's instruction. The eluted peptides were dried under vacuum.

GC-TOF-MS analysis of primary metabolites

The dried MeOH phases were re-dissolved in 500 µl of 50% MeOH from which 50 µl was taken and mixed with 10 µl of ¹³C₆-sorbitol (0.1 mg ml⁻¹; internal standard). After drying, samples were methoxymethylated and silylated as described previously (Weckwerth *et al.*, 2004; Mari *et al.*, 2013).

After derivatization, samples were analyzed on an Agilent® 6890 gas chromatograph coupled to a LECO Pegasus® (<http://de.leco-europe.com/category/separation-science-mass-spectrometry/>) 4D GC×GC-TOF mass spectrometer. Instrument parameters were set as described previously (Doerfler *et al.*, 2013).

NanoLC-Orbitrap-MS/MS for secondary metabolite analysis

Here, 10 µl of the re-dissolved MeOH phase were mixed with 2.5 µl of reserpine (5 mg L⁻¹), 10 µl of 1.0% formic acid (FA) solution and 77.5 µl of water. The prepared samples were centrifuged for 8 min at 20 000 *g*. Supernatants were transferred to vials for measurement.

Next, 5 µl of prepared samples were injected using a HTC PAL Autosampler device coupled to an Eksigent nano LC pump and separated on a Waters ACQUITY UPLC HSS T3 nanoACQUITY Column (particle size 1.8 µm, dimension 100 µm × 100 mm) eluted with a non-linear gradient (Mari *et al.*, 2013) at a constant flow rate of 500 nl min⁻¹.

LTQ Orbitrap XL™ mass spectrometer (Thermo Fisher Scientific, <https://www.thermofisher.com/de/de/home.html>) was used for mass analysis. Ionization was performed on a nano ESI source (Thermo Fisher Scientific) in positive mode. Each full scan was followed by one MS/MS scan in which the most abundant ion was selected and fragmented by collision-induced dissociation (CID) with 35% normalized collision energy and 90 msec activation time. Minimum signal threshold was set to 50 000.

NanoLC-Orbitrap-MS/MS for proteome analysis

Here, 1 µg of digested protein was loaded onto a Ascentis Peptide ES-C18 column (particle size 2.7 µm, dimension 15 cm × 100 µm; Sigma-Aldrich, <https://www.sigmaaldrich.com/>) and eluted with a 100 min linear gradient from 5 to 85% of mobile phase B (90% ACN, 0.1% FA; phase A, 0.1% FA in water) at a constant flow rate of 400 nl min⁻¹ according to Chaturvedi *et al.*, 2013 and Ischebeck *et al.*, 2014.

The same mass spectrometry equipment configuration used for metabolite measurement was also used for proteome analysis. Each full scan was followed by 10 MS/MS scans in which the 10 most abundant ions were selected and fragmented by CID with 35% normalized collision energy and 30 msec activation time. Minimum signal threshold was set to 10 000.

Metabolite identification and quantification

Analysis of GC-TOF-MS raw data was performed with the software LECO CHROMATO® (<http://de.leco-europe.com/category/separation-science-mass-spectrometry/>). RTs of peaks were converted into retention indices (RIs) based on the RTs of the spiked alkanes. Peaks were manually annotated by comparing their RIs and mass spectra to those described in the GMD Golm database (Kopka *et al.*, 2005) with a minimum match factor of 700. A reference list containing the information of all the annotated compounds was constructed and applied to all the samples. Peak areas were used for relative quantification. All metabolomics data are available in MetaboLights (<http://www.ebi.ac.uk/metabolights/>) (Haug *et al.*, 2013).

Accurate precursor masses, formulae, RT together with mass accuracy were exported from Xcalibur (Thermo Xcalibur 2.2 SP1.48) for secondary metabolite identification from nanoLC-orbitrap-MS/MS data. A LCQuan (Thermo, v2.6.6.1128) was used for peak isolation and peak area integration. The obtained peak area information was exported and used for further statistical analysis.

Protein identification and quantification

The SEQUEST algorithm in PROTEOME DISCOVERER (v 1.3, Thermo Scientific) was employed to search raw data against a cacao fasta file from PHYTOZOME version 10.3 (<http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10>) including 44 404 proteins sequences. Processing parameters were set as follows: trypsin as the digestion enzyme with maximum two missed cleavage sites, 10 ppm and 0.8 Da for precursor and fragment mass tolerance respectively, modifications were set as acetylation for N-terminal modification, oxidation of methionine for dynamic modification and carbamidomethylation of cysteine for static modification, false discovery rate (FDR) was set to 0.01. The processed data were then filtered. Only the candidates with high confidence, defined by at least two peptides with Xcorr of 2, 2.25, 2.5, 2.75, 3.0 for peptides of charge 2, 3, 4, 5, 6 respectively and existed at the three biological replicates of at least one stage were considered for the statistical analysis. NSAFs were calculated (Zybailov *et al.*, 2006) for relative quantification. Functional categorization was exploited by assigning the protein candidates to the cacao mapping file released by GoMapMan.org (<http://www.gomapman.org/export/current/mapman>). All proteomics data are available in ProteomXchange (Project accession: PXD002922) (<http://www.proteomexchange.org/>) (Vizcaino *et al.*, 2014).

Bioinformatics, statistics and modelling

Analysis of variance (ANOVA) and *k*-means clustering were performed within the numerical software environment MATLAB® (V8.4.0 R2014b). Significant levels were presented with lower case letters according to the result of Duncan's test ($P < 0.05$). PCA and hierarchical clustering heat map were calculated using COVAIN (Sun and Weckwerth, 2012). The functional integration of GC-MS and LC-MS metabolomics data into a biochemical metabolic network structure was performed by the inverse approximation of the biochemical Jacobian matrix (Nägele, 2014). This approximation directly connects the covariance matrix *C*, which was built from the experimental metabolomics data, with a metabolic network structure of the primary metabolism with an interconnection to secondary metabolism. The metabolic network model is provided in SBML file format in the Data S1 supplement and can be visualized by open source software, e.g. CELLDISIGNER™ (<http://www.celldesigner.org/>) or COPASI (http://www.copasi.org/tiki-view_articles.php). The linkage of covariance data with the network structure is described by equation 1 (Sun and Weckwerth, 2012):

$$JC + CJ^T = -2D, \quad (1)$$

where *J* represents the Jacobian matrix and *D* is a fluctuation matrix which integrates a Gaussian noise function simulating metabolic fluctuations around a steady-state condition. In a biochemical context, entries of the Jacobian matrix *J* represent the elasticity of reaction rates to any change of metabolite concentrations which are characterized by equation 2:

$$J = N \frac{\partial r}{\partial M} \quad (2)$$

N is the stoichiometric matrix or a metabolic interaction matrix describing the interdependencies of metabolic fluxes and metabolites; *r* represents the rates for each reaction and *M* represents metabolite concentrations. As stated before, the Jacobian

approximation comprises the stochastic term *D*. Therefore, we performed 10×10^3 inverse approximations for each population, finally resulting in 10 technical replicates of the Jacobian matrices. All calculations of Jacobian matrices were performed based on a modified script contained in COVAIN (Sun and Weckwerth, 2012; Nägele *et al.*, 2014).

ACKNOWLEDGEMENTS

We would like to thank our colleagues from the Tropical Research Station La Gamba of the University of Vienna and especially Werner Huber for his profound support. We also thank the reviewers and MoSys members for their helpful comments and suggestions. L.W. was supported by a PhD scholarship provided by China Scholarship Council (CSC) (Grant number: 201206220134). T.N. and H.D. were funded by the Austrian Science Fund (FWF, P 26342-B21 and I 2071). E.N. was supported by the FP7 Marie-Curie ITN MERIT fellowship (GA 264474).

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Heat map showing metabolite dynamics during cacao seeds ripening.

Figure S2. PCA plots of developing cacao beans.

Figure S3. K-means clustering analysis of metabolite and protein variation during cacao bean development (Table S8).

Table S1. Information of metabolite candidates from GC-MS data.

Table S2. Quantification and ANOVA result of candidates from GC-MS data.

Table S3. Quantification and ANOVA result of candidates from LC-MS data.

Table S4. Grouped reactions for Jacobian entries.

Table S5. Sequence information of protein candidates.

Table S6. Qualitative and quantitative information of protein candidates.

Table S7. Loading information of PCA plot.

Table S8. *k*-means clustering result.

Table S9. Summary of fruit weight and size.

Appendix S1. Fragmentation deduction of the annotated candidates.

Data S1. Supplement_SBML_Model_Cacao.txt.

REFERENCES

- Adeyeye, E.I., Akinyeye, R.O., Ogunlade, I., Olaofe, O. and Boluwade, J.O. (2010) Effect of farm and industrial processing on the amino acid profile of cocoa beans. *Food Chem.* **118**, 357–363.
- Alemanno, L., Ramos, T., Gargadenec, A., Andary, C. and Ferriere, N. (2003) Localization and identification of phenolic compounds in *Theobroma cacao* L. somatic embryogenesis. *Ann. Bot.* **92**, 613–623.
- Angelovici, R., Fait, A., Zhu, X., Szymanski, J., Feldmesser, E., Fernie, A.R. and Galili, G. (2009) Deciphering transcriptional and metabolic networks associated with lysine metabolism during Arabidopsis seed development. *Plant Physiol.* **151**, 2058–2072.
- Appeldoorn, M.M., Sanders, M., Vincken, J.-P., Cheynier, V., Le Guernevé, C., Hollman, P.C.H. and Gruppen, H. (2009) Efficient isolation of major procyanidin A-type dimers from peanut skins and B-type dimers from grape seeds. *Food Chem.* **117**, 713–720.

- Argout, X., Salse, J., Aury, J.M. *et al.* (2011) The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108.
- Badrinarayanan, S. and Sperry, J. (2012) Pyrazine alkaloids via dimerization of amino acid-derived α -amino aldehydes: biomimetic synthesis of 2,5-diisopropylpyrazine, 2,5-bis(3-indolylmethyl)pyrazine and actinopolymorphol C. *Org. Biomol. Chem.* **10**, 2126–2132.
- Bucheli, P., Rousseau, G., Alvarez, M., Laloi, M. and McCarthy, J. (2001) Developmental variation of sugars, carboxylic acids, purine, alkaloids, fatty acids, and endoproteinase activity during maturation of *Theobroma cacao* L. seeds. *J. Agric. Food Chem.* **49**, 5046–5051.
- Calderon, A.I., Wright, B.J., Hurst, W.J. and van Breemen, R.B. (2009) Screening antioxidants using LC-MS: case study with cocoa. *J. Agric. Food Chem.* **57**, 5693–5699.
- Caligiani, A., Cirilini, M., Palla, G., Ravaglia, R. and Arlorio, M. (2007) GC-MS detection of chiral markers in cocoa beans of different quality and geographic origin. *Chirality*, **19**, 329–334.
- Caligiani, A., Palla, L., Acquotti, D., Marseglia, A. and Palla, G. (2014) Application of ^1H NMR for the characterisation of cocoa beans of different geographical origins and fermentation levels. *Food Chem.* **157**, 94–99.
- Chaturvedi, P., Ischebeck, T., Egelhofer, V., Lichtscheidl, I. and Weckwerth, W. (2013) Cell-specific analysis of the tomato pollen proteome from pollen mother cell to mature pollen provides evidence for developmental priming. *J. Proteome Res.* **12**, 4892–4903.
- Cooper, K.A., Donovan, J.L., Waterhouse, A.L. and Williamson, G. (2008) Cocoa and health: a decade of research. *Br. J. Nutr.* **99**, 1–11.
- Counet, C., Callemien, D., Ouwerx, C. and Collin, S. (2002) Use of gas chromatography-olfactometry to identify key odorant compounds in dark chocolate, comparison of samples before and after conching. *J. Agric. Food Chem.* **50**, 2385–2391.
- Counet, C., Ouwerx, C., Rosoux, D. and Collin, S. (2004) Relationship between procyanidin and flavor contents of cocoa liquors from different origins. *J. Agric. Food Chem.* **52**, 6243–6249.
- Doerfler, H., Lyon, D., Nägele, T., Sun, X., Fragner, L., Hadacek, F., Egelhofer, V. and Weckwerth, W. (2013) Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics*, **9**, 564–574.
- Doerfler, H., Sun, X., Wang, L., Engelmeier, D., Lyon, D. and Weckwerth, W. (2014) mzGroupAnalyzer-predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. *PLoS ONE*, **9**, e96188.
- Dong, J.Z., Yan, W., Bock, C., Nokhrina, K., Keller, W. and Georges, F. (2013) Perturbing the metabolic dynamics of *myo*-inositol in developing *Brassica napus* seeds through *in vivo* methylation impacts its utilization as phytate precursor and affects downstream metabolic pathways. *BMC Plant Biol.* **13**, 84.
- Dorenkott, M.R., Griffin, L.E., Goodrich, K.M. *et al.* (2014) Oligomeric cocoa procyanidins possess enhanced bioactivity compared to monomeric and polymeric cocoa procyanidins for preventing the development of obesity, insulin resistance, and impaired glucose tolerance during high-fat feeding. *J. Agric. Food Chem.*, **62**, 2216–2227.
- Elwers, S., Zambrano, A., Rohsius, C. and Lieberei, R. (2009) Differences between the content of phenolic compounds in Criollo, Forastero and Trinitario cocoa seed (*Theobroma cacao* L.). *Eur. Food Res. Technol.* **229**, 937–948.
- Fait, A., Angelovici, R., Less, H., Ohad, I., Urbanczyk-Wochniak, E., Fernie, A.R. and Galili, G. (2006) Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. *Plant Physiol.* **142**, 839–854.
- Fiehn, O. (2002) Metabolomics-the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171.
- Field, B., Cardon, G., Traka, M., Botterman, J., Vancanneyt, G. and Mithen, R. (2004) Glucosinolate and amino acid biosynthesis in Arabidopsis. *Plant Physiol.* **135**, 828–839.
- Galili, G. (2011) The aspartate-family pathway of plants: linking production of essential amino acids with energy and stress regulation. *Plant Signal Behav.* **6**, 192–195.
- Galili, G., Avin-Wittenberg, T., Angelovici, R. and Fernie, A.R. (2014) The role of photosynthesis and amino acid metabolism in the energy status during seed development. *Front. Plant Sci.* **5**, 447.
- Götz, K.P., Staroske, N., Radchuk, R., Emery, R.J., Wutzke, K.D., Herzog, H. and Weber, H. (2007) Uptake and allocation of carbon and nitrogen in *Vicia narbonensis* plants with increased seed sink strength achieved by seed-specific expression of an amino acid permease. *J. Exp. Bot.* **58**, 3183–3195.
- Grassi, D., Necozione, S., Lippi, C., Croce, G., Valeri, L., Pasqualetti, P., Desideri, G., Blumberg, J.B. and Ferri, C. (2005) Cocoa reduces blood pressure and insulin resistance and improves endothelium-dependent vasodilation in hypertensives. *Hypertension*, **46**, 398–405.
- Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L. and Bino, R. (2002) Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell*, **14**, 1437–1440.
- Haug, K., Salek, R.M., Conesa, P. *et al.* (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, D781–D786.
- Henderson, J.S., Joyce, R.A., Hall, G.R., Hurst, W.J. and McGovern, P.E. (2007) Chemical and archaeological evidence for the earliest cacao beverages. *Proc. Natl Acad. Sci.* **104**, 18937–18940.
- Hensel, A., Deters, A.M., Müller, G., Stark, T., Wittschier, N. and Hofmann, T. (2007) Occurrence of *N*-phenylpropenoyl-L-amino acid amides in different herbal drugs and their influence on human keratinocytes, on human liver cells and on adhesion of *Helicobacter pylori* to the human stomach. *Planta Med.* **73**, 142–150.
- Hoehenwarter, W., van Dongen, J.T., Wienkoop, S., Steinfath, M., Hummel, J., Erban, A., Sulpice, R., Regierer, B., Kopka, J. and Geigenberger, P. (2008) A rapid approach for phenotype – screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics*, **8**, 4214–4225.
- Ischebeck, T., Valledor, L., Lyon, D., Gingl, S., Nagler, M., Meijón, M., Egelhofer, V. and Weckwerth, W. (2014) Comprehensive cell-specific protein analysis in early and late pollen development from diploid microsporocytes to pollen tube growth. *Mol. Cell Proteomics*, **13**, 295–310.
- Karner, U., Peterbauer, T., Raboy, V., Jones, D.A., Hedley, C.L. and Richter, A. (2004) *myo*-inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds. *J. Exp. Bot.* **55**, 1981–1987.
- Kopka, J., Schauer, N., Krueger, S. *et al.* (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21**, 1635–1638.
- de Kraker, J.W. and Gershenzon, J. (2011) From amino acid to glucosinolate biosynthesis: protein sequence changes in the evolution of methylthioalkylmalate synthase in Arabidopsis. *Plant Cell*, **23**, 38–53.
- Maeda, H. and Dudareva, N. (2012) The shikimate pathway and aromatic amino acid biosynthesis in plants. *Annu. Rev. Plant Biol.* **63**, 73–105.
- Maeda, H., Yoo, H. and Dudareva, N. (2011) Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. *Nat. Chem. Biol.* **7**, 19–21.
- Magi, E., Bono, L. and Di Carro, M. (2012) Characterization of cocoa liquors by GC-MS and LC-MS/MS: focus on alkylpyrazines and flavanols. *J. Mass Spectrom.* **47**, 1191–1197.
- Mari, A., Lyon, D., Fragner, L., Montoro, P., Piacente, S., Wienkoop, S., Egelhofer, V. and Weckwerth, W. (2013) Phytochemical composition of *Potentilla anserina* L. analyzed by an integrative GC-MS and LC-MS metabolomics platform. *Metabolomics*, **9**, 599–607.
- Misnawi, and Ariya, B.T.S. (2011) Use of gas chromatography-olfactometry in combination with solid phase micro extraction for cocoa liquor aroma analysis. *Int. Food Res. J.* **18**, 829–835.
- Mounet, F., Lemaire-Chamley, M., Maucourt, M. *et al.* (2007) Quantitative metabolic profiles of tomato flesh and seeds during fruit development: complementary analysis with ANN and PCA. *Metabolomics*, **3**, 273–288.
- Murphy, K.J., Chronopoulos, A.K., Singh, I., Francis, M.A., Moriarty, H., Pike, M.J., Turner, A.H., Mann, N.J. and Sinclair, A.J. (2003) Dietary flavanols and procyanidin oligomers from cocoa (*Theobroma cacao*) inhibit platelet function. *Am. J. Clin. Nutr.* **77**, 1466–1473.
- Nägele, T. (2014) Linking metabolomics data to underlying metabolic regulation. *Front. Mol. Biosci.* **1**, 1–22.
- Nägele, T., Mair, A., Sun, X., Fragner, L., Teige, M. and Weckwerth, W. (2014) Solving the differential biochemical Jacobian from metabolomics covariance data. *PLoS ONE*, **9**, e92299.
- Noah, A.M., Niemenak, N., Sunderhaus, S., Haase, C., Omokolo, D.N., Winkelman, T. and Braun, H.P. (2013) Comparative proteomic analysis of early somatic and zygotic embryogenesis in *Theobroma cacao* L. *J. Proteomics*, **78**, 123–133.

- Oksuz, M., Sadikoglu, H. and Cakir, T. (2013) Sparsity as cellular objective to infer directed metabolic networks from steady-state metabolome data: a theoretical analysis. *PLoS ONE*, **8**, e84505.
- Patzold, R. and Bruckner, H. (2006) Gas chromatographic determination and mechanism of formation of D-amino acids occurring in fermented and roasted cocoa beans, cocoa powder, chocolate and cocoa shell. *Amino Acids*, **31**, 63–72.
- Pereira-Caro, G., Borges, G., Nagai, C., Jackson, M.C., Yokota, T., Crozier, A. and Ashihara, H. (2013) Profiles of phenolic compounds and purine alkaloids during the development of seeds of *Theobroma cacao* cv. Trinitario. *J. Agric. Food Chem.* **61**, 427–434.
- Peterbauer, T. and Richter, A. (2001) Biochemistry and physiology of raffinose family oligosaccharides and galactosyl cyclitols in seeds. *Seed Sci. Res.* **1**, 185–197.
- Pichersky, E. and Gang, D.R. (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends Plant Sci.* **5**, 439–445.
- Ramsak, Z., Baebler, S., Rotter, A., Korbar, M., Mozetic, I., Usadel, B. and Gruden, K. (2014) GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic Acids Res.* **42**, D1167–D1175.
- Rodriguez-Campos, J., Escalona-Buendía, H.B., Orozco-Avila, I., Lugo-Cervantes, E. and Jaramillo-Flores, M.E. (2011) Dynamics of volatile and non-volatile compounds in cocoa (*Theobroma cacao* L.) during fermentation and drying processes using principal components analysis. *Food Res. Int.* **44**, 250–258.
- Rolland, F., Baena-Gonzalez, E. and Sheen, J. (2006) Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annu. Rev. Plant Biol.* **57**, 675–709.
- Saito, K. (2013) Phytochemical genomics—a new trend. *Curr. Opin. Plant Biol.* **16**, 373–380.
- Sanbongi, C., Osakabe, N., Natsume, M., Takizawa, T. and Gomi, S.O.T. (1998) Antioxidative polyphenols isolated from *Theobroma cacao*. *J. Agric. Food Chem.* **46**, 454–457.
- Sanchez-Rabaneda, F., Jauregui, O., Casals, I., Andres-Lacueva, C., Izquierdo-Pulido, M. and Lamuela-Raventos, R.M. (2003) Liquid chromatographic/electrospray ionization tandem mass spectrometric study of the phenolic composition of cocoa (*Theobroma cacao*). *J. Mass Spectrom.* **38**, 35–42.
- Scherling, C., Roscher, C., Giavalisco, P., Schulze, E.-D. and Weckwerth, W. (2010) Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *PLoS ONE* **5**, e12569.
- da Silva, F.L., Escribano-Bailon, M.T., Alonso, J.J.P., Rivas-Gonzalo, J.C. and Santos-Buelga, C. (2007) Anthocyanin pigments in strawberry. *LWT-Food Sci. Technol.* **40**, 374–382.
- Sreenivasulu, N., Usadel, B., Winter, A. et al. (2008) Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. *Plant Physiol.* **146**, 1738–1758.
- Stark, T. and Hofmann, T. (2005) Isolation, structure determination, synthesis, and sensory activity of *N*-phenylpropenoyl-L-amino acids from cocoa (*Theobroma cacao*). *J. Agric. Food Chem.* **53**, 5419–5428.
- Sumner, L.W., Amberg, A., Barrett, D. et al. (2007) Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, **3**, 211–221.
- Sun, X. and Weckwerth, W. (2012) COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics*, **8**, 81–93.
- Taji, T., Ohsumi, C., Luchi, S., Seki, M., Kasuga, M., Kobayashi, M., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J.* **29**, 417–426.
- Tegeder, M. and Rentsch, D. (2010) Uptake and partitioning of amino acids and peptides. *Mol. Plant*, **3**, 997–1011.
- Teye, E., Huang, X., Han, F. and Botchway, F. (2014) Discrimination of cocoa beans according to geographical origin by electronic tongue and multivariate algorithms. *Food Anal. Method.* **7**, 360–365.
- Tzin, V. and Galili, G. (2010) New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Mol. Plant*, **3**, 956–972.
- Urpi-Sarda, M., Llorach, R., Monagas, M., Khan, N., Rotches-Ribalta, M., Roura, E., Lamuela-Raventos, R.M., Estruch, R. and Andres-Lacueva, C. (2011) Effect of cocoa powder in the prevention of cardiovascular disease: Biological, consumption and inflammatory biomarkers. A metabolomic approach. *Rec. Adv. Pharm. Sci.* **37**, 121–132.
- Valledor, L., Escandón, M., Meijón, M., Nukarinen, E., Cañal, M.J. and Weckwerth, W. (2014) A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. *Plant J.* **79**, 173–180.
- Vizcaino, J.A., Deutsch, E.W., Wang, R. et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotech.* **32**, 223–226.
- Wang, L. and Ruan, Y.L. (2013) Regulation of cell division and expansion by sugar and auxin signaling. *Front. Plant Sci.* **4**, 163.
- Weckwerth, W. (2003) Metabolomics in systems biology. *Annu. Rev. Plant Biol.* **54**, 669–689.
- Weckwerth, W. (2010) Metabolomics: an integral technique in systems biology. *Bioanalysis*, **2**, 829–836.
- Weckwerth, W. (2011a) Green systems biology—from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J. Proteomics*, **75**, 284–305.
- Weckwerth, W. (2011b) Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. *Anal. Bioanal. Chem.* **400**, 1967–1978.
- Weckwerth, W., Wenzel, K. and Fiehn, O. (2004) Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, **4**, 78–83.
- Wienkoop, S., Morgenthal, K., Wolschin, F., Scholz, M., Selbig, J. and Weckwerth, W. (2008) Integration of metabolomic and proteomic phenotypes analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in *Arabidopsis thaliana*. *Mol. Cell Proteomics*, **7**, 1725–1736.
- Yoo, H., Widhalm, J.R., Qian, Y., Maeda, H., Cooper, B.R., Jannasch, A.S., Gonda, I., Lewinsohn, E., Rhodes, D. and Dudareva, N. (2013) An alternative pathway contributes to phenylalanine biosynthesis in plants via a cytosolic tyrosine: phenylpyruvate aminotransferase. *Nat. Commun.* **4**, 2833.
- Zhang, L., Tan, Q., Lee, R., Trethewy, A., Lee, Y.H. and Tegeder, M. (2010) Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in *Arabidopsis*. *Plant Cell*, **22**, 3603–3620.
- Zheng, X.Q., Koyama, Y., Nagai, C. and Ashihara, H. (2004) Biosynthesis, accumulation and degradation of theobromine in developing *Theobroma cacao* fruits. *J. Plant Physiol.* **161**, 363–369.
- Zybailov, B., Mosley, A.L., Sardu, M.E., Coleman, M.K., Florens, L. and Washburn, M.P. (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* **5**, 2339–2347.

2. Metabolome and proteome of developing lotus seed

Title: Metabolomic and proteomic profiles reveal the dynamics of primary metabolism during seed development of lotus (*Nelumbo nucifera*)

2.1 Preamble

Sacred lotus (*Nelumbo nucifera*) seeds are popular in Asia countries as snacks as well as traditional medicine. They are consumed either freshly or after desiccation. To figure out the proper harvesting time for different consumption purposes, central primary metabolism of developing lotus seeds was investigated on metabolome and proteome levels. Morphological and biochemical changes during developmental process of sacred lotus seeds were described in detail in the following publication.

2.2 Declaration of authorship

The results of central primary metabolism dynamics during sacred lotus seed development were published in *Frontiers in Plant Science* as an original research paper. This work is a cooperation with Prof. Dr. Pingfang Yang from Wuhan Botanical Garden. Jinlei Fu and I contributed equally to this work. I prepared samples for GC-TOF-MS measurement, analyzed metabolomics data, integrated metabolome and proteome data for exploration of central metabolism dynamics during the lotus seed filling period and wrote the manuscript.

2.3 Published manuscript



Metabolomic and Proteomic Profiles Reveal the Dynamics of Primary Metabolism during Seed Development of Lotus (*Nelumbo nucifera*)

Lei Wang^{1†}, Jinlei Fu^{2†}, Ming Li², Lena Fragner¹, Wolfram Weckwerth^{1,3*} and Pingfang Yang^{2,4*}

¹ Department of Ecogenomics and Systems Biology, University of Vienna, Vienna, Austria, ² Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China, ³ Vienna Metabolomics Center, University of Vienna, Vienna, Austria, ⁴ Sino-African Joint Research Center, Chinese Academy of Sciences, Wuhan, China

OPEN ACCESS

Edited by:

Dominique Job,
Centre National de la Recherche
Scientifique, France

Reviewed by:

Loïc Rajjou,
AgroParisTech, France
Tai Wang,
Institute of Botany, China

*Correspondence:

Pingfang Yang
yangpf@wbpcas.cn;
Wolfram Weckwerth
wolfram.weckwerth@univie.ac.at

[†] These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Plant Proteomics,
a section of the journal
Frontiers in Plant Science

Received: 30 March 2016

Accepted: 17 May 2016

Published: 07 June 2016

Citation:

Wang L, Fu J, Li M, Fragner L,
Weckwerth W and Yang P (2016)
Metabolomic and Proteomic Profiles
Reveal the Dynamics of Primary
Metabolism during Seed
Development of Lotus (*Nelumbo
nucifera*). *Front. Plant Sci.* 7:750.
doi: 10.3389/fpls.2016.00750

Sacred lotus (*Nelumbo nucifera*) belongs to the Nelumbonaceae family. Its seeds are widely consumed in Asian countries as snacks or even medicine. Besides the market value, lotus seed also plays a crucial role in the lotus life cycle. Consequently, it is essential to gain a comprehensive understanding of the development of lotus seed. During its development, lotus seed undergoes cell division, expansion, reserve accumulation, desiccation, and maturation phases. We observed morphological and biochemical changes from 10 to 25 days after pollination (DAP) which corresponded to the reserve synthesis and accumulation phase. The volume of the seed expanded until 20 DAP with the color of the seed coat changing from yellow-green to dark green and gradually fading again. Starch and protein rapidly accumulated from 15 to 20 DAP. To further reveal metabolic adaptation, primary metabolites and proteins profiles were obtained using mass spectrometry based platforms. Metabolites and enzymes involved in sugar metabolism, glycolysis, TCA cycle and amino acid metabolism showed sequential dynamics enabling the clear separation of the different metabolic states during lotus seed development. The integration of the data revealed a highly significant metabolic switch at 15 DAP going through a transition of metabolically highly active tissue to the preparation of storage tissue. The results provide a reference data set for the evaluation of primary metabolism during lotus seed development.

Keywords: mass spectrometry, primary metabolism, GC-TOF-MS, LC-Orbitrap-MS, seed development, *Nelumbo nucifera*

INTRODUCTION

Sacred lotus (*Nelumbo nucifera*) belongs to the small plant family of Nelumbonaceae (The Angiosperm Phylogeny Group, 2009), which is found at a basal position of eudicot plants and contains only one genus (*Nelumbo*) with only two species, sacred lotus and American lotus (*Nelumbo lutea*) (Pan et al., 2010). *N. nucifera* is distributed in Asia and Northern Australia,

whereas *N. lutea* inhabits the eastern areas of North America and the northern area of South America. *N. nucifera* is a diploid ($2n = 16$) with an estimated 929 Mb genome (Ming et al., 2013). In Asia, sacred lotus is an economically important crop with a very long cultivation history. It is widely used as food, medicine, and ornamental plant. Based on its agricultural and horticultural utilization, it can be divided into three groups, namely flower, seed, and rhizome lotus.

The seed plays an important role in plants life cycle. It stores genetic information and nutrients and also develops protective and defense strategies to guarantee the reproduction of the next generation. Sacred lotus seeds are the products of its sexual reproduction, and primarily used to reproduce and distribute its progeny. It has been reported that lotus seeds have an extreme longevity, and can remain viable for about 1300 years (Shen-Miller, 2002). The stored carbohydrates, proteins, lipids and other compounds not only provide energy for seed germination but also for human and other animals in the form of food. The seeds of sacred lotus are widely consumed in Asian countries as snacks or in some cultures for medicinal purposes (Yen et al., 2005, 2006). Sacred lotus blossoms and sets seeds in the hot summer days, which makes its seed development responsive to high temperatures. Based on this economic importance, it is important to study its seed formation and development.

Carbohydrates, proteins and oils are three major reserves accumulating in plant seeds (Weber et al., 2005). Compared to other plants, sacred lotus seeds mainly accumulate starch, which accounts for about 60% of its total dry weight. It also accumulates about 8% protein in immature seeds and as high as 24% in the mature desiccated seeds (Zheng et al., 2003; Bhat and Sridhar, 2008). In contrast to cereals, starch is mainly synthesized and accumulated in cotyledons in sacred lotus. Previous studies on lotus seed was mainly focused on the identification of its nutritional constituents and medicinal components (Yen et al., 2005, 2006; Mukherjee et al., 2010). However, there is thus far no report exploring the regulation of biosynthesis and accumulation of these reserves. Because the accumulation of reserves is important in both nutrition and in an economical sense, studies on seed filling have been widely conducted in various of crops, including rice (Xu et al., 2008), barley (Finnie et al., 2002), wheat (Laino et al., 2010), maize (Mechin et al., 2007), soybean (Agrawal et al., 2008), oilseeds (Hajduch et al., 2005, 2006), and *Medicago truncatula* (Gallardo et al., 2003, 2007; Repetto et al., 2008). All these studies revealed that the seeds experienced dramatic changes in morphology and metabolism. In contrast to other crops, sacred lotus seeds can be consumed either freshly or in a mature desiccated form. As fresh food, sacred lotus seed is sweet, which indicates it contains high contents of soluble sugars. After this, it enters the filling stage, during which starch is quickly synthesized and accumulated. When consumed maturely, the seeds are desiccated and cotyledons are filled with starch. Understanding when this transition from the stage suitable for fresh consumption to the filling stage and how this transition happens, are very important questions in sacred lotus seed production. With the development of sequencing technologies, great success has been achieved in genome sequencing. Recently, the sacred lotus whole genome was sequenced (Ming et al., 2013),

which has provided ample information for the analysis of the transcriptome and proteome of this species (Deng et al., 2015; Yang et al., 2015a,b; Liu et al., 2016). Moro et al. (2015b) profiled the endosperm proteome of mature lotus seed and they also compared the proteome profiles of the immature and mature seed endosperm. However, these data did not provide a clear answer on the transition from the fresh consumption stage to the reserves filling stage. With this in mind, we combined a label-free quantitative proteomics and gas-chromatography-mass spectrometry (GC-MS) based metabolomic studies on the developing sacred lotus seeds in an effort to address this question. The first objective is to identify key enzymes important for sacred lotus seed development and reserve filling. The second is to reveal the metabolic dynamics during seed development, and the third is to shed light on the mechanisms underlying the switches in metabolism during sacred seed development and maturation.

MATERIALS AND METHODS

Plant Material

The sacred lotus (*N. nucifera*) cultivar “Baijianlian” were cultivated in the experimental field of Wuhan Botanical Garden (30°32′45″N 114°24′52″E), Chinese Academy of Sciences in late April. In July, most of the sacred lotus plants blossomed. The healthy plants were selected for experiments. In the first day of blossoming, the flowers were manually pollinated and capped with labeled paper bags. Seeds at different days-after-pollination (DAP) were harvested, and then either fixed in formalin-acetic acid-alcohol (FAA) solution for microscopy analysis or frozen with liquid nitrogen and stored at -80°C for metabolite and protein extraction.

Histological Procedures

The harvested samples were fixed in FAA solution for at least 24 h, and then vacuum-dried for 1 h. Thereafter, the samples were subjected to sequential dehydration with 30–100% ethanol, and then embedded into paraffin. The embedded samples were cut into 6–10 μm slices using a LEICA2150 rotary microtome (Leica, German). After dewaxing, the slices were subjected to Periodic Acid Schiff (PAS) solution and Coomassie Brilliant Blue (CBB) staining for the observation of starch and protein accumulation, respectively. An Olympus BX-61 upright metallurgical microscope (Olympus, Japan) was used for the observation.

Metabolite Extraction and Analysis

The harvested lotus seeds were ground to powder in liquid nitrogen with a mortar and pestle. 100 mg material was extracted with 1.4 ml precooled methanol. The extraction mixture was homogenized by vortexing for 10 s. 60 μl of ribitol (0.2 mg ml^{-1}) was added to the mixture as internal standard followed by 10 s of vortexing. This solution was ultrasonicated for 10 min, followed by centrifugation at 11 000 g for 10 min. The supernatant was transferred to a new tube and mixed with 750 μl of precooled chloroform and 1.5 ml of precooled water. The mixture was then centrifuged at 2200 g for 15 min. 150 μl of extraction solution

from upper phase was dried under vacuum and stored at -80°C until derivatization.

The extract of lotus seed was methoxyaminated and silylated according to the derivatization method from Weckwerth et al. (2004) with minor modifications. The dried extract was dissolved in 20 μl of methoxyamine hydrochloride pyridine solution (40 mg ml^{-1}) and incubated at 30°C with vigorous shaking for 90 min. Before silylation, 30 μl of even-numbered alkanes (C10–C40, 50 mg ml^{-1} each, Fluka, Germany) was spiked in 1 ml of *N*-methyl-*N*-(trimethylsilyl) trifluoroacetamide (MSTFA, MACHEREY-NAGEL, Germany) solution. Then 80 μl of this MSTFA-Alkane solution was added to the sample followed by 30 min incubation at 37°C with vigorous shaking. The derivatized sample was centrifuged at 20 000 g for 8 min. The supernatant was then transferred to vials for measurement.

Samples were measured with an Agilent 6890 gas chromatography coupled to a LECO Pegasus[®] 4D GC $\times$ GC-TOF spectrometry (GC-TOF-MS). Instrument parameter settings were consistent with a previous report (Doerfler et al., 2013). Each sample was injected under both splitless and split 25 times mode for better quantification of candidates with a wide capacity range.

The obtained raw files were deconvoluted with LECO Chroma TOF[®]. The retention times (RTs) of alkanes were applied to calibrate the RTs of candidates. Candidates were manually annotated by comparing their RTs and mass spectra to those of standards in GMD database (Kopka et al., 2005) with a minimum match factor of 700. Peak areas of annotated candidates corresponding to specific masses were integrated and used for relative quantification. The peak areas of the same metabolite with different derived groups were merged and normalized to internal standard ribitol.

Protein Extraction and Analysis

Proteins were extracted using the phenol method according to previously description (Liu et al., 2013). Briefly, the samples were ground to power with liquid nitrogen, homogenized in pre-cooled homogenization buffer (20 mM Tris/HCl (pH 7.5), 250 mM sucrose, 10 mM ethylene glycol-bis(b-aminoethylether)-*N,N,N',N'*-tetraacetic acid (EGTA), 1 mM PMSF, 1 mM DTT, and 1% Triton X-100), followed by centrifugation at 12 000 g for 20 min at 4°C . Supernatants were collected, and mixed with equal volume of Tris-phenol (pH > 8.0) by vortexing, and centrifuged again. The top phenol phase was carefully transferred to a new tube, and mixed with 5 volumes of 0.1 M methanolic ammonium acetate in 100% methanol, then incubated overnight at -20°C . The precipitated proteins were washed one time with 0.1 M methanolic ammonium acetate and two times with acetone. The protein pellets were vacuum-dried and then stored at -80°C for further use.

For one-dimensional electrophoresis (DE), proteins were dissolved in lysis buffer (7 M urea, 2 M thiourea, 4% w/v CHAPS, and 65 mM DTT) and quantified using the Bradford method (Bradford, 1976). Twenty-five micro gram of total proteins were loaded onto 12.5% sodium dodecyl sulfate–polyacrylamide gel (SDS-PAGE) and run at 50 V until bromophenol blue line enter into the resolving gel for about 1 cm. After running, the gels

were stained with (CBB R-250) solution, and then destained until protein bands were clearly visualized. The gel was cut into small slices, and subjected to in-gel trypsin digestion (Liu et al., 2016).

For in-solution digestion, 100 μg of protein was reduced with 5 mM dithiothreitol (DTT) at 37°C for 45 min then alkylated with 10 mM iodoacetamide (IAA) at 23°C in dark for 60 min. Finally the reaction was finished with addition of 5 mM of DTT (23°C , dark, 15 min). Endoproteinase LysC and trypsin were applied for digestion based on the protocol reported before (Hoehenwarter et al., 2008).

Peptides from each sample were thoroughly dissolved in 4% acetonitrile and 0.1% formic acid, and desalted. Afterward, 10 μg peptides sample was subjected to analysis using nanoHPLC coupled to LTQ-Orbitrap-MS according to previously report (Liu et al., 2016). MS analysis was performed on an Orbitrap LTQ XL mass spectrometer (Thermo, Germany) with a controlled flow rate of 500 nl min^{-1} . Specific tune settings were 1.8 kV for spray voltage and 180°C for temperature of the heated transfer capillary. MS raw data were searched using the SEQUEST algorithm from Proteome Discoverer version 1.3 (Thermo, Germany). The filter criteria were set as follows. (I) 1% false discovery rate (FDR); (II) a variable modification of acetylation of N-terminus and oxidation of methionine; (III) mass tolerance of 10 ppm for parent ion and 0.8 Da for fragment ion.

The MS data were searched against the newly annotated genome database of lotus¹ to identify proteins. The assembled genome of lotus was 804 Mb, representing 86.5% of the estimated lotus genome. Following repeat-masking and annotation, 26,685 unigenes were inferred, 82% of them have similarity to proteins in SwissProt (Ming et al., 2013). The searching parameters are as follows: (i) medium or high peptide confidence; (ii) minimum Xcorr of 1 for +1, 2 for +2 and 3 for +3 charged peptides, etc. Subsequently, a label-free approach based on spectral counts (SpCs) was used to quantify the identified proteins followed by a normalized spectral abundance factor (NSAF) normalization strategy (Zybailov et al., 2006). The NSAF was calculated based on the formula: $\text{NSAF}_i = (\text{SpC}_i/L_i) / \sum_{i=1}^N (\text{SpC}_i/L_i)$. The NSAF for a protein is the number of spectral counts (SpCs) divided by protein length (L) and then divided by the sum of SpC/L. The obtained raw files and sequence information of the identified proteins were submitted to the public repository ProteomeXchange (Vizcaino et al., 2014) with the dataset identifier PXD003768² (Username: reviewer72127@ebi.ac.uk, Password: LHImlH6). Furthermore all lotus proteomic data, annotated peptide spectra and experimental meta information is stored at the proteomics database PROMEX³ (Hummel et al., 2007; Wienkoop et al., 2012).

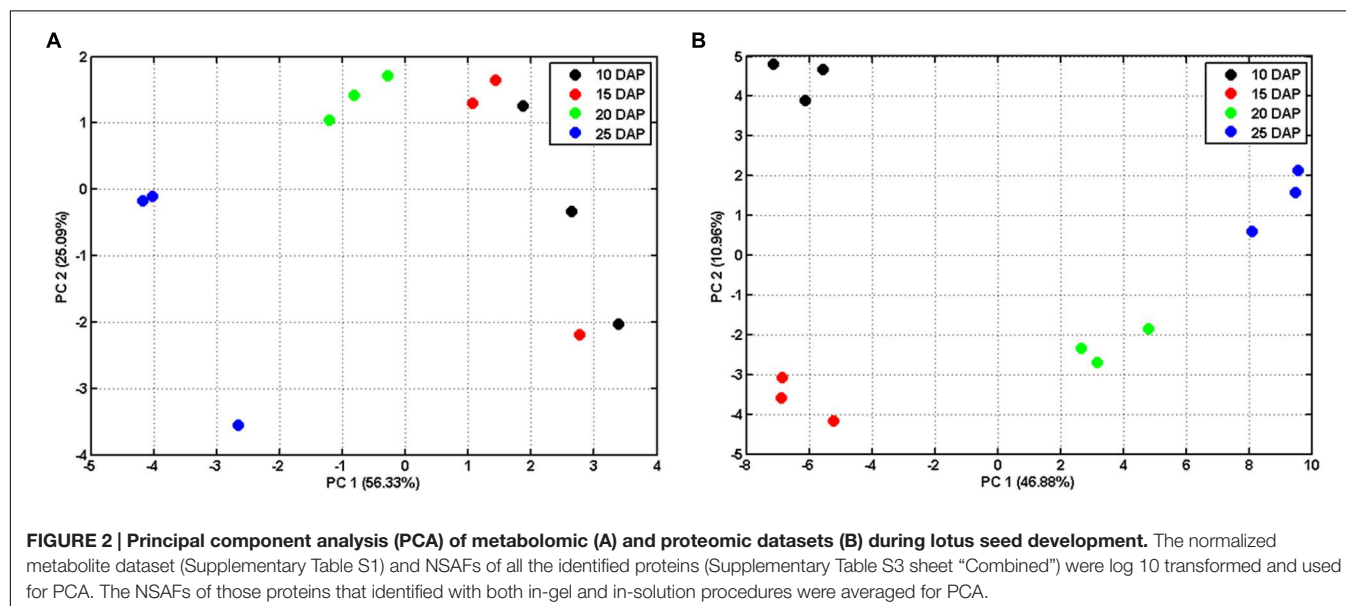
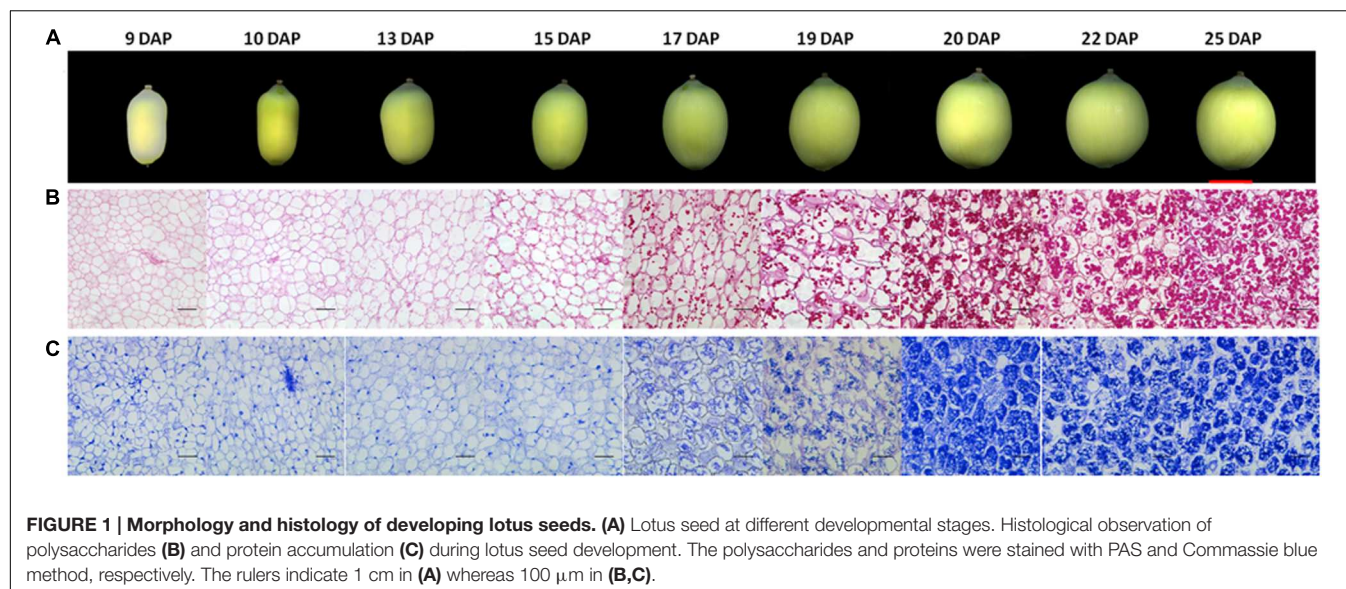
Multivariate Statistical Analysis

Principal component analysis (PCA) and hierarchical clustering analysis were performed with the statistical tool box COVAIN (Sun and Weckwerth, 2012). Analysis of variance (one-way ANOVA) was performed with SPSS Software Package and the

¹<http://lotus-db.wbgcas.cn>

²<http://www.proteomexchange.org>

³<http://promex.pph.univie.ac.at/promex/>



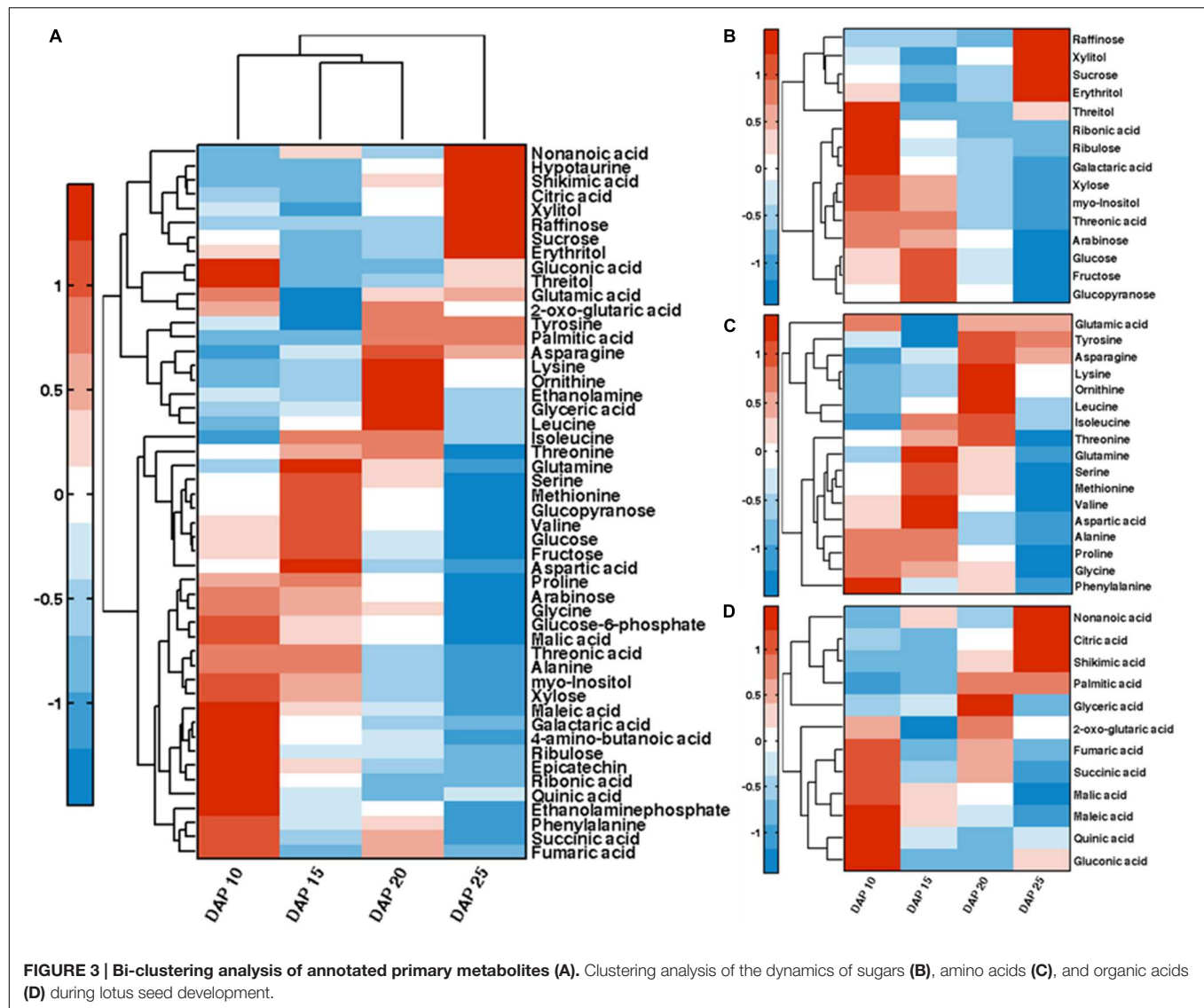
significant levels were indicated with lower case letters according to Duncan's test ($p < 0.05$). Duncan's test was done in SPSS with an option in *post hoc* window named "Duncan" at 0.05 level. Venn diagram was performed with online free software Venny 2.1⁴.

RESULTS AND DISCUSSION

Morphological Changes during Lotus Seed Development

It has been reported that the development of lotus seed could be divided into four phases, phase I cell division and organ formation (1–3 days after pollination, DAP), phase II cell enlargement (4–9 DAP), phase III reserves synthesis and

accumulation (10–25 DAP), phase IV desiccation and maturation (Tang et al., 1989). Based on the objectives of this study, we focused on phase III. Sacred lotus seeds at different time points of this phase were harvested with three biological replicates. For comparison, seeds at 9 DAP were also collected. The size of lotus seeds expanded gradually until 20 DAP, and then slowed down (Figure 1A). The seed coat color gradually changed from yellow-green to dark-green until 17 DAP, and then gradually faded (Figure 1A). To further reveal the accumulation of different reserves, light microscope observations were applied on the cotyledons of these samples. No starch accumulation was observed before 10 DAP, whereas starch accumulated rapidly after 15 DAP, and the accumulation rate slowed after 20 DAP (Figure 1B). Proteins accumulated in a similar pattern (Figure 1C). Interestingly, clear nuclei could be observed in the



cells before 15 DAP, while this signal disappeared after 15 DAP, which might be correlated with the accumulation of starch and proteins. Based on these data, we selected the cotyledons at 10, 15, 20, and 25 DAP for further metabolomic and proteomic studies in order to explore the metabolic switch during this phase.

Metabolomic Profile of Developing Lotus Seed

In order to explore the dynamic changes of primary metabolites such as sugars and organic acids, we analyzed metabolites of developing lotus seed. The extracted metabolites were derivatized before measuring by gas chromatography-time of flight-mass spectrometry (GC-TOF-MS). In total, 95 candidates were annotated and quantified based on their spectra and retention times (Supplementary Table S1). Bi-cluster analysis

of the metabolites presented distinctive metabolic profiles of lotus seed at different developmental stages with the changing patterns of the metabolites shown in a hierarchical heat map (Figure 3A). Samples at 25 DAP were most distinctive compared to samples at other stages. PCA plots of the metabolite data show the trajectory of lotus seed development (Figure 2A). Samples of 10 and 15 DAP are separated from those of 20 and 25 DAP on PC 1, which accounts for over 50% of the separation. Loading values in Supplementary Table S2 indicate sugars e.g. glucose, fructose, xylose, *myo*-inositol contribute greatly to this separation. Fructose and glucose, the main sugars in lotus seed, with peak accumulation at 15 DAP, notably contributed to the separation of developmental samples at the first three major PCs. Other sugars, e.g., ribulose, *myo*-inositol, glucose-6-phosphate, arabinose, and xylose were most abundant at 10 DAP and decreased during lotus seed development. Raffinose, xylitol, sucrose, and erythritol greatly accumulated in seeds at 25 DAP (Figure 3B).

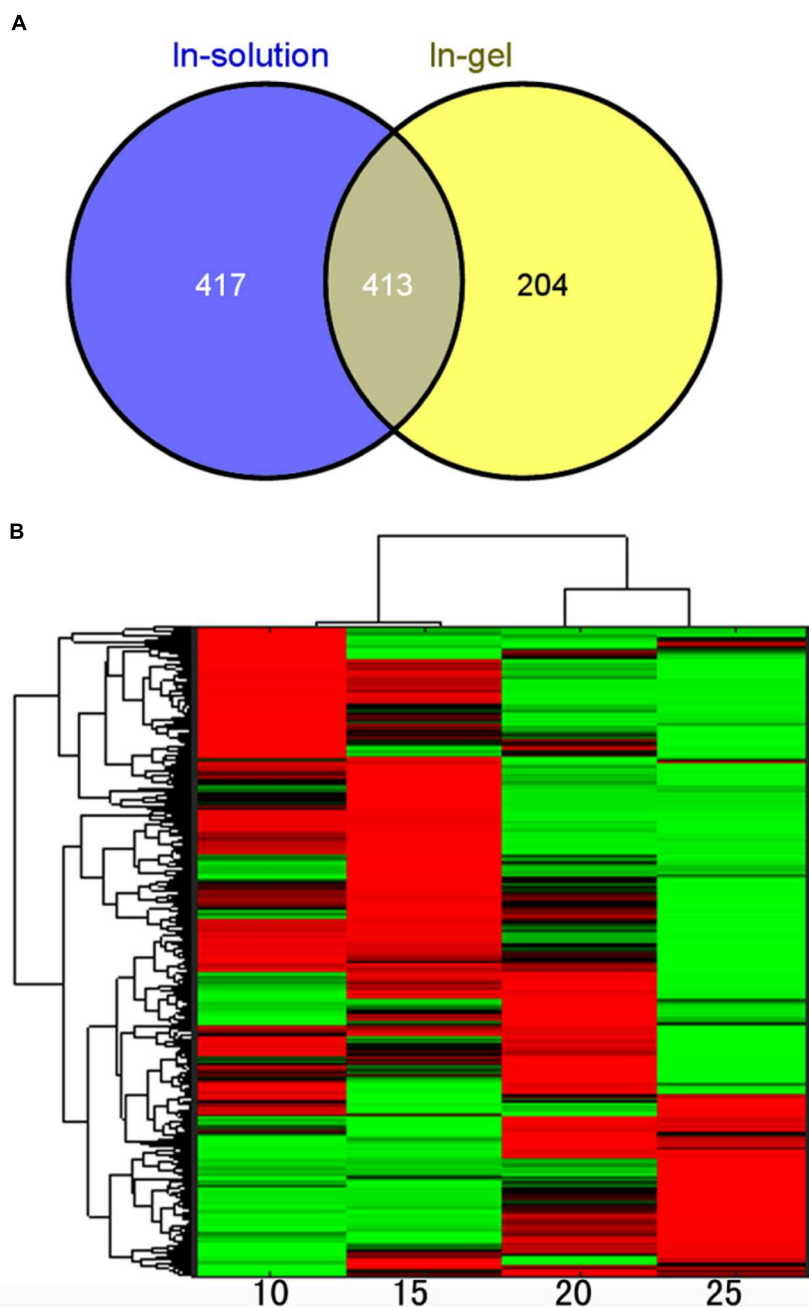
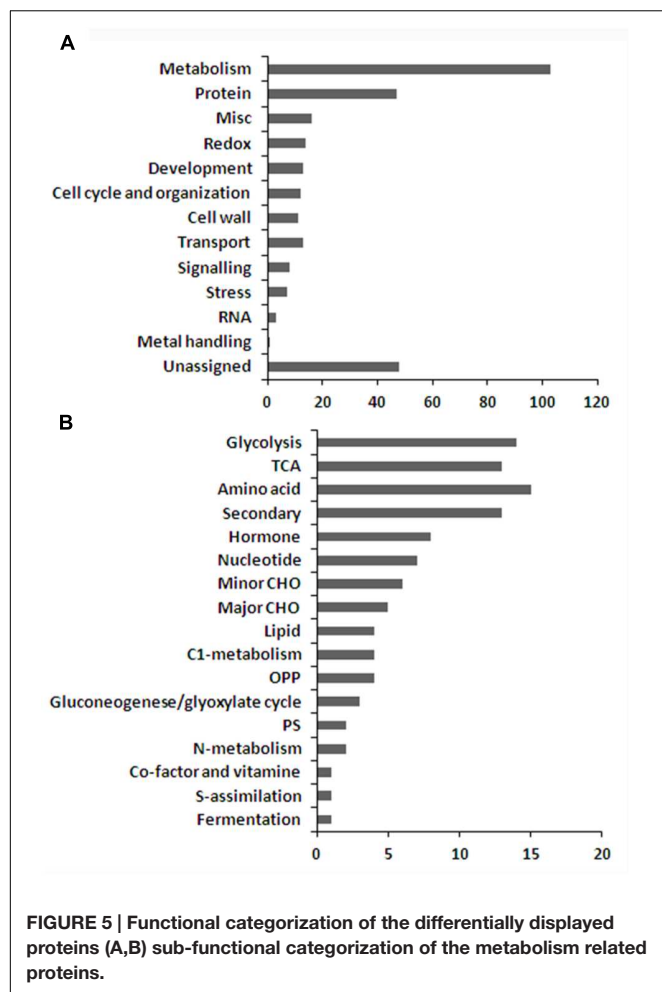


FIGURE 4 | Quantitative analysis of the proteomic data. (A) Venn diagram showing the overlap of proteins identified in samples using in-gel and in-solution digestion methods; **(B)** Clustering analysis of the identified proteins.

Amino acids like alanine, glutamine and lysine were also essential contributors to the characterization of developmental stages. Most of the amino acids detected showed highest level at 15 DAP or 20 DAP except phenylalanine, glycine, and glutamic acid, which were highest at 10 DAP (**Figure 3C**). 10 out of 17 detected amino acids showed lowest level at 25 DAP, which may have resulted from the incorporation of amino acids into storage protein synthesis. This is consistent with the histological

observations (**Figure 1C**) of considerable protein accumulation at 25 DAP.

Organic acids are another class of metabolites involved in essential aspects of metabolism in plants. The predominant organic acid in lotus seed is malic acid which decreased in phase III. 2-oxo-glutarate (also known as α -ketoglutarate), succinate and fumarate involved in TCA cycle presented similar fluctuating pattern. These three metabolites decreased from 10 to 15 DAP



followed by an increase and another decrease during lotus seed development (Figure 3D). Citric acid was the only TCA cycle intermediate that significantly increased from 10 to 25 DAP. Shikimic acid, the precursor for aromatic amino acids also increased notably at this developmental phase.

Proteomic Profile of Developing Lotus Seed

To further understand the molecular mechanisms underlying the dynamic changes of metabolites and the shift from physiological active status to reserves accumulation, comparative proteomic analysis was conducted on the developing lotus seeds at 10, 15, 20 and 25 DAP. In order to obtain as much information as possible, an in-gel digestion via LC-MS/MS strategy was conducted in parallel with an in-solution digestion strategy. Based on the criteria described in the Methods, a total of 1140 and 1610 proteins were identified through in-gel (Valledor and Weckwerth, 2014) and in-solution digestion techniques, respectively, among which 617 and 830 proteins were quantified, respectively, (Supplementary Table S3). The number of identified proteins is much more than that from the other groups working with lotus (Moro et al., 2015a,b). The reasons might be as

follows. Firstly, samples from more stages were used in this study. Secondly, the MS data were searched against a self-generated lotus protein database in this study rather than that available in NCBI database, which will increase the rate of successful protein identifications. However, there are 36 proteins that were previously identified by 1-D and 10 by 2-D (Moro et al., 2015a,b) methods that were not identified in this study (Supplementary Table S4). This might because the lotus genome is not well annotated, which will result in the loss of some information for protein identifications. Furthermore, we used the seeds including embryos and endosperms, which may result in the loss of some embryo specific proteins. Among all the annotated protein candidates, 413 proteins were commonly identified and quantified in both strategies (Figure 4A). To evaluate the reproducibility of biological replicates, PCA was performed as previously reported (Liu et al., 2016) with all the identified proteins. The PCA plot shows clear separation between samples from different developmental stages (Figure 2B). Specifically, samples from 10 and 15 DAP were not separated by PC1 which accounts for 46.88%, but by PC2 accounting for 10.96% (Figure 2B), which is similar to the PCA result from the metabolomic data. Interestingly, clustering analysis based on the expressional patterns of all the quantified proteins also shows that samples of 10 and 15 DAP were separated from those of 20 and 25 DAP (Figure 4B). These data indicate that seeds are metabolically active at the first two stages, and enter into reserve accumulation at the latter two stages.

Quantitative analysis showed that 296 proteins were significantly changed during seed development (Supplementary Table S5). Except for 48 function-not-assigned proteins, the rest of the proteins were affiliated to 12 functional groups based on MapMan descriptions (Figure 5A). 103 out of 296 significantly changed protein candidates, which accounts for about 34.8%, are metabolism related proteins indicating that the seeds are subject to dramatic metabolic changes during these developmental stages. These proteins could be further classified into 17 sub-groups (Figure 5B), including glycolysis, TCA, OPP, gluconeogenesis/glyoxylate cycle, photosynthesis, S-assimilation, fermentation, and metabolism of amino acids, secondary metabolites, hormone, nucleotides, minor carbohydrates (CHO), major CHO, lipids, C1-, nitrogen, co-factor, and vitamin.

Integration of Metabolomic and Proteomic Profiles to Reveal Primary Metabolism Processes during Lotus Seed Development

The seed of lotus, as a sink organ, plays a central role in the lifecycle of lotus. Primary metabolism plays an essential role in plant metabolism and is vital for the plant to survive and develop. Thus, the understanding of primary metabolism of the seed is crucial. The morphological description of the processes alone is not sufficient to interpret such a complex process. Here we explored metabolic regulation and adjustments during this process at the molecular level by integrating metabolomic and proteomic profiles of the developing seeds of lotus (Figure 6). Central primary metabolism pathways

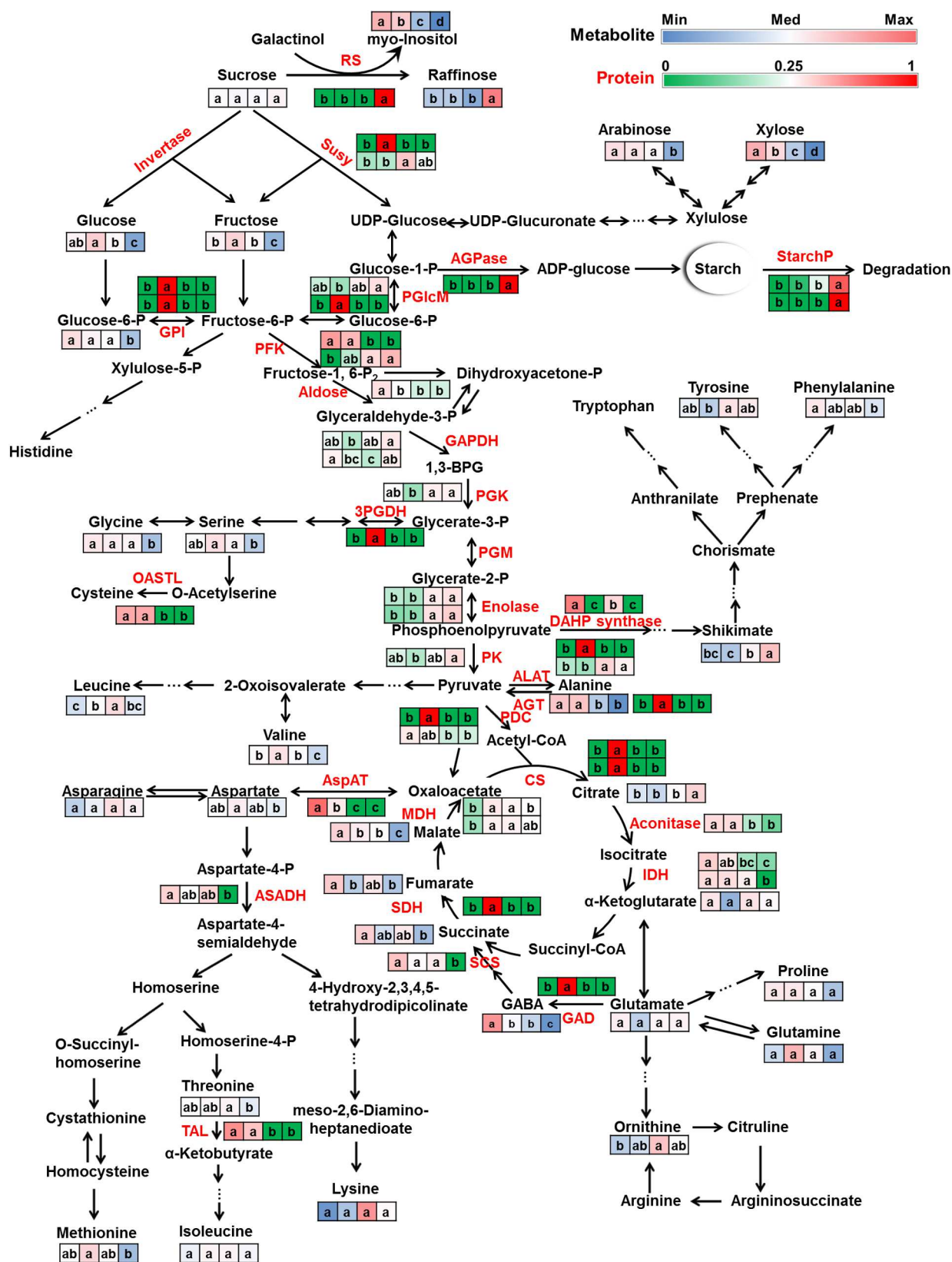
**FIGURE 6 | Continued**

FIGURE 6 | Continued

Primary metabolism dynamics during lotus seed development. Metabolites (black) and proteins (red) involved in sugar metabolism, glycolysis, TCA cycle and amino acid metabolism were mapped on their metabolism pathways. The relative ratios of normalized metabolite peak areas were colored with blue-white-red color bar corresponding to the values from the minimum (Min) to median (Med) to maximum (Max). The relative ratios of the protein NSAFs were colored with green-white-red color bar corresponding to the ratios from 0 to 0.25 to 1. All the protein candidates that annotated as the same enzyme were list here in consecutive rows. Four squares from left to right indicate samples of 10, 15, 20, and 25 DAP, respectively. Lowercase characters in the squares indicate significant levels according to Duncan's test ($p < 0.05$). Abbreviations: RS, raffinose synthase; Susy, sucrose synthase; AGPase, ADP glucose pyrophosphorylase; StarchP, starch phosphorylase; PGlcM, phosphoglucomutase (plastid); GPI, glucose-6-phosphate isomerase (plastid); PFK, phosphofructokinase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; 1,3-BPG, 1,3-Bisphosphoglyceric acid; PGK, 3-phosphoglycerate kinase; PGM, phosphoglycerate mutase; PK, pyruvate kinase; PDC, pyruvate dehydrogenase complex; CS, citrate synthase; IDH, isocitrate dehydrogenase; SCS, succinyl coenzyme A synthetase; SDH, succinate dehydrogenase; MDH, malate dehydrogenase; GAD, glutamic acid decarboxylase; GABA, 4-aminobutanoic acid (γ -Aminobutyric acid); AspAT, aspartate aminotransferase; ALAT, alanine aminotransferase; AGT, alanine-glyoxylate aminotransferase; ASADH, aspartate semialdehyde dehydrogenase; TAL, threonine ammonia-lyase; 3PGDH, 3-phosphoglycerate dehydrogenase; OASTL, O-acetylserine(thiol)lyase; DAHP synthase, 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase.

including sugar metabolism, glycolysis, TCA cycle, and amino acids metabolism are discussed below.

Sugar Metabolism

The predominant sugars in lotus seed include sucrose, glucose, and fructose. Sucrose is mainly produced in leaves via photosynthesis and then transported to seeds and other organs for energy and carbon skeleton production. At early developmental stages, the seed embryo undergoes fast cell division and expansion. During this period, the transported sucrose is predominantly hydrolyzed to glucose and fructose by sucrose synthase and invertase which results in an increase of these hexoses. The incorporation of these hexoses into glycolysis and TCA cycle provides energy and precursors for seed growth. Since 20 DAP, the volume expansion of the seed slowed down and the seed moved into its maturation phase. At this stage, sucrose metabolism, instead of mainly hydrolyzing to hexoses, switched to raffinose synthesis and resulted in significant accumulation of raffinose in mature seeds as well as significant reductions of glucose and fructose. The accumulated raffinose serves as an energy and carbon storage reserve as well as protectants (Peterbauer and Richter, 2001; Karner et al., 2004) for further germination and against stresses. Enzymes involved in these processes, e.g., sucrose synthase (Susy), raffinose synthase (RS) showed consistent changing patterns with these metabolites. Susy, which catalyzes the reversible conversion of sucrose to fructose and UDP-glucose, showed highest level at 15 and 20 DAP while RS was highest in mature seeds (Figure 6). Invertase, which was reported to be crucial for carbohydrate metabolism during seed development (Weber et al., 1997; Sturm, 1999) was not detected. Sucrose is either converted to hexoses or to larger polymers such as raffinose or starch. ADP glucose pyrophosphorylase (AGPase), the key enzyme in the starch biosynthetic pathway (Sakulsingharoj et al., 2004; Weigelt et al., 2009), was highly abundant at 25 DAP (Figure 6). Previous studies indicated elevated cytoplasmic AGPase activity resulted in increased flux of sucrose into starch (Sakulsingharoj et al., 2004) whereas AGPase repression stimulated glycolytic and TCA cycle metabolism (Weigelt et al., 2009). In the present study, starch accumulation was accompanied with a high expression of AGPase, which is consistent with these reports. Starch phosphorylases (StarchP), which was reported to be associated with initiating seed reserve starch accumulation in

endosperm (plastidic StarchP) and starch degradation (cytosolic StarchP) during germination (Schupp and Ziegler, 2004), also significantly accumulated in mature lotus seed. During seed development, sugar composition and concentration highly varied to accomplish different steady states of metabolism. Such sugar partitioning was also observed in other developing seeds, e.g., *Arabidopsis* (Fait et al., 2006), barley (LaBerge et al., 1973), and wheat (Schupp and Ziegler, 2004). From 10 to 25 DAP, sucrose content slightly increased. The relatively constant sucrose level indicated a balance between sucrose import and consumption.

Glycolysis

Glucose and fructose further metabolize along the glycolytic pathway to produce energy (ATP), reductant (NADH) and pyruvate production. The consumption together with the aforementioned reduced production of these hexoses caused a decrease of glucose and fructose after 15 DAP. In the present study, enzymes involved in seven steps, i.e., glucose-6-phosphate isomerase (GPI), phosphofructokinase (PFK), aldose, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), 3-phosphoglycerate kinase (PGK), enolase, and pyruvate kinase (PK) changed significantly (Figure 6) indicating momentous adjustment of glycolytic metabolism during lotus seed development. These enzymes did not change with identical patterns, e.g., aldose was highest at 10 DAP, GPI 15 DAP whereas PGK, enolase and PK accumulated at 20 DAP and/or 25 DAP. One of the reasons might be related with the “bottom up” regulation of plant glycolysis (Plaxton, 1996), since glycolysis also plays a biosynthetic role in plants, especially in actively growing tissues. The intermediates generated during glycolysis are precursors for amino acids, fatty acids (Qi et al., 1995; Schwender et al., 2015), and other secondary metabolites. The incorporation of the intermediates to other anabolic pathways caused the fluctuation of these substances and a feedback to the system. Enzyme accumulation then adjusted based on these feedback signals and caused various changing patterns of enzyme capacities. Another point to be considered is that posttranscriptional modification can also regulate metabolism by enhancing or attenuating enzyme activity (Daran-Lapujade et al., 2007; Oliveira and Sauer, 2012; Tripodi et al., 2015). This means enzyme capacity is not a unique parameter that affects metabolic

regulation. In future studies, enzyme activity information should be complemented with the protein abundance data.

TCA Cycle

The TCA cycle operates as a bridge that connects carbohydrate, amino acid, lipid, and protein metabolism and also as an engine that generates energy and reductive power to drive metabolism, especially in sink organs. During lotus seed development, the intermediates were most abundance at 10 DAP and lowest in mature seed, except citrate which increased during seed ripening. Enzymes involved in the TCA cycle showed their lowest level in mature seed, except malate dehydrogenase (MDH) which was lowest at 10 DAP but also decreased from 15 to 25 DAP. The reduction of metabolic capacity of the TCA cycle at both the metabolic and proteomic levels indicated a decrease in energy, reducing power, and carbon skeleton generation during lotus seed development. This reduction in TCA cycle flux during seed development was also observed in developing *Arabidopsis* seed (Fait et al., 2006). One reason could be the shortage of oxygen, which is required by respiration to generate ATP and reducing equivalents by consuming ADP and reducing agents (Fait et al., 2006; Araujo et al., 2012). Another reason might be the decreasing requirement of energy and precursors from the TCA cycle as the seed at 25 DAP accomplished storage nutrition accumulation and prepared to enter the dormancy stage.

Amino Acid Metabolism

Amino acid metabolism is closely connected with the glycolytic pathway and the TCA cycle. On the one hand, intermediates generated along glycolysis and the TCA cycle can be precursors for amino acid synthesis. Nitrogen assimilation commonly happens in leaves, the biosynthesized amino acids are transported to sink organs, e.g., seed. Amino acids further serve as basic building blocks for protein synthesis and precursors for lipids and secondary metabolites, i.e., flavonoids, terpenoids, and alkaloid synthesis. At 25 DAP, the reduction of most of the amino acids might due to the incorporation of the amino acids into storage protein synthesis. On the other hand, amino acids can be catabolized and incorporate into glycolysis and TCA cycle for energy status maintenance (Angelovici et al., 2011; Galili et al., 2014; Hildebrandt et al., 2015). Lysine, a member of aspartate family, was reported to negatively correlate with TCA cycle activity in developing *Arabidopsis* seed which might cause energy shortage for seed development and germination in high lysine mutants (Angelovici et al., 2011). Junker and coworkers observed an redirection of fluxes leading into and out of the TCA cycle when cultivating *Brassica napus* embryos on a medium with ammonium nitrate as inorganic nitrogen source comparing with those with glutamine and alanine as organic nitrogen source (Junker et al., 2007). Their results indicated the effect of amino acids on metabolic flux. In summary, amino acid metabolism is affected and/or regulated by diverse factors, e.g., transportation, synthesis, and catabolism, together with interactions with other metabolic pathways.

CONCLUSION

Previous studies divided the development of lotus seed into four phases with phase III (10–25 DAP) being defined as the stage of reserve synthesis and accumulation (Tang et al., 1989). Based on our microscopic observations, there was a very clear switch in the metabolic status in lotus seed cotyledons during this phase. As mentioned above, sacred lotus seeds can be consumed either freshly or in a mature desiccated form, which is determined by the transition from the stage containing a high content of soluble sugars to a high starch content and desiccated stage. To know when this transition occurs is important for lotus seed production, and this information can provide guidelines, especially for fresh seed consumption. Integration of metabolome and proteome dynamics indicated that lotus seed was still metabolically from 10 to 20 DAP. Peak accumulation of soluble sugars at 15 DAP together with the significant accumulation of starch at 20 DAP suggests lotus seeds at the time of 15–20 DAP with a relatively high sweetness are most suitable for fresh consumption. Central primary metabolism switched from an energetic mode to a storage mode with an accumulation of starch and protein at 15 to 17 DAP. After 20 DAP, seed volume slowed and lotus seeds reached their final size. Soluble sugars decreased whereas starch and protein accumulated at the end of phase III when lotus seed turned to a storage tissue. The mature lotus seeds with high content of starch and protein are more suitable for drying and snack production or to be used in some other cuisine. This study provides solid evidence outlining the shift in metabolic pathways during lotus seed filling, and also clear clues for the harvesting of sacred lotus seeds as a fleshy fruit.

AUTHOR CONTRIBUTIONS

PY and WW conceived and designed the experiments. LW, JF, LF, and PY performed the experiments. LW, PY, and ML analyzed the data. WW and PY provided the reagents, materials, and analytical tools. LW and PY wrote the manuscript. All the authors approved the final manuscript.

FUNDING

This work was funded by The Knowledge Innovation Project of Chinese Academy of Sciences (Y455421Z02). LW was supported by a Ph.D. scholarship provided by China Scholarship Council (CSC) (Grant number: 201206220134).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fpls.2016.00750>

REFERENCES

- Agrawal, G. K., Hajdich, M., Graham, K., and Thelen, J. J. (2008). In-depth investigation of the soybean seed-filling proteome and comparison with a parallel study of rapeseed. *Plant Physiol.* 148, 504–518. doi: 10.1104/pp.108.119222
- Angelovici, R., Fait, A., Fernie, A. R., and Galili, G. (2011). A seed high-lysine trait is negatively associated with the TCA cycle and slows down *Arabidopsis* seed germination. *New Phytol.* 189, 148–159. doi: 10.1111/j.1469-8137.2010.03478.x
- Araujo, W. L., Nunes-Nesi, A., Nikoloski, Z., Sweetlove, L. J., and Fernie, A. R. (2012). Metabolic control and regulation of the tricarboxylic acid cycle in photosynthetic and heterotrophic plant tissues. *Plant Cell Environ.* 35, 1–21. doi: 10.1111/j.1365-3040.2011.02332.x
- Bhat, R., and Sridhar, K. R. (2008). Nutritional quality evaluation of electron beam-irradiated lotus (*Nelumbo nucifera*) seeds. *Food Chem.* 107, 174–184. doi: 10.1016/j.foodchem.2007.08.002
- Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254. doi: 10.1016/0003-2697(76)90527-3
- Daran-Lapujade, P., Rossell, S., Van Gulik, W. M., Luttik, M. A., De Groot, M. J., Slijper, M., et al. (2007). The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15753–15758. doi: 10.1073/pnas.0707476104
- Deng, J., Fu, Z., Chen, S., Damaris, R. N., Wang, K., Li, T., et al. (2015). Proteomic and Epigenetic analyses of lotus (*Nelumbo nucifera*) petals between red and white cultivars. *Plant Cell Physiol.* 56, 1546–1555. doi: 10.1093/pcp/pcv077
- Doerfler, H., Lyon, D., Nagele, T., Sun, X., Fragner, L., Hadacek, F., et al. (2013). Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics* 9, 564–574. doi: 10.1007/s11306-012-0470-0
- Fait, A., Angelovici, R., Less, H., Ohad, I., Urbanczyk-Wochniak, E., Fernie, A. R., et al. (2006). *Arabidopsis* seed development and germination is associated with temporally distinct metabolic switches. *Plant Physiol.* 142, 839–854. doi: 10.1104/pp.106.086694
- Finnie, C., Melchior, S., Roepstorff, P., and Svensson, B. (2002). Proteome analysis of grain filling and seed maturation in barley. *Plant Physiol.* 129, 1308–1319. doi: 10.1104/pp.003681
- Galili, G., Avin-Wittenberg, T., Angelovici, R., and Fernie, A. R. (2014). The role of photosynthesis and amino acid metabolism in the energy status during seed development. *Front. Plant Sci.* 5:447. doi: 10.3389/fpls.2014.00447
- Gallardo, K., Firnhaber, C., Zuber, H., Hericher, D., Belghazi, M., Henry, C., et al. (2007). A combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds: evidence for metabolic specialization of maternal and filial tissues. *Mol. Cell. Proteomics* 6, 2165–2179. doi: 10.1074/mcp.M700171-MCP200
- Gallardo, K., Le Signor, C., Vandekerckhove, J., Thompson, R. D., and Burstin, J. (2003). Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiol.* 133, 664–682. doi: 10.1104/pp.103.025254
- Hajdich, M., Casteel, J. E., Hurrelmeyer, K. E., Song, Z., Agrawal, G. K., and Thelen, J. J. (2006). Proteomic analysis of seed filling in *Brassica napus*. Developmental characterization of metabolic isozymes using high-resolution two-dimensional gel electrophoresis. *Plant Physiol.* 141, 32–46. doi: 10.1104/pp.105.075390
- Hajdich, M., Ganapathy, A., Stein, J. W., and Thelen, J. J. (2005). A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. *Plant Physiol.* 137, 1397–1419. doi: 10.1104/pp.104.056614
- Hildebrandt, T. M., Nunes Nesi, A., Araujo, W. L., and Braun, H. P. (2015). Amino acid catabolism in plants. *Mol. Plant* 8, 1563–1579. doi: 10.1016/j.molp.2015.09.005
- Hoehnewarter, W., Van Dongen, J. T., Wienkoop, S., Steinfath, M., Hummel, J., Erban, A., et al. (2008). A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* 8, 4214–4225. doi: 10.1002/pmic.200701047
- Hummel, J., Niemann, M., Wienkoop, S., Schulze, W., Steinhauser, D., Selbig, J., et al. (2007). ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics* 8:216. doi: 10.1186/1471-2105-8-216
- Junker, B. H., Lonien, J., Heady, L. E., Rogers, A., and Schwender, J. (2007). Parallel determination of enzyme activities and in vivo fluxes in *Brassica napus* embryos grown on organic or inorganic nitrogen source. *Phytochemistry* 68, 2232–2242. doi: 10.1016/j.phytochem.2007.03.032
- Karner, U., Peterbauer, T., Raboy, V., Jones, D. A., Hedley, C. L., and Richter, A. (2004). myo-Inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds. *J. Exp. Bot.* 55, 1981–1987. doi: 10.1093/jxb/erh216
- Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., et al. (2005). GMD@CSB.DB: the golm metabolome database. *Bioinformatics* 21, 1635–1638. doi: 10.1093/bioinformatics/bti236
- LaBerge, D. E., Macgregor, A. W., and Meredith, W. O. S. (1973). Changes in the free sugar content of barley kernels during maturation. *J. Instit. Brew.* 79, 471–477. doi: 10.1002/j.2050-0416.1973.tb03568.x
- Laino, P., Shelton, D., Finnie, C., De Leonardi, A. M., Mastrangelo, A. M., Svensson, B., et al. (2010). Comparative proteome analysis of metabolic proteins from seeds of durum wheat (cv. Svevo) subjected to heat stress. *Proteomics* 10, 2359–2368. doi: 10.1002/pmic.200900803
- Liu, Y., Chaturvedi, P., Fu, J., Cai, Q., Weckwerth, W., and Yang, P. (2016). Induction and quantitative proteomic analysis of cell dedifferentiation during callus formation of lotus (*Nelumbo nucifera* Gaertn.spp. baijianlian). *J. Proteomics* 131, 61–70. doi: 10.1016/j.jprot.2015.10.010
- Liu, Y., Li, M., Han, C., Wu, F., Tu, B., and Yang, P. (2013). Comparative proteomic analysis of rice shoots exposed to high arsenate. *J. Integr. Plant Biol.* 55, 965–978. doi: 10.1111/jipb.12058
- Mechin, V., Thevenot, C., Le Guilloux, M., Prioul, J. L., and Damerval, C. (2007). Developmental analysis of maize endosperm proteome suggests a pivotal role for pyruvate orthophosphate dikinase. *Plant Physiol.* 143, 1203–1219. doi: 10.1104/pp.106.092148
- Ming, R., Vanburen, R., Liu, Y., Yang, M., Han, Y., Li, L. T., et al. (2013). Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* 14:R41. doi: 10.1186/gb-2013-14-5-r41
- Moro, C. F., Fukao, Y., Shibato, J., Rakwal, R., Agrawal, G., Shioda, S., et al. (2015a). Immature seed endosperm and embryo proteomics of the lotus (*Nelumbo Nucifera* Gaertn.) by one-dimensional gel-based tandem mass spectrometry and a comparison with the mature endosperm proteome. *Proteomes* 3, 184–235. doi: 10.3390/proteomes3030184
- Moro, C. F., Fukao, Y., Shibato, J., Rakwal, R., Timperio, A. M., Zolla, L., et al. (2015b). Unraveling the seed endosperm proteome of the lotus (*Nelumbo nucifera* Gaertn.) utilizing 1DE and 2DE separation in conjunction with tandem mass spectrometry. *Proteomics* 15, 1717–1735. doi: 10.1002/pmic.201400406
- Mukherjee, D., Khatua, T. N., Venkatesh, P., Saha, B. P., and Mukherjee, P. K. (2010). Immunomodulatory potential of rhizome and seed extracts of *Nelumbo nucifera* Gaertn. *J. Ethnopharmacol.* 128, 490–494. doi: 10.1016/j.jep.2010.01.015
- Oliveira, A. P., and Sauer, U. (2012). The importance of post-translational modifications in regulating *Saccharomyces cerevisiae* metabolism. *FEMS Yeast Res.* 12, 104–117. doi: 10.1111/j.1567-1364.2011.00765.x
- Pan, L., Xia, Q., Quan, Z., Liu, H., Ke, W., and Ding, Y. (2010). Development of novel EST-SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. *J. Hered.* 101, 71–82. doi: 10.1093/jhered/esp070
- Peterbauer, T., and Richter, A. (2001). Biochemistry and physiology of raffinose family oligosaccharides and galactosyl cyclitols in seeds. *Seed Sci. Res.* 11, 185–197.
- Plaxton, W. C. (1996). The organization and regulation of plant glycolysis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 47, 185–214. doi: 10.1146/annurev.arplant.47.1.185
- Qi, Q., Kleppinger-Sparace, K. F., and Sparace, S. A. (1995). The utilization of glycolytic intermediates as precursors for fatty acid biosynthesis by pea root plastids. *Plant Physiol.* 107, 413–419.
- Repetto, O., Rogniaux, H., Firnhaber, C., Zuber, H., Kuster, H., Larre, C., et al. (2008). Exploring the nuclear proteome of *Medicago truncatula* at the switch towards seed filling. *Plant J.* 56, 398–410. doi: 10.1111/j.1365-3113.2008.03610.x

- Sakulsingharoj, C., Choi, S.-B., Hwang, S.-K., Edwards, G. E., Bork, J., Meyer, C. R., et al. (2004). Engineering starch biosynthesis for increasing rice seed weight: the role of the cytoplasmic ADP-glucose pyrophosphorylase. *Plant Sci.* 167, 1323–1333. doi: 10.1016/j.plantsci.2004.06.028
- Schupp, N., and Ziegler, P. (2004). The relation of starch phosphorylases to starch metabolism in wheat. *Plant Cell Physiol.* 45, 1471–1484. doi: 10.1093/pcp/pch170
- Schwender, J., Hebbelmann, I., Heinzel, N., Hildebrandt, T., Rogers, A., Naik, D., et al. (2015). Quantitative multilevel analysis of central metabolism in developing oilseeds of oilseed rape during in vitro culture. *Plant Physiol.* 168, 828–848. doi: 10.1104/pp.15.00385
- Shen-Miller, J. (2002). Sacred lotus, the long-living fruits of China Antique. *Seed Sci. Res.* 12, 131–143. doi: 10.1007/s12042-013-9124-2
- Sturm, A. (1999). Invertases. primary structures, functions, and roles in plant development and sucrose partitioning. *Plant Physiol.* 121, 1–7. doi: 10.1104/pp.121.1.1
- Sun, X., and Weckwerth, W. (2012). COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* 8, 81–93. doi: 10.1007/s11306-012-0399-3
- Tang, P. H., Sun, D. L., Zhao, J., and Su, X. Z. (1989). Light and electron microscopic observation on the developing pattern of cotyledons of *Nelumbo nucifera*. *Acta Bot. Sin.* 31, 750–756.
- The Angiosperm Phylogeny Group (2009). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Bot. J. Linnean Soc.* 161, 105–121. doi: 10.1016/j.jep.2015.05.035
- Tripodi, F., Nicastro, R., Reghellin, V., and Cocetti, P. (2015). Post-translational modifications on yeast carbon metabolism: regulatory mechanisms beyond transcriptional control. *Biochim. Biophys. Acta* 1850, 620–627. doi: 10.1016/j.bbagen.2014.12.010
- Valledor, L., and Weckwerth, W. (2014). An improved detergent-compatible gel-fractionation LC-LTQ-Orbitrap-MS workflow for plant and microbial proteomics. *Methods Mol. Biol.* 1072, 347–358. doi: 10.1007/978-1-62703-631-3_25
- Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32, 223–226. doi: 10.1038/nbt.2839
- Weber, H., Borisjuk, L., and Wobus, U. (1997). Sugar import and metabolism during seed development. *Trends Plant Sci.* 2, 169–174. doi: 10.1016/S1360-1385(97)85222-3
- Weber, H., Borisjuk, L., and Wobus, U. (2005). Molecular physiology of legume seed development. *Annu. Rev. Plant Biol.* 56, 253–279. doi: 10.1146/annurev.arplant.56.032604.144201
- Weckwerth, W., Wenzel, K., and Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4, 78–83. doi: 10.1002/pmic.200200500
- Weigelt, K., Kuster, H., Rutten, T., Fait, A., Fernie, A. R., Miersch, O., et al. (2009). ADP-glucose pyrophosphorylase-deficient pea embryos reveal specific transcriptional and metabolic changes of carbon-nitrogen metabolism and stress responses. *Plant Physiol.* 149, 395–411. doi: 10.1104/pp.108.129940
- Wienkoop, S., Staudinger, C., Hoehenwarter, W., Weckwerth, W., and Egelhofer, V. (2012). ProMEX - a mass spectral reference database for plant proteomics. *Front. Plant Sci.* 3:125. doi: 10.3389/fpls.2012.00125
- Xu, S. B., Li, T., Deng, Z. Y., Chong, K., Xue, Y., and Wang, T. (2008). Dynamic proteomic analysis reveals a switch between central carbon metabolism and alcoholic fermentation in rice filling grains. *Plant Physiol.* 148, 908–925. doi: 10.1104/pp.108.125633
- Yang, M., Xu, L., Liu, Y., and Yang, P. (2015a). RNA-Seq Uncovers SNPs and alternative splicing events in asian lotus (*Nelumbo nucifera*). *PLoS ONE* 10:e0125702. doi: 10.1371/journal.pone.0125702
- Yang, M., Zhu, L., Pan, C., Xu, L., Liu, Y., Ke, W., et al. (2015b). Transcriptomic analysis of the regulation of rhizome formation in temperate and tropical lotus (*Nelumbo nucifera*). *Sci. Rep.* 5:13059. doi: 10.1038/srep13059
- Yen, G.-C., Duh, P.-D., and Su, H.-J. (2005). Antioxidant properties of lotus seed and its effect on DNA damage in human lymphocytes. *Food Chem.* 89, 379–385. doi: 10.1016/j.foodchem.2004.02.045
- Yen, G.-C., Duh, P.-D., Su, H.-J., Yeh, C.-T., and Wu, C.-H. (2006). Scavenging effects of lotus seed extracts on reactive nitrogen species. *Food Chem.* 94, 596–602. doi: 10.1016/j.foodchem.2004.11.052
- Zheng, B. D., Zheng, J. G., and Zeng, S. X. (2003). Analysis of the nutritional composition in Chinese main lotus seed varieties. *Acta Nutri. Sin.* 25, 153–156.
- Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., and Washburn, M. P. (2006). Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 5, 2339–2347. doi: 10.1021/pr060161n

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Wang, Fu, Li, Fragner, Weckwerth and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

3. mzGroupAnalyzer

Title: mzGroupAnalyzer-Predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data

3.1 Preamble

High throughput metabolomics platforms greatly improved efficiency of data acquisition. However, robust metabolite annotation is still a challenge in metabolomic field. In the following paper, an algorithm called mzGroupAnalyzer was developed for mass spectrometry based metabolite structure exploration. Additionally, combining information of (bio)chemical transformation of novel structures from known structures and the biosynthetic pathway information of known metabolites allows mapping newly annotated candidates to their related pathways. mzGroupAnalyzer was applied to analysis metabolomic dataset of cold and high light treated *Arabidopsis thaliana* which resulted in annotation of 15 new flavonoids.

3.2 Declaration of authorship

mzGroupAnalyzer algorithm, user guideline and application results were published in *PLoS ONE*. I have participated in critical discussion, fragmentation rules searching, algorithm testing, though this work is dominantly performed by coauthors.

3.3 Published manuscript



mzGroupAnalyzer-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data

Hannes Doerfler[‡], Xiaoliang Sun[‡], Lei Wang, Doris Engelmeier, David Lyon, Wolfram Weckwerth*

Department of Ecogenomics and Systems Biology, University of Vienna, Vienna, Austria

Abstract

The metabolome is a highly dynamic entity and the final readout of the genotype x environment x phenotype (GxExP) relationship of an organism. Monitoring metabolite dynamics over time thus theoretically encrypts the whole range of possible chemical and biochemical transformations of small molecules involved in metabolism. The bottleneck is, however, the sheer number of unidentified structures in these samples. This represents the next challenge for metabolomics technology and is comparable with genome sequencing 30 years ago. At the same time it is impossible to handle the amount of data involved in a metabolomics analysis manually. Algorithms are therefore imperative to allow for automated *m/z* feature extraction and subsequent structure or pathway assignment. Here we provide an automated pathway inference strategy comprising measurements of metabolome time series using LC-MS with high resolution and high mass accuracy. An algorithm was developed, called *mzGroupAnalyzer*, to automatically explore the metabolome for the detection of metabolite transformations caused by biochemical or chemical modifications. Pathways are extracted directly from the data and putative novel structures can be identified. The detected *m/z* features can be mapped on a van Krevelen diagram according to their H/C and O/C ratios for pattern recognition and to visualize oxidative processes and biochemical transformations. This method was applied to *Arabidopsis thaliana* treated simultaneously with cold and high light. Due to a protective antioxidant response the plants turn from green to purple color via the accumulation of flavonoid structures. The detection of potential biochemical pathways resulted in 15 putatively new compounds involved in the flavonoid-pathway. These compounds were further validated by product ion spectra from the same data. The *mzGroupAnalyzer* is implemented in the graphical user interface (GUI) of the metabolomics toolbox COVAIN (Sun & Weckwerth, 2012, *Metabolomics* 8: 81–93). The strategy can be extended to any biological system.

Citation: Doerfler H, Sun X, Wang L, Engelmeier D, Lyon D, et al. (2014) *mzGroupAnalyzer*-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data. *PLoS ONE* 9(5): e96188. doi:10.1371/journal.pone.0096188

Editor: Akos Vertes, The George Washington University, United States of America

Received: October 29, 2013; **Accepted:** April 4, 2014; **Published:** May 20, 2014

Copyright: © 2014 Weckwerth et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors thank the Austrian Science Fund (FWF) for financial support to Hannes Dörfler and David Lyon (FWF project P25488 and P23441, respectively). Lei Wang was supported by a grant of the China Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wolfram.weckwerth@univie.ac.at

‡ These authors contributed equally to this work.

Introduction

Metabolomic techniques have been recently established and refined to characterize the widely heterogeneous small molecules present at a specific time in a living tissue. Several analytical approaches exist for the application of metabolomics to various biological questions, for instance gas chromatography coupled to mass spectrometry (GC-MS) for the detection of small and volatile compounds [1] or capillary electrophoresis coupled to mass spectrometry (CE-MS) for charged compounds [2]. The method with the highest preference for larger and more hydrophobic metabolites and complementary to GC-MS is liquid chromatography coupled to mass spectrometry (LC-MS) [3,4]. Recently, we showed that the GC-MS and LC-MS techniques can be integrated into a combined platform to increase the total coverage of the metabolome, as well as to provide insights into the mutual regulation of both primary and secondary metabolism by analysis of the same sample and subsequent data merging and processing [4–6]. Metabolomics-via-LC-MS approaches have yet to hit the

ranks of other analytical techniques with respect to their robustness and database availability, but no other method has the potential to achieve a better coverage of the metabolome whilst maintaining high resolution and inferring additional structural information in the process. In contrast to metabolite profiling on the GC-MS platform, where standard operating procedures and large databases are present and being improved continuously, there exists little standardization on how to approach the analysis of larger metabolites using LC-MS [7]. A significant step forward in the field of untargeted metabolomics is the advent of instruments capable of sub-ppm mass accuracy measurements as well as precursor fragmentation, enabling the acquisition of the exact mass as well as obtaining molecule fragment information in order to construct a meaningful sum formula [8,9]. Owing to these technological advances, metabolomics has already proven to be a valuable tool in fields like biomarker discovery and functional genomics [10–12]. In this study, we introduce an algorithm called *mzGroupAnalyzer* to provide characterization and identification of data signals acquired by an LC-Orbitrap-FT-MS system utilizing

sub-ppm mass measurements and intelligent sum formula query. We applied this strategy to plant secondary metabolism. Plants are most important resources for natural products, providing the highest diversity of chemical structures in the range of 200,000–400,000 different compounds and a high *in vivo* plasticity in response to environmental conditions. Due to this diversity of chemical structures ranging from simple hydrocarbons to complex heteroatomic molecules, the elucidation of the metabolome of higher plants and in general of natural products of any origin poses a challenge for metabolomic techniques. Traditionally, the metabolism of higher plants is subclassified into the primary metabolism, which comprises molecules with a low molecular weight that are involved in the central energy conversion cycles of the organism, and the secondary metabolism, which is not per se involved in energy homeostasis but rather involved in the chemical communication with the environment. Only recently, the field of plant secondary metabolites has begun moving more and more into the focus of new bio-analytical techniques and their stereotypic role as simple chemical weapons is being revised as numerous findings indicate their ability to stimulate vital processes in the cell by regulating the concentration of reactive oxygen species *in vivo* [13,14]. Also, secondary metabolites are of vast interest to the area of medicine and nutrition, as these phytochemicals often possess physiological activity in the human body, for example antioxidant activity or cancer chemoprevention [15].

To provide a suitable experimental system, we stressed *Arabidopsis thaliana* plants by parallel cold and light treatment introducing high oxidative stress over a 3-week interval. This led to a comprehensive switch of the vegetative growth metabolism to protective accumulation of secondary metabolites. We show a way of embedding the multitude of signals into a biochemical context through automated detection of metabolic steps between the acquired *m/z* features. Putative structures are inferred by analysis of the product ions from the same data. By applying *mzGroupAnalyzer* to LC-MS raw data, we are able to prove the existence and relation of known molecules as well as propose novel compounds and pathways, in the present case molecules arising during oxidative stress within the secondary metabolism of *Arabidopsis thaliana*. This strategy can be applied systematically and conveniently to any kind of LC-MS data set and is expected to improve identification and structural elucidation in complex metabolomics data, which is currently the limiting step in large-scale metabolomics studies.

Materials and Methods

Plant material and harvest

Arabidopsis thaliana Col-0 was cultivated in a growth chamber under controlled conditions: light intensity was $280 \mu\text{mol m}^{-2} \text{s}^{-1}$ in an 8-hour light/16-hour dark day cycle; relative humidity was 60% with an average temperature of 22°C. Time point “zero” of cold stress consisted of replicates of non-stressed plants, while every 2 days another sample batch of cold-acclimated plants in a 4°C cold chamber was taken. Rosette leaves were harvested approximately 2 hours after the beginning of the light period. Metabolic activity in the leaves was quenched by immediately putting the plant material into liquid nitrogen after harvesting. Deep-frozen leaf material was ground to a fine powder with a pestle and mortar under steady addition of liquid nitrogen and subsequently stored at -80°C before measurement.

Chemicals

Methanol (HPLC-grade), chloroform (anhydrous, >99%, p.a.) and acetonitrile (UHPLC-grade) were purchased from Sigma-Aldrich (Vienna, Austria). Formic acid (98–100%) was purchased from Merck (Vienna, Austria). Chloramphenicol (>98%) and Ampicillin trihydrate (analytical standard) were purchased from Fluka (Vienna, Austria).

Extraction procedure and sample preparation for secondary metabolite analysis

For LC-MS analysis, about 50 mg of frozen plant-leaf material was extracted by 1 ml pre-chilled 80/20 v:v MeOH/H₂O solution containing 1 μg of the internal standards Ampicillin and Chloramphenicol. Samples were centrifuged at 15,000 g for 15 minutes. The supernatant was dried out overnight in a new tube and re-dissolved in 100 μl of 50/50 v:v MeOH/H₂O solution and centrifuged again for 15 minutes at 20,000 g. The supernatant was then filtered through a STAGE tip (Empore/Disk C18, diameter 47 mm) before it was conveyed into a GC vial with a micro insert tip. Plant extracts were further extracted with 500 μl of chloroform to remove the highly abundant lipid components. The LC-MS method for secondary metabolite analysis has been described before [6].

Data processing, *mzGroupAnalyzer* and pathway viewer

Both *mzGroupAnalyzer* and *Pathway Viewer* have been integrated into the GUI of the COVAIn toolbox [16]. The standalone version of COVAIn can be downloaded at <http://www.univie.ac.at/mosys/software.html>. The data processing strategy and subsequent analysis of the data using *mzGroupAnalyzer* and *Pathway viewer* are explained in the *mzGroupAnalyzer*-Tutorial (Figure S1). *m/z*-values acquired by LC-MS were exported to Excel data sheets using the Xcalibur software. Elemental composition determination was enabled, with a maximum of 10 possible sum formulas for each compound and a ppm deviation of 1.

The single excel-sheets can be uploaded to *mzGroupAnalyzer* via the GUI of COVAIn. Furthermore, a user-defined rules-file is uploaded and the folder for storage of the result-files is provided. By starting the *mzGroupAnalyzer*, the lists of *m/z* values with associated chemical compositions from Xcalibur output are read and the atomic differences between *m/z*-pairs are calculated and compared with the putative chemical modifications provided by the rules-file.

Based on all chemical modifications provided by the rules-file, *mzGroupAnalyzer* searches pathways between pairs of *m/z* features. Regarding each *m/z* feature as a node in a metabolic network, an edge connects two nodes if a chemical modification exists. These edges and nodes generate large networks. A pathway can therefore be constructed by searching the shortest path between two nodes. Redundant paths that are included in other longer paths are removed. The pathway searching algorithm can deal with time series data by filtering out pathways that do not reflect the correct chronological order of the given measurements. For example, there is no path from *m/z* feature *A* occurring on Day 2 to *m/z* feature *B* occurring on Day 1, although theoretically a chemical modification from *A* to *B* is possible. Finally, to better visualize the pathways, we further developed *Pathway Viewer*, which is integrated in *mzGroupAnalyzer*, and is able to plot the pathways after a series of user-defined filtering options like *m/z* range or time points. The following result files will be created, exported and saved:

1. Transformations corresponding to the rules-file.
2. A ranking of frequently found chemical modifications.

- Putative transformations that were not listed in the rules-file.
- A mzStructure file for *Pathway viewer*.
- A Pajek-file (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>) for network visualization.

Next the *Pathway viewer* is started and a table of transformations and the possibility for visualization of pathways is provided (see Figure S1 – *mzGroupAnalyzer*-Tutorial). The above process is summarized in Figure 1. The list of chemical and biochemical reactions searched by *mzGroupAnalyzer* (currently comprising 56 metabolic reaction steps in the provided rules-file) is shown in Table S1. This list can be easily extended to novel transformations.

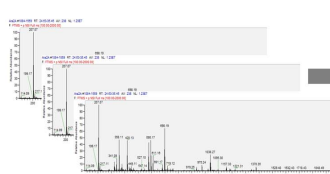
For the construction of the van Krevelen plots, sum formulas with *mzGroupAnalyzer*-predicted metabolic transformations were assorted in an Excel sheet and their O/C and H/C ratios were calculated. These values were exported to SigmaPlot 12.3 and mapped to a multiple scatter plot within the boundaries 0 to 1 for the O/C ratio and 0 to 2 for H/C ratio.

Results and Discussion

Development of the *mzGroupAnalyzer* algorithm to identify biochemical and chemical transformations in non-targeted metabolomics data

The development and application of algorithms is essential to systematically search for biochemical and chemical transformations of compounds and to find putative pathways in highly complex LC-MS based metabolomics data. We have established an algorithm which is able to extract putative chemical transformations from high mass accuracy metabolome data. The algorithm generates multiple potential pathways directly from raw-LC-MS data and visualizes these as networks. The entire approach is implemented as a graphical user interface (GUI) in COVAIN (Figure 1). COVAIN is a toolbox for statistical data mining in metabolomics and other OMICS approaches [16] in the Matlab environment. In order to evaluate the algorithm, we measured various reference compounds of typical plant secondary metabolites with nano-UPLC-Orbitrap-MS. Subsequently, sum formulas were generated in Xcalibur 2.0 using the integrated sum formula

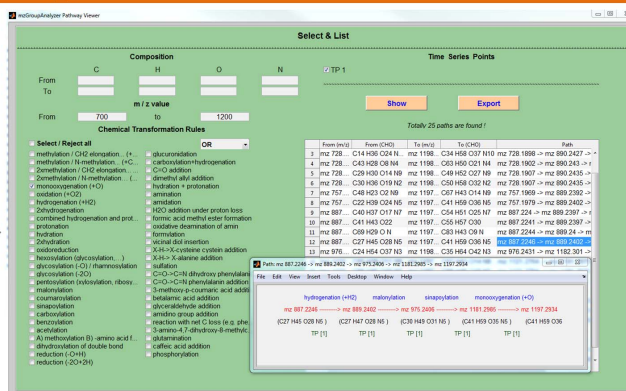
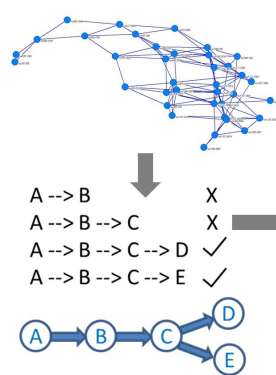
Read m/z data and reaction rules



A	B	C	D	E	F
m/z	Intensity	Relative	Thes. Mass	Delta (ppm)	Composition
448.11559	25978.6	0.86	448.11578	-0.41	C11 H22 O14 N5
448.11527			448.11527	0.72	C23 H30 O8
448.11526			448.11526	0.73	C24 H34 O3 N7
449.10783	18821.2	6.2	449.10783	-0.01	C20 H15 O8 N7
449.10784			449.10784	-0.02	C12 H21 O11
594.18977	9240.6	0.3	594.18977	-0.01	C36 H58 O5 N6
594.19028			594.19028	-0.86	C22 H28 O11 N9
594.19028			594.19028	-0.87	C23 H34 O16 N2
594.35642	8594.1	0.28	594.35644	-0.04	C37 H46 O3 N6
594.35695			594.35695	-0.89	C23 H48 O9 N9
594.35696			594.35696	-0.9	C24 H54 O14 N2
595.14462			595.14462	0.21	C30 H27 O13
595.14461			595.14461	0.22	C29 H24 O8 N7
595.14512			595.14512	-0.64	C16 H29 O19 N5

Reaction step / chemical modification	A	B	C	D	E	F
net C change	net H change	net O change	net N change			
methylation / CH3 elongation... (-CH2)	1	2	0	0		
demethylation / N-methylation... (-CH2)	1	1	0	0		
demethylation / CH2 elongation... (-CH2)	2	4	0	0		
demethylation / N-methylation... (-CH2)	2	6	0	0		
monooxygenation (+O)	0	0	1	0		
oxidation (+O2)	0	0	2	0		
hydrogenation (+H2)	0	2	0	0		
hydroxylation	0	4	0	0		
N-alk-oxime system addition	3	4	2	1		
N-alk-3-amine addition	1	5	2	1		
sulfation	0	0	3	0		
combined hydrogenation and protonation	0	0	0	0		
protonation	0	1	0	0		
hydrolysis	0	2	1	0		
hydroxylation	0	4	2	0		
hydroxylation	0	12	1	0		
hydroxylation (glycosylation...)	4	40	5	0		
glycosylation (+O2) / phosphorylation	6	30	4	0		

mzGroupAnalyzer + Pathway Viewer



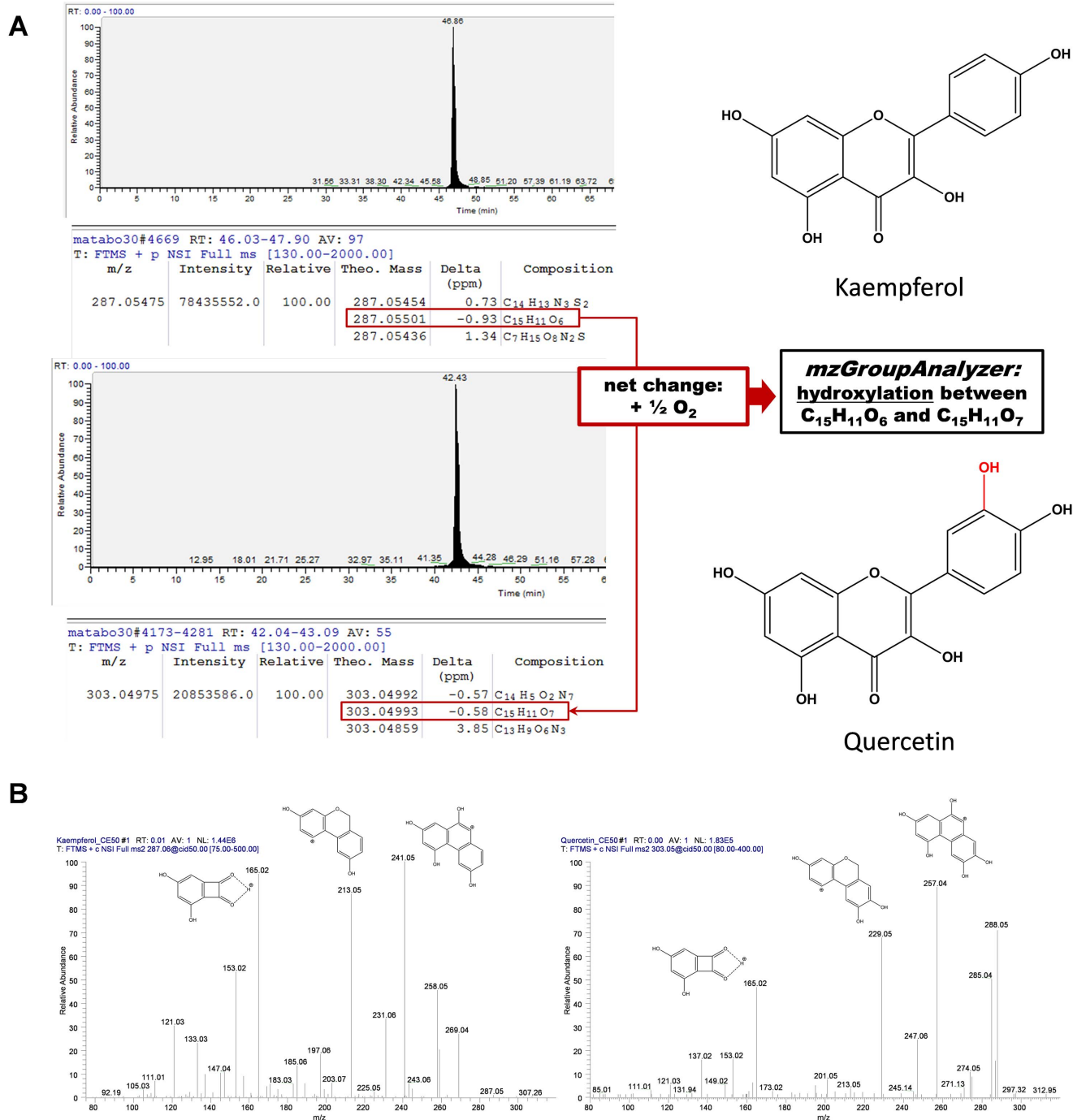


Figure 2. *mzGroupAnalyzer* is able to detect possible metabolic steps out of various proposed sum formulas for a measured m/z feature. **A** Kaempferol and Quercetin standards measured by LC-MS result into several sum formula suggestions for the measured mass-to-charge-ratio (m/z). Because a low ppm deviation of assigned elemental composition to the mass is not the decisive factor, the correct sum formula might not be the first one proposed and thus several must be looked at, which is handled by the program automatically. If *mzGroupAnalyzer* finds a possible reaction step (out of a list of reactions which can be altered manually), it is reported to the user. **B** MS² spectra of Kaempferol (left) and Quercetin (right). The fragmentation schemes are in accordance with published literature [42,43]. The difference of one oxygen atom (nominal mass 16) is visible in the fragments m/z 213→229 and 241→257, while m/z 165 occurs in both product scans. doi:10.1371/journal.pone.0096188.g002

calculator. Within a 1 ppm mass accuracy window and using the monoisotopic masses of several common elements, possible sum formulas were obtained (Figure 2 and Figure S1 – *mzGroupAnalyzer*-Tutorial). Many of the sum formulas did not result in feasible chemical structures but had a very high mass accuracy according

to the measured compound, thus leading to false positives. As a consequence the correct sum formula prediction was not among the top first hits. The fact that high mass accuracy alone is not sufficient for correct sum formula prediction has been shown before [17,18]. This is especially true for metabolites which often

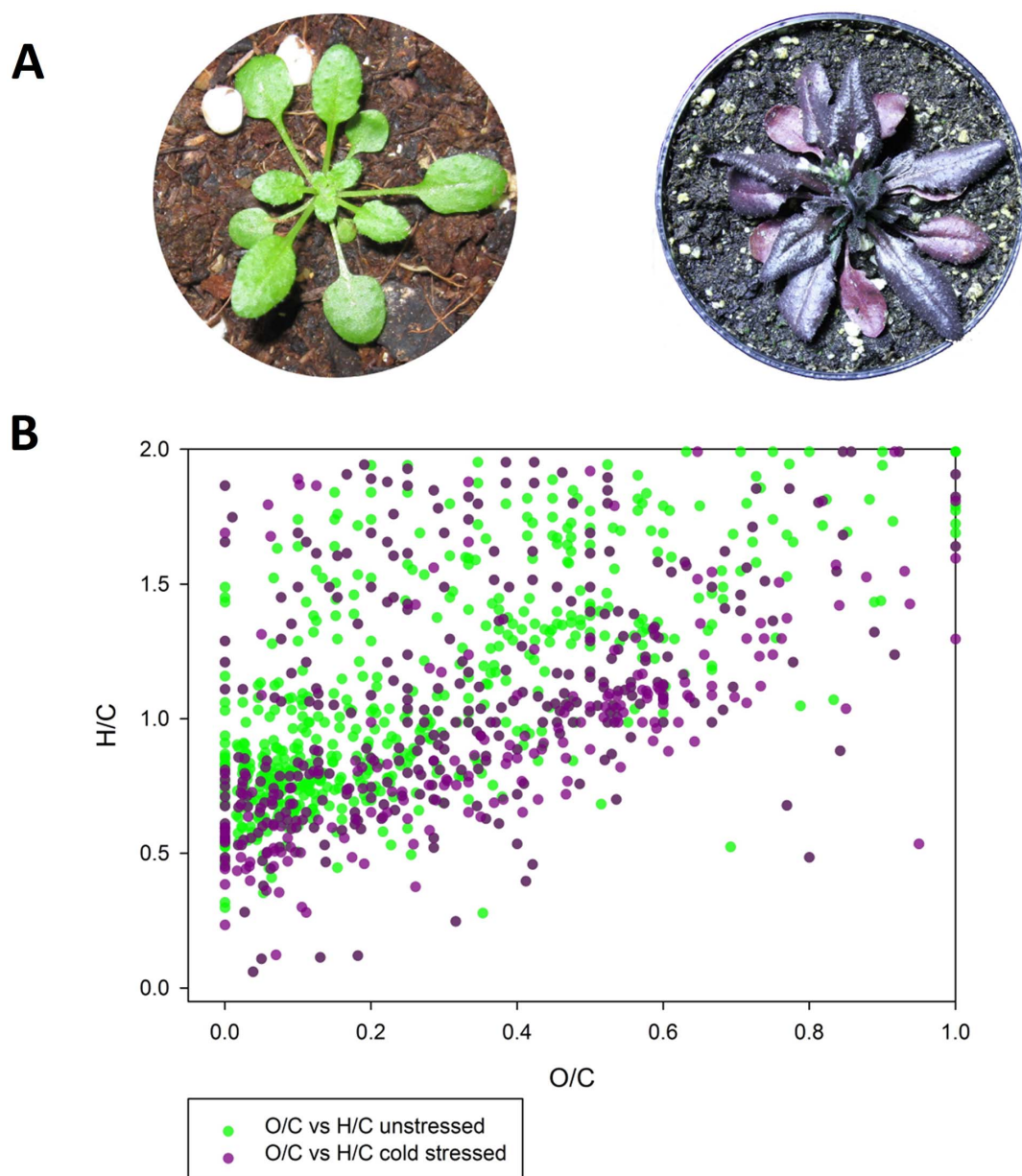


Figure 3. After oxidative stress the *Arabidopsis thaliana* plants turn from green into purple indicating a dramatic shift in metabolism, specifically elevated flavonoid biosynthesis involved in oxidative stress protection [6]. **A** Plants turn from green to purple under high light and cold temperature treatment. **B** Van Krevelen diagram of the most abundant m/z values of unstressed (green dots) and 20-day cold stressed (purple dots) *Arabidopsis* plants. A clear shift of metabolism in the stressed plants is visible.
doi:10.1371/journal.pone.0096188.g003

contain elements like S and P, or even halogens. The utilization of electrospray ionization tends to produce Na^+ , K^+ and other adducts in positive ionization mode. Because *mzGroupAnalyzer* looks for putative chemical transformations of compounds the considered number of sum formulas can be reduced. The application of *mzGroupAnalyzer* revealed metabolic conversions of the protonated reference compound such as the addition of a hydroxyl group to Kaempferol leading to Quercetin (Figure 2). Due to this chemical transformation the sum formula pair of Kaempferol and Quercetin is automatically detected. Indeed, this reaction occurs in the flavonol biosynthetic pathway catalyzed by a flavonoid 3'-monooxygenase. The constraints for detectable metabolic reactions (e.g. +1C+2H denotes a net methylation) are uploaded to the

program before performing the analysis and can be customized by the user (Figure S1 – *mzGroupAnalyzer*-Tutorial). *mzGroupAnalyzer* is also able to recognize the frequency of equidistant steps between m/z values and exports these data as suggestions for novel modifications (see Methods section). Furthermore, whole series of reaction steps in the data can be detected and analyzed in the context of metabolic pathways. This will be discussed in the next sections.

Cold- and light-induced stress has a dramatic effect on the *Arabidopsis thaliana* metabolome

To test the performance of the *mzGroupAnalyzer* algorithm we designed the following experiment. *Arabidopsis thaliana* plants were

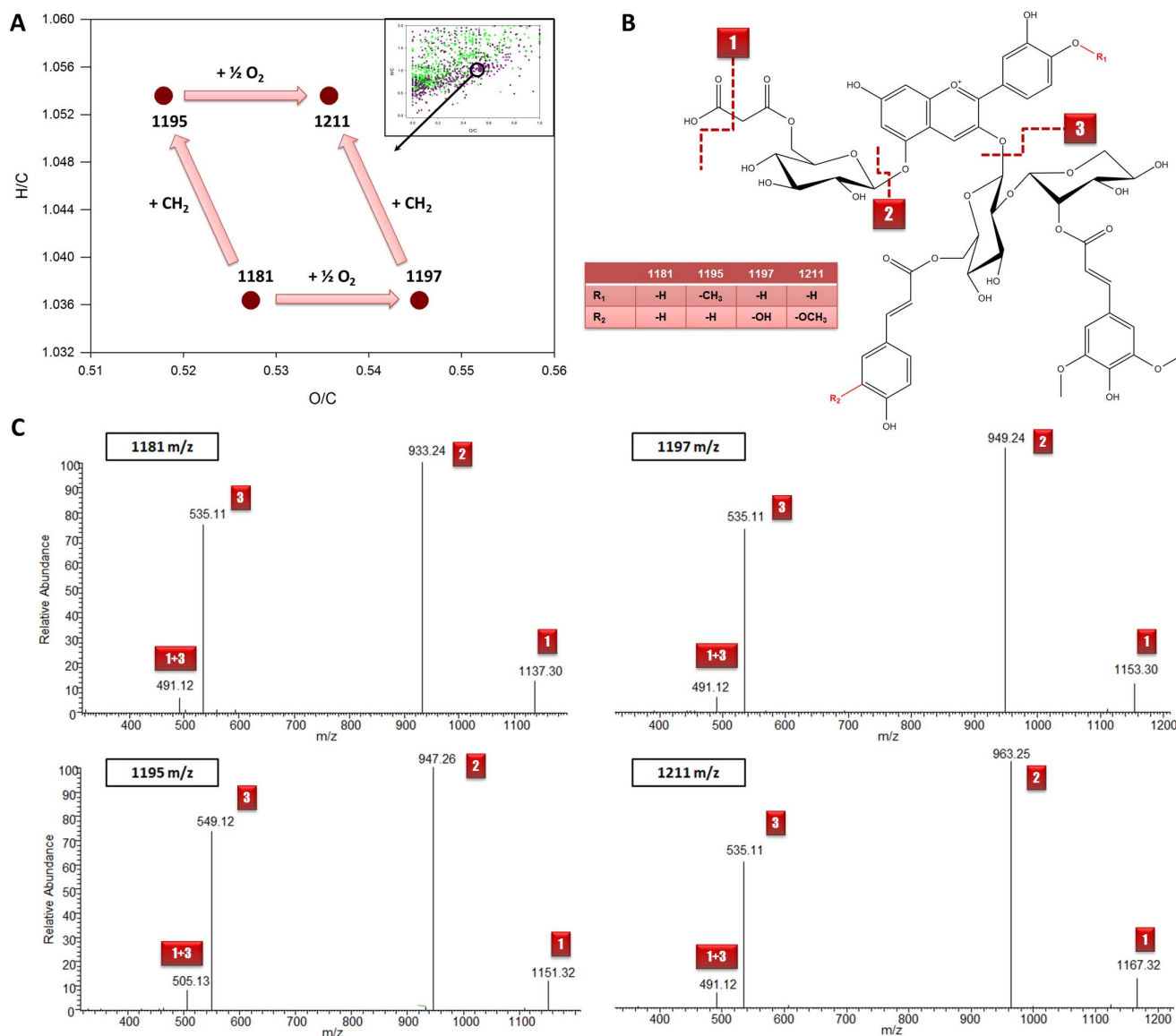


Figure 4. Exploration of the van Krevelen diagram created by sum formulas with chemical transformations detected by *mzGroupAnalyzer*. **A** *m/z* 1181, 1195, 1197 and 1211 are interconnected with net shifts of $\frac{1}{2} O_2$ and a CH_2 group and form a rhombic pattern. **B** Proposed fragmentation scheme of these compounds under the chosen conditions. **C** Product ion scans show similar fragmentation behavior of the polysubstituted anthocyanins. The spectrum of *m/z* 1195 shows a peak at *m/z* 549, pointing to a methyl group at the cyanidin core. A putative methylation site is shown.

doi:10.1371/journal.pone.0096188.g004

exposed to excess irradiation in a 4°C environment. This treatment is described in the literature to produce high levels of reactive oxygen species (ROS) [19–21]. ROS, such as hydrogen peroxide (H_2O_2), hydroxyl radicals (OH), superoxide anion (O_2^-) and singlet oxygen ($^1O_2$), are unavoidable by-products of photosynthetic organisms occurring in organelles with a high oxidative turnover rate during normal metabolic activity and can be highly damaging for cells and tissues under stress [13,22,23]. These stress conditions require an effective scavenging system in order to prevent the organism from being damaged by free radicals [24]. Especially biomolecules from the phenylpropanoid family, such as flavonoids and anthocyanins, have been recognized as effective radical-scavenging compounds [25,26]. After several days of stress in our experimental systems, the plants began to produce violet pigments in the rosette leaves (Figure 3A). The

purple color is the result of the accumulation of anthocyanidins as a response to oxidative stress [25–28]. The enhanced production of anthocyanidins under these stress conditions requires a large-scale metabolic reprogramming as recently described by us [6]. Leaf extracts of the cold/light-treated *Arabidopsis thaliana* plants were analyzed with LC/MS. From these analyses sum formulas of putative metabolites were generated based on the acquired *m/z* values focusing only on C, H and O elements (see also Figure S1 – *mzGroupAnalyzer*-Tutorial). The generated sum formulas were loaded into a van Krevelen plot to visualize the chemical and biochemical transformations of metabolites during cold and light stress (Figure 3B). Oxidative shifts in the plot induced by cold-related stress might be explained by the incorporation of oxygen, as well as radical scavenging by redox-active compounds, such as aromatic hydroxyl groups which can stabilize radicals after

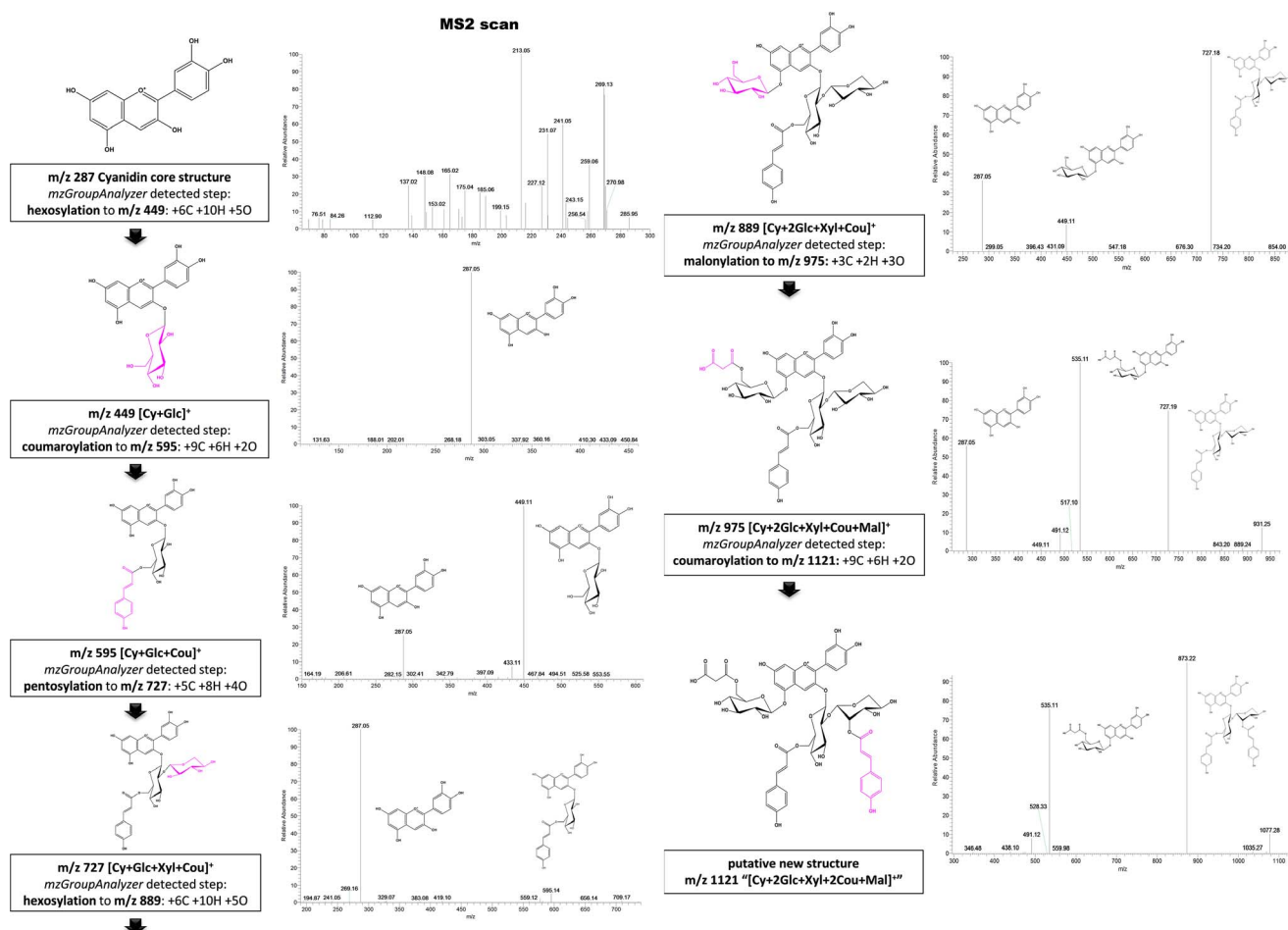


Figure 5. Identification of biochemical transformations of *in vivo* data using *mzGroupAnalyzer*. A metabolic pathway leading to a putative new compound *m/z* 1121 is revealed. Amongst several hundreds of other interlinked *m/z* values in the data, the figure shows metabolic transitions derived from sub-ppm accuracy measurements on the left side and their corresponding MS² product ion scans on the right. Comparison of the spectral information from step to step reveals the possible location of metabolic structural changes. Stereochemistry is assumed due to literature findings [33].

doi:10.1371/journal.pone.0096188.g005

deprotonation. Van Krevelen plots were originally introduced to characterize carbon-based resources like mineral oils and coal according to their possible chemical composition acquired by high-resolution mass spectrometry [29,30]; only recently, van Krevelen plots have been applied as useful tools for visualization of metabolic processes and pathways [31] as well as for sum formula annotation of natural organic matter (NOM) [32].

The investigation of van Krevelen plots composed of untargeted metabolomics data can validate structural familiarity between compounds through a metabolic pattern as depicted in Figure 4. In the O/C ratio range from 0.51 to 0.55 and H/C range of 1.036 to 1.056, the chemical relation of compounds *m/z* 1181, 1195, 1197 and 1211 is visible (Figure 4A), while their structures are validated by MS² product ion scans (Figure 4B and 4C). These compounds were also automatically detected by the *mzGroupAnalyzer* approach when investigating the putative chemical and biochemical transformations from the generated sum formula hits (see also Figure S1 – *mzGroupAnalyzer*-Tutorial). In the following section we explored the potential of *mzGroupAnalyzer* to reveal full pathways leading to novel structures.

Broad-scale analysis of metabolic conversions and novel structure prediction by *mzGroupAnalyzer* in the cold/light stress metabolome of *Arabidopsis thaliana*

Time-dependent sampling of *Arabidopsis thaliana* leaf samples in cold stress yielded strong alterations in the metabolic profiles (see above). All precursor ions from the LC/MS analysis obtained from one sampling time point were assigned to ten possible sum formulas with the following parameters: 100 max C, 200 max H, 50 max O, 10 max N, 1 ppm maximum deviation from suggested sum formula. Data from all time points were exported from Xcalibur 2.0 software as single Excel sheets and uploaded into *mzGroupAnalyzer* using the graphical user interface (see above and methods; see Figure S1 – *mzGroupAnalyzer*-Tutorial). Further, a user-defined rules-file of the molecular shifts of chemical and biochemical transformations needs to be uploaded via the GUI. Here, we provide a rules-file with 56 reactions. By starting the *mzGroupAnalyzer* via the GUI, the algorithm detects metabolic transformations between pairs of *m/z* values in the uploaded list of suggested sum formulas. A table of detected pathways is generated using the GUI option “Pathway Viewer”. Individual pathways can be visualized by clicking on the corresponding cell. Furthermore, the whole pathway network can be exported as a Pajek-file for

Table 1. Putative compounds including their *mzGroupAnalyzer*-predicted sum formulas, the corresponding exact mass as well as dominant MS² product ion fragments.

name or <i>m/z</i> value	sum formula detected by <i>mzGroupAnalyzer</i>	exact mass	main fragments	mass accuracy (ppm)	reference
A1	C ₃₂ H ₃₉ O ₂₀	743.20292	287, 581	0.35	Tohge et al. [33]
A2	C ₃₅ H ₄₁ O ₂₃	829.20331	287, 535, 581, 785	0.29	Tohge et al.
A3	C ₄₁ H ₄₅ O ₂₂	889.23970	287, 449, 727	−0.11	Tohge et al.
A4	C ₄₃ H ₄₉ O ₂₄	949.26083	287, 449, 787	0.48	Tohge et al.
A5	C ₄₄ H ₄₇ O ₂₅	975.24009	491, 535, 727, 931	−0.23	Tohge et al.
A6	C ₄₇ H ₅₅ O ₂₇	1051.29252	449, 889	−0.27	Tohge et al.
A7	C ₅₂ H ₅₅ O ₂₆	1095.29761	449, 933	0.41	Tohge et al.
A8	C ₅₀ H ₅₇ O ₃₀	1137.29292	491, 535, 889	−0.13	Bloor & Abrahams [40]
A9	C ₅₅ H ₅₇ O ₂₉	1181.29800	491, 535, 933, 1137	−0.28	Bloor & Abrahams
A10	C ₅₈ H ₆₅ O ₃₁	1257.35043	449, 1095	−0.07	Tohge et al.
A11	C ₆₁ H ₆₇ O ₃₄	1343.35083	491, 535, 1095, 1299	0.51	Bloor & Abrahams
A12	C ₂₆ H ₂₉ O ₁₅	581.15010	287, 449	0.43	Tohge et al.
A13	C ₃₅ H ₃₅ O ₁₇	727.18688	287, 581	0.23	Tohge et al.
A14	C ₃₇ H ₃₉ O ₁₉	787.20801	287	0.21	Tohge et al.
A15	C ₄₁ H ₄₅ O ₂₂	889.23970	287	−0.14	Tohge et al.
A16	C ₄₆ H ₄₅ O ₂₁	933.24478	287, 727	0.28	Tohge et al.
A17	C ₅₂ H ₅₅ O ₂₆	1095.29761	449, 933	0.56	Tohge et al.
919	C ₄₂ H ₄₇ O ₂₃	919.25026	287, 757	0.16	-
991	C ₄₄ H ₄₇ O ₂₆	991.23501	491, 535, 743, 947	0.17	-
1005	C ₄₅ H ₄₉ O ₂₆	1005.25066	491, 535, 757, 961	0.28	-
1035	C ₄₆ H ₅₁ O ₂₇	1035.26122	491, 535, 787, 991	0.31	-
1065	C ₅₁ H ₅₃ O ₂₅	1065.28704	449, 903	0.29	-
1081	C ₅₁ H ₅₃ O ₂₆	1081.28196	449, 919	0.68	-
1111	C ₅₂ H ₅₅ O ₂₇	1111.29252	449, 949	0.36	-
1121	C ₅₃ H ₅₃ O ₂₇	1121.27687	491, 535, 873, 1077	0.13	-
1125	C ₅₃ H ₅₇ O ₂₇	1125.30817	449, 963	0.51	Saito et al. [39]
1151	C ₅₄ H ₅₅ O ₂₈	1151.28744	491, 535, 903, 1107	0.48	-
1167	C ₅₄ H ₅₅ O ₂₉	1167.28235	491, 535, 919, 1123	0.54	Kasai et al. [41]
1195	C ₅₆ H ₅₉ O ₂₉	1195.31365	505, 549, 947, 1151	0.49	-
1197	C ₅₅ H ₅₇ O ₃₀	1197.29292	491, 535, 949, 1153	0.28	Saito et al.
1211	C ₅₆ H ₅₉ O ₃₀	1211.30857	491, 535, 963, 1167	0.05	Saito et al.
1313	C ₆₀ H ₆₅ O ₃₃	1313.34026	491, 535, 1065, 1269	0.13	-
1329	C ₆₀ H ₆₅ O ₃₄	1329.33518	491, 535, 1081, 1285	0.43	-
1359	C ₆₁ H ₆₇ O ₃₅	1359.34574	491, 535, 1111, 1315	0.58	-
1373	C ₆₂ H ₆₉ O ₃₅	1373.37156	491, 535, 1125, 1329	0.67	-
1549	C ₇₂ H ₇₇ O ₃₈	1549.40873	535, 1301, 1505	0.57	-

The nomenclature is according to [33]. Compounds *m/z* 1125, 1197 and 1211 were found in *Matthiola incana* by [39].
doi:10.1371/journal.pone.0096188.t001

visualization (see Figure S1 – *mzGroupAnalyzer*-Tutorial and Figure 1). Over the course of the cold and light stress, several highly frequent reactions were identified. Overall, the 10 predominant reactions were mono-oxygenations, followed by hydrogenations, hydrations, methoxylations, methylations, oxidoreductions, oxidations, malonylations, hexosylations, and dihydroxylations of double bonds. Due to possible false positives within the detected metabolic steps and reactions, these reaction lists have to be carefully validated manually and present preliminary results. More evaluation of the *mzGroupAnalyzer* and further proof-of-concept studies are needed in the future.

Nevertheless, *mzGroupAnalyzer* reported many reactions, which were validated to be metabolic pathways: By investigating the information from the MS² spectra as well as comparing them with published literature (e.g. [33], see Table 1), the presence of several compounds of the anthocyanin family in the *Arabidopsis* plants was revealed. Additionally, the program reported metabolic steps from these compounds to previously undescribed *m/z* features, which are suspected to be new anthocyanins. In Figure 5, it is shown how a metabolic pathway is extracted by *mzGroupAnalyzer*. The prediction of the pathway is validated by corresponding MS² product ion spectra from the same data set. This pathway leads to

Ara3A #1400 RT: 31.79 AV: 1 NL: 1.36E5
T: FTMS + c NSI d Full ms2 1121.28@cid50.00 [295.00-1135.00]

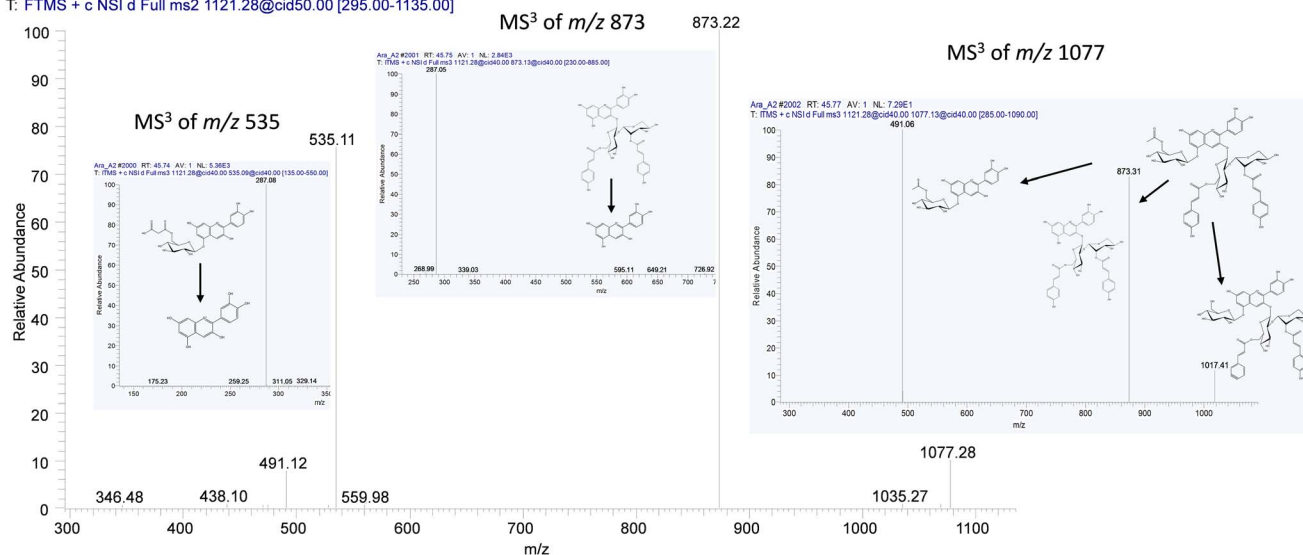


Figure 6. Structure validation of m/z 1121 by MS^3 product ion scans. Both MS^2 fragments m/z 535 and 873 result in the core cyanidin structure by undergoing MS^3 fragmentation. m/z 1077, the putatively decarboxylized form of 1121, yields m/z 491, as observed in the MS^2 spectrum already, by scission of the 3-O-glycosidic bond. Fragment m/z 873 arises again from the breaking of the glycosidic bond at 5-O. m/z 1017 would comply with the complete removal of the rest of the former malonyl group together with a water loss (-60 u). A putative structure is given. doi:10.1371/journal.pone.0096188.g006

a putatively novel compound m/z 1121 which was detected automatically by *mzGroupAnalyzer*.

mzGroupAnalyzer detected reactions between the m/z signals 287, 449, 595, 727, 889, 975 and 1121. m/z 287, the core cyanidin structure, is hexosylated to structure m/z 449, which itself shows the intact cyanidin structure in its MS^2 product ion scan and which has been reported as cyanidin 5-O-glycoside [33]. A coumaroyl group is then added to the sugar group in m/z 449, resulting in compound m/z 595, while showing the previous two compounds in the MS^2 spectrum. To m/z 595, a 5-C sugar is then added to the already existing hexose group – judged from the MS^2 spectrum – forming compound m/z 727. In the MS^2 spectrum of m/z 727 (“A13”) only m/z 595 is visible, indicating that the sugar-sugar bond is more prone for collision-induced dissociation (CID) than the hexose-coumaroyl bond. Next, another metabolic shift of $+6C$, $+10H$ and $+5O$ from m/z 727 to m/z 889 (“A3”) was detected by *mzGroupAnalyzer*, which indicates a hexosylation reaction. Indeed, MS^2 fragmentation scans again showed peaks at m/z 287 and m/z 449, as well as m/z 727 itself in the spectrum, proving our assumptions that CID is happening in both positions, 3-O as well as 5-O. From m/z 889 to m/z 975, a malonylation step was detected. In the product ion scans, the fragment m/z 727 is again present, as well as a new peak at m/z 535, with m/z 449 nearly disappearing. m/z 975 has been reported as molecule “A5” before [33], and coincides with all our findings both in the metabolic route as well as in the MS^2 spectra. The last step in this reaction list was detected to be a coumaroylation from m/z 975 to m/z 1121. Again, fragment 535 is found in the MS^2 scan, while now a higher peak, m/z 873, emerges. We assume this peak is the structure of m/z 727 with another coumaroyl group in position 2 of the xylose, as this position tends to carry further groups. The peaks at m/z 1077 and m/z 931 in the MS^2 product ion scan of m/z 1121 and m/z 975, respectively, correspond to a decarboxylation of the malonyl group in the native molecule. Furthermore the m/z 491 in these two MS^2 product ion scans supports the decarboxylation reaction of m/z 535. Figure 6 shows additional MS^3

product ion scans of the isolated MS^2 product ions m/z 535, 873 and 1077. Both MS^2 peaks m/z 535 and 873 result in the core cyanidin structure m/z 287 by undergoing MS^3 fragmentation. m/z 1077, the putatively decarboxylized form of m/z 1121, yields m/z 491, as observed in the MS^2 spectrum already, by CID-cleavage of the 3-O-glycosidic bond. Fragment m/z 873 arises again from the CID-cleavage of the glycosidic bond at 5-O. m/z 1017 is a putative structure generated by CID cleavage of an acetyl-group together with a water loss (-60 u). These findings lead us to propose m/z 1121 as compound “[Cy+2Glc+Xyl+2Cou+Mal] $^{+}$ ”, or systematically, cyanidin 3-O-[2"-O-(6"-O-(p-coumaroyl) xylosyl) 6"-O-(p-O-(glucosyl)-p-coumaroyl) glucoside] 5-O-(6"-O-malonyl) glucoside. The resulting putative structure is shown in Figure 5.

mzGroupAnalyzer detected a differential appearance of various anthocyanidins (see Table 1) during the different time points. m/z 287 was detected in all time points, m/z 449 after 2 days, and 595, 727, 889, 975 and 1121 only after 4 days of oxidative stress. Following this strategy, 15 new compounds in *Arabidopsis thaliana* could be proposed by investigating the *mzGroupAnalyzer* pathway suggestions together with their product ion spectra m/z . Putative compounds with their *mzGroupAnalyzer*-predicted sum formulas, the corresponding exact mass as well as dominant MS^2 product ion fragments are summarized in Table 1.

Using these new substances in combination with confirmed compounds from the literature, a network of the anthocyanin family starting with the KEGG pathway was reconstructed (Figure 7). Product ion scans of the new compounds and their reconstructed structures can be found in the Figure S2.

Conclusions

In this study, we showed that the application of *mzGroupAnalyzer* – a novel algorithm for untargeted identification of chemical modifications in metabolome data – on time-dependent, high-throughput LC-Orbitrap-FT-MS metabolomic profiles can give new insights into biochemical pathways, and combined with MS^n scans has the power to validate known compounds and predict

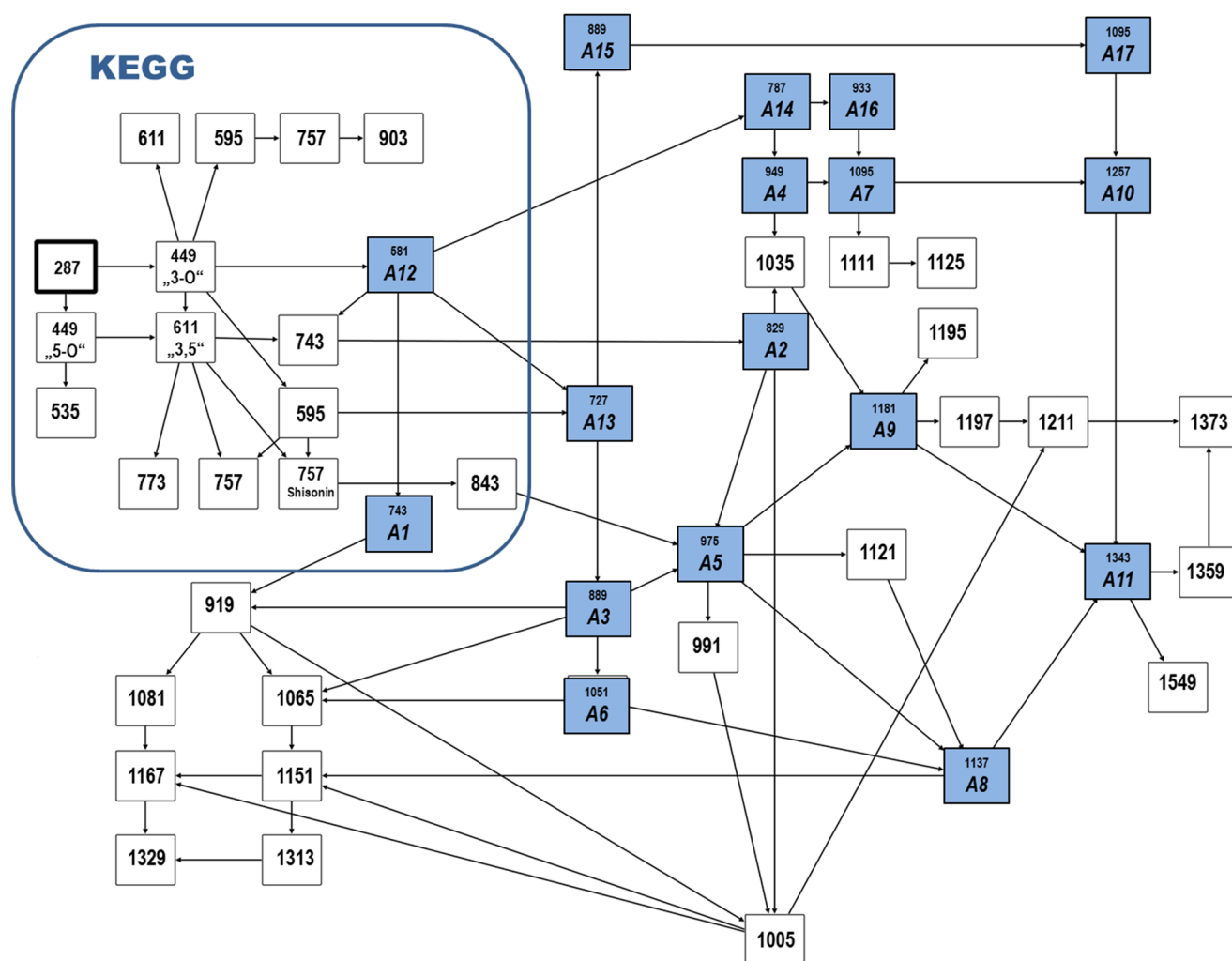


Figure 7. A proposed network of the detected anthocyanin family featuring putatively novel compounds as well as known structures including the KEGG pathway of anthocyanin biosynthesis [38,44]. Only compounds from the KEGG anthocyanin pathway are depicted, for which a suitable precursor mass was found in the data. Exact masses, sum formulas, and main MS² fragments of the new compounds are compiled in Table 1; reconstructed structures together with MS² scans are in the supporting information. The network was created with VANTED [45,46].

new structures. Attempts at assigning sum formulas to highly resolved metabolomic data have been done before and have proven to be very fruitful [34,35], yet it is clear that those approaches ultimately have to be automatized. The visualization of the elemental composition of metabolites on a van Krevelen diagram is useful for recognition of metabolic patterns, which can point to a structural similarity between those molecules.

Several strategies will be implemented in future to further improve the algorithm. First, the transformation frequency will be used in future to rank sum formulas corresponding to the same m/z feature. Secondly, more strict sum formula filtering criteria will be applied in selecting correct sum formulas from m/z features [36]. Thirdly, more reaction rules, and how these rules are connected in a pathway, will be learned from metabolic pathway databases ([37,38] for KEGG and MetaCyc).

Supporting Information

Figure S1 “S1 mzGA tutorial.pptx”; tutorial for mzGroupAnalyzer in pptx format. (PPTX)

Figure S2 “S2 MS2 putative cyanidins.pptx”; recorded MS2 product ion scans of the putative new cyanins in pptx format. (PPTX)

Table S1 Rules file for chemical and biochemical transformations. (XLSX)

References

1. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, et al. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18: 1157–1161.
2. Soga T (2007) Capillary electrophoresis-mass spectrometry for metabolomics. *Methods Mol Biol* 358: 129–137.
3. Weckwerth W (2011) Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. *Analytical and Bioanalytical Chemistry* 400: 1967–1978.
4. Scherling C, Roscher C, Giallisco P, Schulze ED, Weckwerth W (2010) Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *Plos One* 5: e12569.
5. Mari A, Lyon D, Fragner L, Montoro P, Piacente S, et al. (2013) Phytochemical composition of *L. analyzed* by an integrative GC-MS and LC-MS metabolomics platform. *Metabolomics* 9: 599–607.
6. Doerfler H, Lyon D, Nagele T, Sun X, Fragner L, et al. (2013) Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics: Official journal of the Metabolomic Society* 9: 564–574.
7. Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, et al. (2012) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*.
8. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, et al. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & cellular proteomics: MCP* 4: 2010–2021.
9. Kind T, Fiehn O (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews* 2: 23–60.
10. Weckwerth W (2011) Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of Proteomics* 75: 284–305.
11. Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* 10: 1551–1558.
12. van der Greef J, Hankemeier T, McBurney RN (2006) Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics* 7: 1087–1094.
13. Mittler R, Vanderauwera S, Gollery M, Van Breusegem F (2004) Reactive oxygen gene network of plants. *Trends in plant science* 9: 490–498.
14. Croteau R, Kuchan TM, Lewis NG (2000) Natural Products (Secondary Metabolites). *Biochemistry and Molecular Biology of Plants*: 1250–1318.
15. Lee KW, Lee HJ, Lee CY (2004) Vitamins, phytochemicals, diets, and their implementation in cancer chemoprevention. *Crit Rev Food Sci Nutr* 44: 437–452.
16. Sun X, Weckwerth W (2012) COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics: Official journal of the Metabolomic Society* 8: 81–93.
17. Kind T, Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* 7: 234.
18. Quenzer T (2002) Automated accurate mass analysis using FTICR mass spectrometry. *Proceedings of the 50th Annual Conference on Mass Spectrometry and Allied Topics*, Orlando, FL.
19. Kimura M, Yamamoto YY, Seki M, Sakurai T, Sato M, et al. (2003) Identification of Arabidopsis genes regulated by high light-stress using cDNA microarray. *Photochemistry and Photobiology* 77: 226–233.
20. Dong CH, Zolman BK, Bartel B, Lee BH, Stevenson B, et al. (2009) Disruption of Arabidopsis CHY1 reveals an important role of metabolic status in plant cold stress signaling. *Molecular plant* 2: 59–72.
21. Huang X, Li Y, Zhang X, Zuo J, Yang S (2010) The Arabidopsis LSD1 gene plays an important role in the regulation of low temperature-dependent cell death. *The New phytologist* 187: 301–312.
22. Nishiyama Y, Yamamoto H, Allakhverdiev SI, Inaba M, Yokota A, et al. (2001) Oxidative stress inhibits the repair of photodamage to the photosynthetic machinery. *The EMBO journal* 20: 5587–5594.
23. Asada K (2006) Production and scavenging of reactive oxygen species in chloroplasts and their functions. *Plant physiology* 141: 391–396.
24. Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annual review of plant biology* 55: 373–399.
25. Hernandez I, Alegre L, Van Breusegem F, Munne-Bosch S (2009) How relevant are flavonoids as antioxidants in plants? *Trends in plant science* 14: 125–132.
26. Seyoum A, Asres K, El-Fiky FK (2006) Structure-radical scavenging activity relationships of flavonoids. *Phytochemistry* 67: 2058–2070.
27. Bashandy T, Taconnat L, Renou JP, Meyer Y, Reichheld JP (2009) Accumulation of flavonoids in an ntra ntrb mutant leads to tolerance to UV-C. *Mol Plant* 2: 249–258.
28. Kimura M, Yamamoto YY, Seki M, Sakurai T, Sato M, et al. (2003) Identification of Arabidopsis genes regulated by high light-stress using cDNA microarray. *Photochem Photobiol* 77: 226–233.
29. Wu Z, Rodgers RP, Marshall AG (2004) Two- and three-dimensional van krevelen diagrams: a graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance mass measurements. *Analytical chemistry* 76: 2511–2516.
30. van Krevelen DW (1950) Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel*: 269–284.
31. Kai K, Takahashi H, Saga H, Ogawa T, Kanaya S, et al. (2011) Metabolomic characterization of the possible involvement of a Cytochrome P450, CYP81F4, in the biosynthesis of indolic glucosinolate in Arabidopsis. *Plant Biotechnology* 28: 379–385.
32. Reemtsma T (2009) Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: status and needs. *Journal of chromatography A* 1216: 3687–3701.
33. Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, et al. (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *The Plant journal: for cell and molecular biology* 42: 218–235.
34. Giallisco P, Hummel J, Lisek J, Inostroza AC, Catchpole G, et al. (2008) High-resolution direct infusion-based mass spectrometry in combination with whole ¹³C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Analytical chemistry* 80: 9417–9425.
35. Rogers S, Scheltema RA, Girolami M, Breidling R (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* 25: 512–518.
36. Kind T, Fiehn O (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* 8: 105.
37. Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. *Nucleic acids research* 30: 59–61.
38. Kanehisa M (2002) The KEGG database. *Novartis Foundation symposium* 247: 91–101; discussion 101–103, 119–128, 244–152.
39. Saito N, Tatsuzawa F, Nishiyama A, Yokoi M, Shigihara A, et al. (1995) Acylated cyanidin 3-sambubioside-5-glucosides in *Matthiola incana*. *Phytochemistry* 38: 1027–1032.
40. Bloor SJ, Abrahams S (2002) The structure of the major anthocyanin in Arabidopsis thaliana. *Phytochemistry* 59: 343–346.
41. Kasai HF, Saito N, Honda T (2011) Structural features of polyacylated anthocyanins using matrix-assisted laser desorption/ionization and electrospray ionization time-of-flight mass spectrometry. *Rapid communications in mass spectrometry: RCM* 25: 1051–1060.
42. March RE, Miao X-S (2004) A fragmentation study of kaempferol using electrospray quadrupole time-of-flight mass spectrometry at high mass resolution. *International Journal of Mass Spectrometry* 231: 157–167.
43. Abad-Garcia B, Garmon-Lobato S, Berrueta LA, Gallo B, Vicente F (2009) A fragmentation study of dihydroquercetin using triple quadrupole mass spectrometry and its application for identification of dihydroflavonols in Citrus juices. *Rapid communications in mass spectrometry: RCM* 23: 2785–2792.
44. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* 42: D199–205.
45. Junker BH, Klukas C, Schreiber F (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC bioinformatics* 7: 109.
46. Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, et al. (2012) VANTED v2: a framework for systems biology applications. *BMC systems biology* 6: 139.

Acknowledgments

We would like to thank the gardeners for the excellent plant cultivation and the members of the Department of Ecogenomics and Systems Biology for the constructive discussions.

Author Contributions

Conceived and designed the experiments: HD XS WW. Performed the experiments: HD XS. Analyzed the data: HD XS WW. Contributed reagents/materials/analysis tools: XS LW DE DL WW. Wrote the paper: HD XS LW DE DL WW.

4. Metabolism dynamics of developing grape berry

Title: Granger causality analysis of metabolite and protein dynamics during Early Campbell grape berry development

4.1 Preamble

Grapevine is a fruit crop with great economic value. In the following work, multivariate statistical analysis of integrated metabolome and proteome data revealed metabolism dynamics as well as interactions and correlations between primary and secondary metabolisms during grape berry development. Metabolic features of developing Early Campbell grape berries were described and compared with other reports.

4.2 Declaration of authorship

The primary and secondary metabolic features of developing grape berries were described and discussed in the following manuscript. This work is mainly my own. I harvested samples, performed experiments, analyzed data and wrote the manuscript.

4.3 Manuscript

Granger causality analysis of metabolite and protein dynamics during Early Campbell grape berry development

Lei Wang¹, Xiaoliang Sun¹, Jakob Weiszmänn¹, Wolfram Weckwerth^{1,2,*}

¹. Department of Ecogenomics and Systems Biology; University of Vienna, Vienna, Austria.

². Vienna Metabolomics Center (VIME); University of Vienna, Vienna, Austria.

*Correspondence:

Wolfram Weckwerth

wolfram.weckwerth@univie.ac.at

Abstract

Grapevine is a popular fruit crop worldwide with essential economic importance. The grape berry undergoes complex biochemical changes from fruit set until ripening. To better understand this dynamic process, we applied mass spectrometry based platforms to analysis the metabolome and proteome of grape berries at 12 developmental stages covering the whole developmental process of grape berries. Primary metabolites involved in central carbon metabolism such as sugars, organic acids and amino acids metabolism together with various bioactive secondary metabolites like flavonols, flavan-3-ols and anthocyanins were annotated and quantified. At the same time, the proteomic analysis revealed the protein dynamics of the developing grape berries. Multivariate statistical analysis of the integrated metabolomic and proteomic dataset revealed a growth trajectory of the developmental process and distinctions among different stages. The k-means clustering analysis and Granger causality analysis presented the specificities of the metabolism profile of each stage and directed metabolism correlations between them. The incorporation of the metabolome and proteome profiles together with their corresponding metabolism pathway information revealed the dynamics of sugar, glycolysis, TCA cycle, amino acid, phenylpropanoid, flavonoid biosynthetic pathways. The overview of the metabolism dynamics on both protein and metabolite level unveiled the metabolism switch and adjustments during grape berry development.

28

29 **Keywords:** *Vitis vinifera*, berry development, mass spectrometry, primary metabolism,
30 secondary metabolism, flavonoids

31

32 **Running title:** Metabolome and proteome of developing grape berries

33

1. Introduction

Grapevine (*Vitis vinifera* L.) is one of the most important and widely cultivated economic crops. Grape berries are consumed either as fresh fruit or processed to raisins, juice and wine. Besides its enormous economical and nutritional values, grapes and grape productions possess a wide variety of health benefits, such as antioxidation (Sánchez-Moreno et al., 1999;Doshi et al., 2006;Sánchez-Alonso et al., 2007;Sáyago-Ayerdi et al., 2009;Anastasiadi et al., 2010), cardiovascular protection (Tebib et al., 1994;Adisakwattana et al., 2010;Razavi et al., 2013), neuroprotection (Feng et al., 2007), anti-obesity (Kim et al., 2013;Zhang et al., 2013), etc.

The grape berry is a non-climacteric fruit. From fruit set to ripening, grape berries undergo three main developmental phases including two sigmoidal growth phases with an intermediate lag phase (Kennedy, 2002). The performance of grape berry development is characterized by dramatic changes in both physiology and biochemistry, including increases in volume and weight, changes in texture, color, aroma, acidity, sugar contents, susceptibility, etc. The first growth phase (phase I) is characterized by fruit formation and enlargement due to the active cell division and expansion. In this phase, a notable accumulation of organic acids, especially malic and tartaric acid was observed (Conde et al., 2007). Phase II, which is defined as a lag phase features a slow enlargement of berry volume caused by a stop in cell division. The grape berry is still green and hard at this phase. Organic acids continuously accumulate until veraison, which marks the beginning of phase III. During the last phase, the grape berries undergo a second sigmoidal growth accompanied by a decrease in acidity and increase in sugar content (Conde et al., 2007;Deluc et al., 2007;Fortes et al., 2011;Liang et al., 2011;Dai et al., 2013;Degu et al., 2014;Fraige et al., 2015;Cuadros-Inostroza et al., 2016). The peel of red varieties colors as a result of the accumulation of anthocyanins (Boss et al., 1996;Ali et al., 2011b;Degu et al., 2014;Fraige et al., 2015). The grape berry becomes soft in the final phase and is ready to be harvested. Another generally adopted descriptive system is the E-L system which was proposed firstly by Eichhorn and Lorenz (Eichhorn and Lorenz, 1978) with a more detailed description of grape berry development stages.

Fruit development is an intricate process, featuring complex regulation and fine-tuned changes in metabolism. Its analysis requires the use of sensitive methods, which allow high sample throughput to cope with the amount of samples necessary to examine a time continuous process.

Since the release of the grapevine genome sequence in 2007 (Jaillon et al., 2007; Velasco et al., 2007), studies of developing grape berry based on transcriptomic (Deluc et al., 2007; Palumbo et al., 2014), proteomic (Giribaldi et al., 2007; Negri et al., 2008; Martinez-Esteso et al., 2011; Fraige et al., 2015), metabolomic (Ali et al., 2011a; Dai et al., 2013; Degu et al., 2015) techniques contributed extensively to our understanding of berry growing and ripening process. These studies not only enhanced and supplemented the morphological and physiological descriptions but also promoted the work to molecular level. Exploring the developmental process basing on a single level data results in a partial view of the progress. Several studies described the developmental process by combining transcriptomic and metabolomic profiles (Fortes et al., 2011; Agudelo-Romero et al., 2013; Degu et al., 2014). Considering proteome and metabolome are closer to the phenotype of plants, integration of these data will provide additional information for the understanding of the morphological and physiological changes. Nonetheless, systematic analysis of integrated metabolome and proteome profiles of developing grape berries is still absence so far. It is also problematic to schematize the metabolism dynamics of developing grape berry by summarizing or comparing those studies due to the coverage limitation of either developmental stages or metabolism branches. For instance, some studies only focus on primary metabolism (Dai et al., 2013) whereas some only target on flavonoid accumulation during grape berry ripening (Ali et al., 2011b). Zamboni and coworkers integrated the transcriptomic, proteomic and metabolomic data of developing grape berry into a network for the identification of putative, stage-specific biomarkers (Zamboni et al., 2010). However, most of the suggested candidates were not sequenced or identified. In addition, some researches worked on individual parts of grape berries, such as skin (Negri et al., 2008; Ali et al., 2011b; Degu et al., 2014; Wu et al., 2014; Degu et al., 2015) or berries depleted of seed or peel (Martinez-Esteso et al., 2011; Fang et al., 2013; Fraige et al., 2015) instead of whole berries. Therefore it is necessary to take the tissue into account when comparing the results.

In this study, we harvested samples according to the modified E-L system (Coombe, 1995) from fruit set to ripening at 12 time points. Mass spectrometry based high-throughput platforms were applied for the metabolomic and proteomic analysis of both primary and secondary metabolism dynamics of developing grape berries. Multivariate statistical analysis of the dynamics of metabolites and proteins involved in primary metabolism i.e. glycolysis, tricarboxylic acid (TCA) cycle, amino acid metabolism as well as secondary metabolism i.e. flavonol, flavan-3-ols, anthocyanins, lignin unveiled metabolism interactions during the berry growing period. The holistic approach will contribute to advance our comprehension about grape berry development.

98 2. Material and Methods

99 2.1 Sample collection

100 Berries at 12 developmental stages corresponding to EL 27, 29, 30, 31, 32, 33, 34, 35, 36, 37,
101 37.5 (to distinguish with the samples at early EL 37 stage), 38 (**Figure 1**) were harvested
102 according to the modified Eichhorn-Lorenz system (E-L system) (Eichhorn and Lorenz,
103 1978;Coombe, 1995) from *V. vinifera* (Early Campbell) growing in the plant garden of
104 University of Vienna (48°13'50.2"N 16°21'28.2"E) during the 2014 growing year. Three
105 biological replicates each containing 5 to 10 berries were collected for each developmental
106 phase. The harvested berries were frozen in liquid nitrogen immediately and stored at -80 °C.

107 2.2 Metabolite and protein extraction

108 An integrative extraction of metabolites and proteins was performed according to a universal
109 extraction protocol (Weckwerth et al., 2004) with some modifications. The grape berries were
110 ground to fine powder in liquid nitrogen using mortar and pestle. 50 to 100 mg of material was
111 extracted with 750 µl of extraction solution (methanol: water: formic acid = 70:28:2) and 250
112 µl of hexane. The mixture was homogenized by vigorous vortexing and incubated 30 min on
113 ice. Then the mixture was centrifugated at 20 000 g for 8 min to separate the lipophilic and
114 hydrophilic phases which were subsequently transferred into new tubes, respectively. The
115 extraction procedure was repeated once with the lipophilic and hydrophilic phases pooled
116 together with those from the first extraction, respectively. The extractions were dried under
117 vacuum. Proteins were extracted from the residue pellets according to a previous protocol
118 (Noah et al., 2013).

119 2.3 Metabolite measurement, identification and quantification

120 The dried hydrophilic phases were re-dissolved in 400 µl of 50% methanol. For the primary
121 metabolite analysis, 25 µl of this re-dissolved hydrophilic phase was dried under vacuum and
122 subsequently derivatized according to a modified protocol (Weckwerth et al., 2004;Mari et al.,
123 2013). Agilent® 6890 gas chromatograph coupled to a LECO Pegasus® 4D GC×GC-TOF
124 spectrometer was used for the primary metabolite measurement. Instrument parameters were
125 set as described previously (Doerfler et al., 2013). The metabolite identification and

quantification was performed with LECO Chroma TOF®. Retention times (RTs) of the peaks were converted to retention indices (RIs) according to the RTs of spiked alkanes (C12-C40). Metabolites were annotated by comparing their RIs and mass spectra to those of standards in the GMD Golm database (Kopka et al., 2005) with a minimum match factor set to 700. The peak areas of the annotated metabolites corresponding to specific masses were extracted and used for relative quantification. Mixtures of standard compounds were measured under the same conditions at different concentrations to calculate the standard curves for absolute quantification.

For the secondary metabolite analysis, 10 µl of the re-dissolved hydrophilic phase was mixed with 2.5 µl of reserpine (5 mg l⁻¹), 10 µl of 1.0% formic acid (FA) solution and 77.5 µl of water. After centrifugation at 20 000 g for 8 min, 5 µl of the supernatant were loaded on Waters ACQUITY UPLC HSS T3 nanoACQUITY Column (particle size 1.8 µm, dimension 100 µm×100 mm) via a HTC PAL Autosampler device coupled to an Eksigent nano LC pump and eluted with a non-linear gradient (Mari et al., 2013) at a constant flow rate of 500 nl min⁻¹. Ionization was performed by a nano ESI source (Thermo Scientific, USA) in positive mode with the masses analyzed by a LTQ Orbitrap XL™ mass spectrometer (Thermo, Germany). Each full scan was followed by one MS/MS scan with the most abundant precursor ion fragmented by collision induced dissociation (CID) under 35% of the normalized collision energy during 90 ms activation time. The minimum signal threshold was set to 50 000. For the secondary metabolite identification, accurate precursor masses, formulae, RTs together with mass accuracy were exported from Xcalibur (Thermo Xcalibur 2.2 SP1.48) and compared with the information from literature or standard compounds. The annotation levels was marked according to a standard proposed by the Metabolomics Standards Initiative (Sumner et al., 2007). LCquan (Thermo, v2.6.6.1128) was used for peak isolation and peak area integration.

2.4 Protein digestion and analysis

Protein concentration was determined by the Bradford method (Bradford, 1976) with a BSA standard curve. 100 µg of protein were firstly reduced with dithiothreitol (DTT, 5 mM, 37°C, 45 min); then alkylated with iodoacetamide (IAA, 10 mM, 23°C, dark, 60 min) and finally added with addition of 5 mM of DTT (23°C, dark, 15 min). Endoproteinase LysC and trypsin were applied for digestion based on a previous protocol (Hoehenwarter et al., 2008). After digestion, samples were desalted with C18-SPEC-96 well plate (15 mg, Agilent, USA) according to the manufacturer's instruction. The eluted peptides were dried under vacuum and

dissolved in 500 µl of start gradient solution (4.5% acetonitrile, 0.1% FA). 1 µg of digested protein was loaded on an Ascentis Peptide ES-C18 column (particle size 2.7 µm, dimension 15 cm × 100 µm, Sigma-Aldrich, USA) and eluted with a 90 min linear gradient from 5% to 40% of mobile phase B (90% acetonitrile, 0.1% FA; phase A, 0.1% FA in water) at a constant flow rate of 400 nl min⁻¹. The same ESI-LTQ-Obitrap equipment used for metabolite analysis was applied for peptide measurement. Each full scan was followed by 10 MS/MS scans in which the 10 most abundant ions were selected and fragmented by CID with 35% of the normalized collision energy during a 30 ms activation time. Minimum signal threshold was set to 10 000.

The obtained raw files that containing peptide information were searched against a grape fasta file including 65 448 protein sequences from UniProt with the SEQUEST algorithm in Proteome Discoverer (v 1.3, Thermo Scientific). Searching parameters were set as below: maximum two missed cleavage sites, acetylation for N-terminal modification, oxidation of methionine for dynamic modification and carbamidomethylation of cysteine for static modification were allowed. Mass tolerance for precursors was set to 10 ppm and for fragment masses to 0.8 Da. False discovery rate (FDR) was set to 0.01. Protein candidates were defined by at least two peptides with high confidence. The obtained raw files and sequence information of the identified proteins were submitted to the public repository ProteomeXchange (Vizcaino et al., 2014) with the dataset identifier PXD003769 (<http://www.proteomexchange.org/>, Username: reviewer36585@ebi.ac.uk, Password: iEFYHERU). Normalized spectral abundance factors (NSAFs) were calculated (Zybailov et al., 2006) for relative quantification. The protein candidates that present in all the three biological replicates of at least one stage were considered for the statistical analysis. For the functional analysis, the identified protein sequences were blasted against a protein database of *Arabidopsis thaliana* (from PLAZA with 27 407 protein sequences) and *Theobroma cacao* (from PLAZA with 44 404 protein sequences) with the BLASTP function in NCBI (v 2.2.31, <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>). A cacao mapping file from GoMapMan (“tca_Phytozome9.1_transcript_2015-01-09_mapping.xlsx”, <http://www.gomapman.org/export/current/mapman>) was applied for the functional analysis.

2.5 Statistical analysis

The obtained metabolite data were normalized to fresh weight and dodecane (C12 alkane, GC-MS data) or total ion intensity (LC-MS data). Analysis of variance (ANOVA) and k-means clustering were performed within Matlab® (V8.4.0 R2014b; <http://www.mathworks.com>). The

significant levels of the candidates were presented with lower case letters according to results of the Duncan's test (Duncan, 1955). K-means clustering analysis was repeated 100 times and finally the result with minimal total distance was selected. Principal component analysis (PCA), hierarchical clustering analysis and Granger causality analysis were performed with COVAIN under Matlab environment (Sun and Weckwerth, 2012). Granger causality analysis was performed on all the identified metabolites and proteins as well as the clusters after k-means clustering analysis with their time lag was set to 1 and 1, 2, 3, respectively. The correlations with p values less than 0.05 were recorded. The network of Granger result was visualized in Cytoscape (<http://www.cytoscape.org/>). The Venn diagram was drawn with Venny 2.0.2 (<http://bioinfogp.cnb.csic.es/tools/venny/>).

3. Results and Discussion

3.1 Metabolomic profiles of developing grape berry

The GC-TOF-MS platform allowed the annotation of 87 candidates including sugars, amino acids, organic acids together with simple amine and phenolic compounds, according to their RIs) and mass spectra. Additionally, 49 flavonoids were annotated from LC-Orbitrap-MS data according to the accurate precursor masses, formulae and their fragmentation patterns (**Table 1**). The detailed information (RI and RTs, quantification masses or MS2 fragments and the integrated peak area, one way ANOVA results) of the annotated candidates is listed in **Table S1**. The dynamic patterns of the annotated metabolites were visualized by a hierarchical bi-clustering color map (**Figure 2A**). Sugars (including sugar alcohols, sugar acids), amino acids, organic acids and flavonoids were further shown in **Figure 2 B, C, D and E**.

Sugars

The main sugars in grape berries are fructose, glucose and sucrose. In the present study, fructose constantly accumulated during development with significant increases before veraison and during ripening (**Figure 2B**). Glucose also significantly accumulated around veraison but declined afterwards (**Figure 2B**). The decrease of glucose after veraison has been observed in table grapes, i.e. "Thompson Seedless", "Crimson Seedless" and "Red Globe" (Muñoz-Robredo et al., 2011). In contrast to these findings, both glucose and fructose concentration constantly increased during grape berry development in some grape varieties and cultivars (Wu

et al., 2011;Dai et al., 2013). The discrepancy in glucose accumulation patterns could be explained by the differences in the ripening process among varieties. The content of sucrose fluctuated during grape development (**Figure 2B**) with four inflection points at EL 30, 32, 34 and 36 respectively. Other sugars, sugar alcohols and sugar acids either decreased during development (ribose, xylose, myo-inositol, arabinose, rhamnose, galactaric acid) or featured the highest level at EL 32 (threonic acid, galactonic acid, gluconic acid, erythriol) (**Figure 2B**).

Amino acids

Amino acids showed distinct dynamics during grape berry development (**Figure 2C**). Arginine and asparagine were the most abundant amino acids in young berries whereas alanine and glutamine in mature berries (**Table S2**). Lysine, tyrosine, arginine, ornithine and phenylalanine increased significantly from EL 27 to EL 29, however, decreased dramatically until the end of lag phase (EL 34) and remained at a relatively low level during the second sigmoidal growth period (**Figure 2C**). Asparagine was in high abundance at the first two developing stages followed by a dramatic decline from EL 29 to EL 30 then stayed in low level until the end. Other amino acids fluctuated during grape berry developing and all showed a turning point at EL 34 which is the end of the lag phase and the beginning of the veraison (**Figure 2C**).

Organic acids

The predominant organic acids detected in grape berry were malic acid, tartaric acid and citric acid which increased before veraison (EL 35) and decreased afterwards (**Figure 2D**). The changing patterns of these dominant organic acids were consistent with previous reports (Deluc et al., 2007;Ali et al., 2011a;Muñoz-Robredo et al., 2011;Fraige et al., 2015). Other organic acids showed similar changing pattern except pyruvic acid, gallic acid and caffeic acid which were highest in the young berries then decreased throughout the developmental process (**Figure 2D**). In grape, organic acids are responsible for the titratable acidity which is an index for fruit quality. High amount of organic acids endow young berries sour taste for defense against herbivores. The organic acids in mature berries are essential for wine production as they protect the fermentation process from bacterial contamination. Those in wine are responsible for the sour taste. They are also essential for the color of wine by contributing to the stabilization of anthocyanins (Clemente and Galli, 2011).

Flavonoids

Grape and its products are rich in polyphenolics. These secondary metabolites, especially flavonoids, play multiple roles in grape and attract more and more attentions due to their health benefits (Anastasiadi et al., 2010; Kim et al., 2013; Zhang et al., 2013). During grape berry development, the detected flavonoids presented two distinct changing patterns (**Figure 2E**). All the anthocyanins accumulated during ripening whereas most of the candidates in the other subfamilies like proanthocyanins, flavan-3-ol, flavonol, flavanone and their glycosides were abundant in young berries and decreased during the time course of development (**Figure 2E**). Parallel phenomena were observed before in other varieties, i.e. Cabernet Sauvignon (Ali et al., 2011b; Degu et al., 2014), Shiraz (Degu et al., 2014) and Norton (*V. aestivalis*) (Ali et al., 2011b) and supposed to be caused by the competition of precursors between anthocyanins and other subfamily members of flavonoids (Ali et al., 2011a; Degu et al., 2014). The synthesis of anthocyanins splits into three branches, i.e. the monohydroxylated (pelargonidin, Pg), the dihydroxylated (cyanidin, Cy), and the trihydroxylated (delphinidin, Dp) branch. Cy and peonidin (Pn) glycosides which belong to the dihydroxylated branch were detected from stage EL 37 or even EL 36 whereas derivatives of the other two branches started to appear one stage later. Furthermore, in mature berries (EL 38), the relative abundance of Cy- and Dp- derivatives were higher than the corresponding derivatives of other aglycones (**Figure S1**). For instance, Cy-Cou-diGlc (**30**) and Dp-Cou-diGlc (**25**) were more abundant than petunidin- (Pt, **29**), malvidin- (Mv, **32**) and Pg- (**33**) coumaroyl-diglucoside; Cy-Glc (**2**) and Dp-Glc (**8**) were in higher level than glucoside of other aglycones (**Figure S1**). It was also observed in “Concord” (Liang et al., 2011) and “Pink Sultana” (Boss et al., 1996) that Cy and Dp derivatives were the dominant anthocyanins. In contrast, Mv derivatives were the most abundant anthocyanins in the other varieties in these two studies (Boss et al., 1996; Liang et al., 2011) and other reports (Mazza et al., 1999; He et al., 2010; Papini et al., 2010; Ali et al., 2011b; Degu et al., 2014; Degu et al., 2015). Noticeably, Early Campbell a hybrid of *Vitis vinifera* and *Vitis labrusca* by crossing Moore Early with (Belvidere × Muscat of Hamburg). Both Moore Early and Belvidere are seedlings of Concord. The fruit taste and disease resistance of Early Campbell are similar to Concord (Robinson et al., 2012). The similar anthocyanin profile of Early Campbell with that of Concord is probably also due to its genetic background.

3.2 Protein profiles of developing grape berry

In total, 1313 proteins were identified from all the samples (for sequences information see in **Table S3**). 848 candidates prevalent in all replicates of at least one stage were used for further statistical analysis. NSAFs and the ANOVA analysis result can be found in **Table S4**. The

functions of all the protein candidates were annotated by blasting against protein sequences of *A. thaliana* and *T. cacao*. The matching with the *T. cacao* database yielded a higher amount of hits with a better blast quality and was therefore used for functional analysis (see **Table S4**). Subsequently, the identified protein candidates were assigned to corresponding functional bins according to the cacao mapping file from GoMapMan (**Table S4**).

Hierarchical bi-clustering analysis was applied to visualize the dynamic proteome profiles of developing grape berry (**Figure S2**). Samples of 12 developing stages were clustered into 3 groups indicated with color blue, red and green (**Figure S2**). Samples from stage EL 27, 29, 30 and 31 were assigned to group 1; EL 32, 33, 34, 35 group 2 and EL 36, 37, 37.5, 38 group 3. The Venn diagram (**Figure 3A**) shows 394 proteins were common to all groups and 157, 43 and 80 protein candidates are specific to group 1, 2 and 3, respectively. The functional distribution of these group specific proteins was summarized with pie charts (**Figure 3A**). There were 18 amino acid metabolism related proteins that showed exclusively in group one which accounting for 5.83% of all the group 1 specific proteins whereas only 1 and 2 proteins which accounting for 1.61% and 1.43% respectively of specific to group 2 and 3 (**Figure 3A**) indicating active amino acid metabolism at phase I on proteome level. There were 8 transport related proteins accounting for 2.59 % of group 1 specific proteins and 5 accounting 3.57 % of group 3 specific proteins whereas there was no transport related protein exclusive to group 2 (**Figure 3A**). Another notable point is that more proteins associated with secondary metabolism existed exclusively in group 1 (12, 3.88 %) and 3 (12, 8.57 %) than in group 2 (2, 3.23 %) (**Figure 3A**) indicating the synthesis of secondary metabolites was more active in the beginning and the end of berry developmental stages than in the middle.

The annotated protein candidates were assigned to 50 functional bins. The majority functional categories include candidates involved in protein synthesis (11.41%), protein degradation (7.11%), RNA regulation of transcription (6.09%), signaling (4.38%), and abiotic stress response (4.30%) (**Figure 3B**). The changing patterns of the proteins in these functional categories were summarized to four groups by hierarchical clustering analysis (**Figure 3C**) with their summarized changing patterns shown on the right side. Proteins in 30 out of 50 functional bins were in lowest content around veraison (**Figure 3C** in red and green). In contrast, proteins involved in 7 functional groups were with highest abundance just before veraison (EL 33) (**Figure 3C** in blue). The functional bins in purple group (**Figure 3C**) involve proteins constantly accumulated during grape berry development.

Proteins related to abiotic and biotic stresses (4.30% and 2.03% respectively) showed high abundance at early developmental stages or/and during ripening (**Table 3B, Figure 3C** in red) indicating high resistance ability of grape berry against environmental and developmental stresses during these developmental phases. A larger amount of oxidative stress responsive proteins in young, green berries as well as increasing expression of pathogen responsive proteins after veraison have been previously reported (Giribaldi et al., 2007). Additionally, another study reported a parallel transcript profile of stress/pathogens responsive gene strongly expressed in ripening berries (Davies and Robinson, 2000).

Photosynthesis (6.1%) is another major functional category including candidates involved in light reaction (2.97%), photorespiration (0.86%) and calvin cycle (2.27%) (**Figure 3B**). Proteins involved in light reaction were in higher abundance in the earliest stage and around veraison (**Figure 3C** in blue) whereas those involved in photorespiration and calvin cycle were more abundant in young green berries. The levels of proteins in all of these three subgroups declined after veraison. The decrease in abundance of photosynthesis related proteins throughout grape berry development especially after veraison was consistent with previous proteomic studies (Martinez-Esteso et al., 2011; Fraige et al., 2015) and the physiological situation (Pandey and Farmahan, 1977) of developing grape berries.

High frequency of proteins associated with lipid metabolism was observed (3.1%, **Figure 3B**) increasingly expressed after veraison (**Figure 3C**). Proteins associated with secondary metabolism showed distinct changing patterns. Those involved in phenylpropanoids synthesis were in high abundance in young berries then decreased during development (**Figure 3C** in red) whereas those related with flavonoids and isoprenoids synthesis proteins were strongly accumulated after veraison (**Figure 3C** in purple). Proteins associated with synthesis of N and S containing metabolites were highly expressed around veraison. The distinct arrangement of protein expression reflected the metabolism adjustment during grape berry development.

3.3 Metabolome and proteome data integration

3.3.1 Multivariate statistical analysis reveal trajectory of grape berry development

The PCA plot (**Figure 4A**) of the integrated metabolomics and proteomic dataset reveals a continuous trajectory during grape berry development. The separation of various developmental stages indicating their metabolism distinctions on metabolite and protein levels. The first principal component (PC 1) accounting for 47.44% (**Figure 4A**) of the total variance

characterized metabolic and proteomic specificities of grape berries at different developmental stages. Candidates with high absolute loading scores (**Table S5**) included metabolites, especially flavonoids, caffeic acid, gallic acid, lysine, asparagine, arginine and methionine together with protein candidates involved in development, lipid metabolism, cell wall construction, TCA cycle and protein degradation which accounted most of the separation among developmental stages.

Further, k-means clustering analysis was applied to group candidates according to their changing patterns. **Figure S3** presents the 12 clusters with a bold red line indicating the averaged pattern of all the candidates in each cluster. Cluster 2 and 6 with 86 and 124 candidates, respectively, present candidates with higher abundance at early developmental stages (EL 27 to 31) (**Figure 4B** and **S3**). Candidates in these two clusters include amino acids (methionine, phenylalanine, asparagine, ornithine, arginine, lysine, tyrosine), organic acids (caffeic acid, pyruvic acid, gallic acid), sugars and sugar alcohols (arabinose, rhamnose, myo-inositol), (dihydro)flavonol derivatives (epicatechin, 3, 20, 36, 49, 12, 31, 38, 39, 43, 45, compound number see in **Table 1**) and proteins involved in amino acid metabolism (**Table S6**). The candidates in cluster 10 and 9 are with highest abundance at EL 32 and veraison (EL 34 and 35), respectively. The most abundant organic acids, i.e. malic acids and citric acids, together with some proteins involved in photosynthesis in cluster 5 were in a relatively higher levels during the lag phase and veraison than during the two sigmoidal growth phases. In contrast, those candidates in cluster 3 and 4 showed opposite dynamic patterns including organic acids (shikimic acid, ascorbic acid, fumaric acid, maleic acid), amino acids (valine, leucine, proline isoleucine) and proteins involved in stress response (**Figure S3**, **Table S6**). All the annotated anthocyanins and four flavonol derivatives (5, 10, 21, 26, compound number see in **Table 1**) together with high abundance of proteins involved in secondary metabolism and lipid metabolism in cluster 1, 8 and 12 mainly accumulated after veraison (EL 36 to 38) (**Figure S3**, **Table S6**). The candidates in cluster 7 show a continuous elevation in their abundance whereas those in cluster 11 decreased (**Figure S3**, **Table S6**).

To extent the understanding of the dynamic correlations of all the identified metabolites and proteins during the developmental time course, Granger causality analysis was applied to all the candidates and clusters. Granger causality analysis was firstly introduced by Clive Granger (Granger, 1969) to predict event based on time series data in economics. It was also applied to some biological studies to interpret directed and nonlinear correlations between metabolites and proteins (Doerfler et al., 2013; Valledor et al., 2014). Here the Granger causality analysis results

indicated significant metabolite-metabolite, metabolite-protein and protein-protein correlations. **Figure 4C** shows the directed correlation of phenylalanine and phloretin (p value = 0.02187) which indicating strong effect of phenylalanine concentration on phloretin synthesis. Not only precursors but also enzymes showed significant causal correlation to their product synthesis. One example is the prior accumulation of anthocyanin synthase (ANS, A2ICC9) to anthocyanins (**Figure 4D, Table S7**). **Figure 4E** presents a significant granger correlation (p value = 0.00538) between protein E0CSB6 (malate dehydrogenase) and D7T300 (adenylpyrophosphatase, ATPase) indicating the close relationship between TCA cycle and electron transport in respiration process. Other time series correlations with p values less than 0.05 were summarized in **Supplementary table S7**. The Granger-based network (**Figure S4**) includes 674 nodes (**Table S7**) with 21 neighbors in average. Many amino acids (ornithine, arginine, phenylalanine, lysine, tyrosine, asparagine), organic acids (acid_like2, shikimic acid) and their metabolism related proteins (**Table S7**) show highest node degrees indicating their essential metabolism position during grape berry development. The application of Granger analysis to the 12 clusters obtained from the k-means clustering analysis indicated some time lagged correlations between these clusters. Significant directed connection among these clusters were present in **Figure F, G and H** with the time lag set as 1, 2 and 3, respectively. The Granger causality correlations from clusters 2 to 6, 6 to 10, 10 to 9, 9 to 12 and 12 to 8 (**Figure 4G** indicated with red arrows) were visualized with a line chart (**Figure 4B**) clearly showing their dynamic shift over the developmental time course. Such combined Granger causality analysis with clustering analysis indicated a general metabolism shift from the metabolism of amino acids, sugars and some flavonoids to organic acids accumulation and finally to lipid and anthocyanin synthesis.

PCA and k-means clustering analysis presented the dynamics of the metabolites and proteins during grape berry development. To understand the metabolism progress of developing grape berry in a biochemical context, we mapped the sugars, amino acids, organic acids, flavonoids and the related proteins on their corresponding biosynthetic pathways (**Figure 5**). The integration of dynamics of metabolites as well as proteins involved in both primary and secondary metabolisms presented metabolism adjustments and interactions during grape berry development.

3.3.2 Primary metabolism dynamics

410 Primary metabolism plays an essential role in grape berry development. The products from
411 primary metabolism pathways are not only crucial for grape to survive but also endow grape
412 berry specific characters which further deciding its market values.

413 ***Sugar metabolism***

414 Sugars, especially fructose, glucose and sucrose determine the sweetness of grapes, moreover,
415 the alcohol concentration of wine. In grape berries, sucrose is mainly imported via phloem from
416 source organs. Subsequently, the imported sucrose is either hydrolyzed to glucose and fructose
417 by invertase or converted to glycolysis substrates via sucrose synthase (Susy) and UDP-glucose
418 pyrophosphorylase (UGPase). The fluctuation of sucrose content during grape berry
419 development might be caused by a disproportionate ratio of the import to the consumption.
420 Furthermore, synthesis of sucrose from malate via the gluconeogenic pathway (Ruffner et al.,
421 1975; Dai et al., 2013) might also contribute to the fluctuation of sucrose concentration,
422 especially after veraison. The accumulation of glucose and fructose during grape berry
423 development was reported previously (Wu et al., 2011; Dai et al., 2013). In the present study,
424 fructose accumulated throughout the developmental process whereas glucose concentration did
425 not continue to rise after veraison. Similar glucose dynamic was also observed in some table
426 grape varieties (Muñoz-Robredo et al., 2011). The distinct expression pattern of invertase and
427 Susy might explain the unequal accumulation of glucose and fructose. The decline in abundance
428 of invertase since EL 31 caused a decrease in the production of glucose and fructose whereas
429 the increasing expression of Susy ensured the continuous accumulation of fructose.

430 ***Glycolysis***

431 The substances generated from sugar metabolism subsequently incorporated into glycolysis
432 pathway. Metabolism along this process generates energy (ATP), reductant (NADH) as well as
433 intermediates for amino acids, lipids and secondary metabolites production. The abundance of
434 most glycolytic enzymes increased through the development and ripening process.
435 Phosphoglycerate kinase (PGK) was the only glycolytic enzyme, which declined in abundance.
436 The expression of pyrophosphate-fructose 6-phosphate 1-phosphotransferase (PFP), enolase
437 and pyruvate kinase (PK) in cytosol strongly increased after veraison (**Figure 5**).
438 Phosphoglycerate mutase (PGM) increasingly expressed during the first sigmoidal growing
439 phase (EL 27 to EL 32) and maintained in a relatively constant level afterward. The increase in
440 abundance of glycolytic proteins was consistent with some former reports that studied the

proteomic profile of grape skins and berries removed seeds (Negri et al., 2008;Kambiranda et al., 2014). However, some studies reported decrease in abundance of glycolytic enzymes (Davies and Robinson, 2000;Giribaldi et al., 2007;Martinez-Esteso et al., 2011) or glycolytic intermediates (Dai et al., 2013) during berry ripening. Variety and differences in growth conditions might explain these different observations. Additionally, isoforms of an enzyme may play different roles at a particular time or situation, thus displaying varying dynamics during development. For instance, Fraige et al reported three isoforms of UDPase, of which two candidates decreased in abundance after veraison whereas one increased (Fraige et al., 2015).

TCA cycle

Pyruvate generated from glycolysis follows the respiratory process to incorporate into TCA cycle. Metabolism along TCA cycle generates energy, reducing power and carbon skeletons endowing TCA cycle central importance to many metabolism pathways. Pyruvate dehydrogenase complex (PDC) converts pyruvate to acetyl-CoA which serves as fuel to enter TCA cycle. During the grape berry development, PDC was in high abundance during the first growing phase (EL 27-32) followed by a significant decline at the lag phase (EL 32-34) and with a subsequent increase (**Figure 5**). Similar to PDC, isocitrate dehydrogenase (IDH), oxoglutarate dehydrogenase complex (OGDC), and succinyl coenzyme A synthetase (SCS) were in high abundance during the sigmoidal growth phases whereas in low abundance during the lag phase (**Figure 5**). The high abundance of the enzymes in the young and ripening berries were consistent with the great demand of energy and building blocks at these two phases. Aconitase and succinate dehydrogenase (SDH) strongly expressed after veraison. In a former report, a sharp expression of aconitase was observed in ripening grape skin (Negri et al., 2008). Fumarase is the only enzyme whose expression gradually declined throughout the berry development (**Figure 5**). Malate dehydrogenase (MDH), catalyzing a reversible reaction between oxaloacetate and malate, decreasingly expressed at phase I whereas progressively up-expressed during ripening which was agree with previous reports (Martinez-Esteso et al., 2011;Kambiranda et al., 2014). The transcript levels of MDH and malic enzyme were reported to increase during grape berry ripening which might contribute to the decline of malate concentration after veraison (Deluc et al., 2007).

The contents of the intermediates, citrate, succinate, fumarate and malate gradually increased in early stages of development, up to a peak in concentration EL 31 (fumarate and succinate), EL 33 (malate) or around veraison (citrate) with a subsequent decline. Similar dynamic patterns

were observed in developing Cabernet Sauvignon berries with a peak in accumulation of most TCA cycle intermediates at veraison (Dai et al., 2013). The accumulation of these organic acids before veraison was parallel to the high abundance of enzymes at phase I. However, the gradually increasing expression of TCA cycle associated enzymes was accompanied by a decrease in the intermediates during grape berry ripening (phase III). The discrepancy between the increase in the abundance of the enzymes and the decrease in the content of the intermediates during grape berry ripening might be due to the incorporation of the intermediates in the synthesis of amino acids, lipids and secondary metabolites.

Amino acids metabolism

Amino acids are major transportable nitrogenous compounds in grape. In source organs, intermediates from glycolysis and TCA cycle can be utilized as precursors for the synthesis of amino acids. For instance, phosphoenolpyruvate is the precursor of aromatic amino acids that derived from shikimate pathway; α -ketoglutarate and oxaloacetate are the precursors of glutamate and aspartate family amino acids, respectively. These biosynthesized amino acids together with those assimilated from soil via roots are subsequently transported to sink organs. Asparagine and glutamine were the major amino acids in young and mature grape berries respectively. Both of them carry an extra amide group making them efficient nitrogen-carriers. They play important roles in the nitrogen assimilation, transportation and storage in plants. Asparagine and glutamine can be converted to Asp and Glu to serve as precursors for biosynthesis of many other amino acids, e.g. proline, arginine (from glutamate); methionine, threonine and lysine (from aspartate). Several enzymes involved in amino acid metabolism were detected and mapped on **Figure 5**. Aspartate aminotransferase (AspAT) catalyzes the reversible transfer of an amino group between aspartate and glutamate thus plays important role in nitrogen distribution. In parallel with a previous report (Martinez-Esteso et al., 2011), AspAT abundancy gradually decreased during grape berry development. Methionine synthase (MetH), catalyzing the synthesis of methionine, is another essential amino acid of aspartate family. The abundance of MetH increased until veraison then stayed relatively constant during ripening in the present study. However, it decreased during green developmental stages and increased during ripening in the study of Martinez-Esteso et al (Martinez-Esteso et al., 2011). Glutamine synthetase (GS) is another crucial enzyme involved in nitrogen assimilation. GS that catalyzes the condensation of glutamate and ammonia to generate glutamine gradually declined in abundance from EL 27 to EL 36 and slightly increased afterwards. The transcript level of GS was shown to be significantly higher in phase I in a previous study (Deluc et al., 2007). Martinez-

Esteso et al reported a decline in the level of GS before veraison (Martinez-Esteso et al., 2011) which is consistent with our result. However, the changing trend of GS after veraison was absent in their study.

3.3.3 Secondary metabolism dynamics

Grape is rich in polyphenolic compounds that derived from phenylpropanoid pathway. These bioactive secondary metabolites play essential roles in protecting grape berry against abiotic stresses such as UV radiation (Pontin et al., 2010), high light and high temperature stresses (Ayenew et al., 2015), drought stress (Król et al., 2014) as well as biotic stresses (Gutha et al., 2010; Wallis and Chen, 2012). In addition, they also contribute to organoleptic features of grape berry and wine (Schmidtke et al., 2010; Gutiérrez-Capitán et al., 2014).

Lignin

Simple phenolic compounds that synthesized from phenylalanine can be polymerized to lignin which is an essential component of the cell wall. In the present study, three enzymes, i.e. cinnamoyl-CoA reductase (CCR), cinnamyl-alcohol dehydrogenase (CAD) and ferulate 5-hydroxylase (F5H, Q9M4H8) involved in lignin synthesis were annotated. CCR and CAD, which catalyze the last two steps of monolignol synthesis not only impact lignification but also plant development. Absence of CCR and CAD resulted in dwarfism and sterility in *Arabidopsis* (Thevenin et al., 2011). One protein candidate was annotated as CCR (A5AXM6) and detected after veraison. Four candidates were annotated as CAD. The averaged expression pattern of CAD showed high levels in both young and ripening berries and low level around veraison (**Figure 5**). Aharoni and coworkers (Aharoni et al., 2002) reported a comparable pattern of the transcription level of CAD in developing strawberries. In their study, CAD expression level was high in green strawberries followed by a decreasing during the white and turning stages and finally increased again in red strawberries (Aharoni et al., 2002). Like grape, strawberry is also a non-climacteric fruit. Furthermore, strawberry also undergoes color turning phases. The comparable developmental process might be supported by similar molecular activities.

Flavonoids

Flavonoid is another class of secondary metabolites derived from phenylpropanoid pathway and shares common precursors with lignin. The changing patterns of the detected flavonols, flavan-3-ols, and anthocyanins were described in metabolomic profile part thus we will not

repeat these informations but rather focusing on the enzymes involved in flavonoid synthesis. The annotated proteins that involved in the flavonoid pathway include chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), dihydroflavanol 4-reductase (DFR) and anthocyanidin synthase (ANS). CHS catalyzes the condensation of *p*-coumaroyl-CoA with malonyl-CoA to generate chalcone which further isomerized to naringenin by CHI. Two proteins, annotated as CHS (A2ICC5, G4XGW2), were detected after veraison and increased in abundance during ripening (**Table S4**). In a previous study, three copies of *Chs* were found in the grape genome. The mRNA levels of *Chs2* and *Chs3* significantly coincided with anthocyanin and *Chs1* and *Chs2* with flavonol biosynthesis (Jeong et al., 2008). The expression of A2ICC5 and G4XGW2 were in accordance with anthocyanin accumulation indicating the involvement of these two CHSs in the coordination of anthocyanin synthesis. The expression level of CHI (A5ANT9) gradually increased before veraison then underwent a sharp decline at veraison (EL 35) with a subsequent recovery to the level before veraison. F3H catalyzes the hydroxylation of flavanones at 3-position to form dihydroflavonols. Both protein candidates annotated as F3H (A2ICC8, A2ICC8) were detected from EL 32 and increasingly expressed during grape berry development. The transcription of F3hs appeared to be induced for the biosynthesis of flavonols and anthocyanins (Jeong et al., 2008). Dihydroflavonols are further converted to either flavonols via flavonol synthase (FLS) or to leucoanthocyanidins via DFR. The competition between FLS and DFR influences the contents of flavonols and anthocyanins (Tian et al., 2015) which further affects the color (Lou et al., 2014) and the abilities of plants to cope with stress (Hua et al., 2013; Wang et al., 2013). DFR is a crucial enzyme in the flavonoid pathway involved in the synthesis of anthocyanins, proanthocyanins and tannins (Moyano et al., 1998; Zhang et al., 2008; Hua et al., 2013; Wang et al., 2016). In the present study, two protein candidates (A5BGJ0, A5BIY8) were annotated as DFR. Their combined expression pattern indicated a progressive increased in abundance of DFR before veraison and a slight decrease afterwards (**Figure 5**). ANS catalyzing the conversion of colorless leucoanthocyanins to colored anthocyanidins was detected in ripening berries. The abundance of ANS significantly increased after veraison (**Figure 5**) in accordance with the anthocyanin accumulation.

4. Conclusion

In summary, the analysis of changes in metabolism of grape berries from fruit set to mature fruit by mass spectrometry based platforms revealed dynamic process at proteome and

metabolome levels. The broad coverage of developmental stages included in the present study ensured the intactness of our observation about these dynamic processes. In addition, we reported an extensive dynamic metabolism network covering central carbon metabolism such as sugar metabolism, glycolysis, TCA cycle, amino acid metabolism together with secondary metabolism, especially the flavonoid pathway. Stilbenes are a metabolite family including bioactive metabolites that were reported to be rich in grape. However, we did not detect any compound candidate or related enzyme in this study, which might be due to the genetic background of the variety we studied. It is thus of interest to compare the flavonoid content of this variety with other grape varieties that produce stilbenes to investigate the competition of stilbene and flavonoid biosynthesis. Red varieties of grape are rich in anthocyanins and other flavonoids, which makes them a great model plant for studies of anthocyanin or flavonoid synthesis pathways. The inequality of anthocyanins that derived from different branches was observed at the metabolome level. Further works is needed to explain the regulation of anthocyanin synthesis by the transcriptional and translational evidences. The multivariate statistical analysis such as PCA, clustering analysis and Granger causality analysis assisted the interpretation of the integrated metabolome and proteome dataset and revealed the interactions between metabolites, proteins and various metabolisms pathways.

Conflict of interest statement

The authors have no conflict of interest to declare.

Author contributions

LW and WW conceived and designed the experiments. LW performed the experiments. LW and JW analyzed the data. WW provided the reagents, materials and analytical tools. LW wrote the manuscript. WW and JW revised the manuscript. All the authors approved the final manuscript.

Funding

Lei Wang was supported by a PhD scholarship provided by China Scholarship Council (CSC) (Grant number: 201206220134).

Acknowledgement

We would like to thank the gardeners for their great maintenance of the grapevine. Thank all the MoSys members for their fruitful discussion. Thank Mag. Reinhard Turetschek for his suggestions in protein blast analysis.

REFERENCES

- Abu-Reidah, I.M., Ali-Shtayeh, M.S., Jamous, R.M., Arraez-Roman, D., and Segura-Carretero, A. (2015). HPLC-DAD-ESI-MS/MS screening of bioactive components from *Rhus coriaria* L. (Sumac) fruits. *Food Chem* 166, 179-191.
- Adisakwattana, S., Moonrat, J., Srichairat, S., Chanasit, C., Tirapongporn, H., Chanathong¹, B., Ngamukote¹, S., Mäkynen, K., and Sapwarobol, S. (2010). Lipid-Lowering mechanisms of grape seed extract (*Vitis vinifera* L) and its antihyperlipidemic activity. *Journal of Medicinal Plants Research* 4, 2113-2120.
- Agudelo-Romero, P., Erban, A., Sousa, L., Pais, M.S., Kopka, J., and Fortes, A.M. (2013). Search for transcriptional and metabolic markers of grape pre-ripening and ripening and insights into specific aroma development in three Portuguese cultivars. *PLoS One* 8, e60422.
- Aharoni, A., Keizer, L.C., Van Den Broeck, H.C., Blanco-Portales, R., Munoz-Blanco, J., Bois, G., Smit, P., De Vos, R.C., and O'connell, A.P. (2002). Novel insight into vascular, stress, and auxin-dependent and -independent gene expression programs in strawberry, a non-climacteric fruit. *Plant Physiol* 129, 1019-1031.
- Ali, K., Maltese, F., Fortes, A.M., Pais, M.S., Choi, Y.H., and Verpoorte, R. (2011a). Monitoring biochemical changes during grape berry development in Portuguese cultivars by NMR spectroscopy. *Food Chemistry* 124, 1760-1769.
- Ali, M.B., Howard, S., Chen, S., Wang, Y., Yu, O., Kovacs, L.G., and Qiu, W. (2011b). Berry skin development in Norton grape: distinct patterns of transcriptional regulation and flavonoid biosynthesis. *BMC Plant Biol* 11, 7.
- Anastasiadi, M., Pratsinis, H., Kletsas, D., Skaltsounis, A.-L., and Haroutounian, S.A. (2010). Bioactive non-coloured polyphenols content of grapes, wines and vinification by-products: Evaluation of the antioxidant activities of their extracts. *Food Research International* 43, 805-813.
- Ayenew, B., Degu, A., Manela, N., Perl, A., Shamir, M.O., and Fait, A. (2015). Metabolite profiling and transcript analysis reveal specificities in the response of a berry derived cell culture to abiotic stresses. *Front Plant Sci* 6, 728.
- Boss, P.K., Davies, C., and Robinson, S.P. (1996). Anthocyanin composition and anthocyanin pathway gene expression in grapevine sports differing in berry skin colour. *Aust. J. Grape Wine R.* 2, 163-170.
- Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry* 72, 248-254.
- Castillo-Munoz, N., Fernandez-Gonzalez, M., Gomez-Alonso, S., Garcia-Romero, E., and Hermosin-Gutierrez, I. (2009). Red-color related phenolic composition of Garnacha Tintorera (*Vitis vinifera* L.) grapes and red wines. *J Agric Food Chem* 57, 7883-7891.
- Chen, S.D., Lu, C.J., and Zhao, R.Z. (2015). Identification and quantitative characterization of PSORI-CM01, a Chinese medicine formula for psoriasis therapy, by liquid chromatography coupled with an LTQ Orbitrap mass spectrometer. *Molecules* 20, 1594-1609.
- Clemente, E., and Galli, D. (2011). Stability of the anthocyanins extracted from residues of the wine industry. *Ciênc. Tecnol. Aliment.* 31, 765-768.
- Conde, C., Silva, P., Fontes, N., Dias, A.C.P., Tavares, R.M., Sousa, M.J., Agasse, A., Delrot, S., and Gerós, H. (2007). Biochemical Changes throughout Grape Berry Development and Fruit and Wine Quality. *Food* 1, 1-22.

- Coombe, B.G. (1995). Adoption of a system for identifying grapevine growth stages. *Australian Journal of Grape and Wine Research* 1, 100-110.
- Cuadros-Inostroza, A., Ruiz-Lara, S., Gonzalez, E., Eckardt, A., Willmitzer, L., and Pena-Cortes, H. (2016). GC-MS metabolic profiling of Cabernet Sauvignon and Merlot cultivars during grapevine berry development and network analysis reveals a stage- and cultivar-dependent connectivity of primary metabolites. *Metabolomics* 12, 39.
- Dai, Z.W., Leon, C., Feil, R., Lunn, J.E., Delrot, S., and Gomes, E. (2013). Metabolic profiling reveals coordinated switches in primary carbohydrate metabolism in grape berry (*Vitis vinifera* L.), a non-climacteric fleshy fruit. *J Exp Bot* 64, 1345-1355.
- Davies, C., and Robinson, S.P. (2000). Differential screening indicates a dramatic change in mRNA profiles during grape berry ripening. Cloning and characterization of cDNAs encoding putative cell wall and stress response proteins. *Plant physiology* 122, 803-812.
- Degu, A., Hochberg, U., Sikron, N., Venturini, L., Buson, G., Ghan, G., Plaschkes, I., Batushansky, A., Chalifa-Caspi, V., Mattivi, F., Delledonne, M., Pezzotti, M., Rachmilevitch, S., Cramer, G.R., and Fait, A. (2014). Metabolite and transcript profiling of berry skin during fruit development elucidates differential regulation between Cabernet Sauvignon and Shiraz cultivars at branching points in the polyphenol pathway. *BMC Plant Biology* 14, 188.
- Degu, A., Morcia, C., Tumino, G., Hochberg, U., Toubiana, D., Mattivi, F., Schneider, A., Bosca, P., Cattivelli, L., Terzi, V., and Fait, A. (2015). Metabolite profiling elucidates communalities and differences in the polyphenol biosynthetic pathways of red and white Muscat genotypes. *Plant Physiol Biochem* 86, 24-33.
- Delcambre, A., and Saucier, C. (2012). Identification of new flavan-3-ol monoglycosides by UHPLC-ESI-Q-TOF in grapes and wine. *J Mass Spectrom* 47, 727-736.
- Deluc, L.G., Grimplet, J., Wheatley, M.D., Tillett, R.L., Quilici, D.R., Osborne, C., Schooley, D.A., Schlauch, K.A., Cushman, J.C., and Cramer, G.R. (2007). Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development. *BMC Genomics* 8, 429.
- Doerfler, H., Lyon, D., Nagele, T., Sun, X., Fagner, L., Hadacek, F., Egelhofer, V., and Weckwerth, W. (2013). Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics* 9, 564-574.
- Doshi, P., Adsule, P., and Banerjee, K. (2006). Phenolic composition and antioxidant activity in grapevine parts and berries (*Vitis vinifera* L.) cv. Kishmish Chornyi (Sharad Seedless) during maturation. *International Journal of Food Science and Technology* 41, 1-9.
- Duncan, D.B. (1955). Multiple range and multiple F tests. *Biometrices* 11, 1-42.
- Eichhorn, K.W., and Lorenz, D. (1978). *Phaenologische Entwicklungsstadien der Rebe: Anwendungstermine d. Pflanzenschutzmittel ; Sonderdr. aus "Der deutsche Weinbau".* BASF.
- Fang, F., Tang, K., and Huang, W.-D. (2013). Changes of flavonol synthase and flavonol contents during grape berry development. *European Food Research and Technology* 237, 529-540.
- Feng, Y., Liu, Y.M., Leblanc, M.H., Bhatt, A.J., and Rhodes, P.G. (2007). Grape seed extract given three hours after injury suppresses lipid peroxidation and reduces hypoxic-ischemic brain injury in neonatal rats. *Pediatr Res* 61, 295-300.
- Flamini, R., De Rosso, M., and Bavaresco, L. (2015). Study of Grape Polyphenols by Liquid Chromatography-High-Resolution Mass Spectrometry (UHPLC/QTOF) and Suspect Screening Analysis. *J Anal Methods Chem* 2015, 350259.
- Fortes, A.M., Agudelo-Romero, P., Silva, M.S., Ali, K., Sousa, L., Maltese, F., Choi, Y.H., Grimplet, J., Martinez-Zapater, J.M., Verpoorte, R., and Pais, M.S. (2011). Transcript and metabolite analysis in Trincadeira cultivar reveals novel information regarding the dynamics of grape ripening. *BMC Plant Biol* 11, 149.
- Fraige, K., Gonzalez-Fernandez, R., Carrilho, E., and Jorrin-Novo, J.V. (2015). Metabolite and proteome changes during the ripening of Syrah and Cabernet Sauvignon grape varieties cultured in a nontraditional wine region in Brazil. *J Proteomics* 113, 206-225.
- Giribaldi, M., Perugini, I., Sauvage, F.X., and Schubert, A. (2007). Analysis of protein changes during grape berry ripening by 2-DE and MALDI-TOF. *Proteomics* 7, 3154-3170.

- Granger, C. (1969). Investigation causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424-438.
- Gutha, L.R., Casassa, L.F., Harbertson, J.F., and Naidu, R.A. (2010). Modulation of flavonoid biosynthetic pathway genes and anthocyanins due to virus infection in grapevine (*Vitis vinifera* L.) leaves. *BMC Plant Biol* 10, 187.
- Gutiérrez-Capitán, M., Capdevila, F., Vila-Planas, J., Domingo, C., Büttgenbach, S., Llobera, A., Puig-Pujol, A., and Jiménez-Jorquera, C. (2014). Hybrid Electronic Tongues Applied to the Quality Control of Wines. *Journal of Sensors* 2014, 1-10.
- He, F., Liang, N.N., Mu, L., Pan, Q.H., Wang, J., Reeves, M.J., and Duan, C.Q. (2012). Anthocyanins and their variation in red wines I. Monomeric anthocyanins and their color expression. *Molecules* 17, 1571-1601.
- He, F., Mu, L., Yan, G.L., Liang, N.N., Pan, Q.H., Wang, J., Reeves, M.J., and Duan, C.Q. (2010). Biosynthesis of anthocyanins and their regulation in colored grapes. *Molecules* 15, 9057-9091.
- Hoehenwarter, W., Van Dongen, J.T., Wienkoop, S., Steinfath, M., Hummel, J., Erban, A., Sulpice, R., Regierer, B., Kopka, J., Geigenberger, P., and Weckwerth, W. (2008). A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* 8, 4214-4225.
- Hua, C., Linling, L., Shuiyuan, C., Fuliang, C., Feng, X., Honghui, Y., and Conghua, W. (2013). Molecular cloning and characterization of three genes encoding dihydroflavonol-4-reductase from *Ginkgo biloba* in anthocyanin biosynthetic pathway. *PLoS One* 8, e72017.
- Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F., Wincker, P., and French-Italian Public Consortium for Grapevine Genome, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463-467.
- Jeong, S.T., Goto-Yamamoto, N., Hashizume, K., and Esaka, M. (2008). Expression of multi-copy flavonoid pathway genes coincides with anthocyanin, flavonol and flavan-3-ol accumulation of grapevine. *Vitis* 47, 135-140.
- Kambiranda, D., Katam, R., Basha, S.M., and Siebert, S. (2014). iTRAQ-based quantitative proteomics of developing and ripening muscadine grape berry. *J Proteome Res* 13, 555-569.
- Kennedy, J. (2002). Understanding grape berry development. *Practical Winery and Vineyard Journal* 7, 1-5.
- Kim, Y., Choi, Y., Lee, J., and Park, Y. (2013). Downregulated lipid metabolism in differentiated murine adipocytes by procyanidins from defatted grape seed meal. *Biosci Biotechnol Biochem* 77, 1420-1423.
- Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R., and Steinhauser, D. (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21, 1635-1638.
- Król, A., Amarowicz, R., and Weidner, S. (2014). Changes in the composition of phenolic compounds and antioxidant properties of grapevine roots and leaves (*Vitis vinifera* L.) under continuous of long-term drought stress. *Acta Physiologiae Plantarum* 36, 1491-1499.
- Liang, Z., Sang, M., Fan, P., Wu, B., Wang, L., Duan, W., and Li, S. (2011). Changes of polyphenols, sugars, and organic acid in 5 *Vitis* genotypes during berry ripening. *J Food Sci* 76, C1231-1238.
- Liang, Z., Wu, B., Fan, P., Yang, C., Duan, W., Zheng, X., Liu, C., and Li, S. (2008). Anthocyanin composition and content in grape berry skin in *Vitis* germplasm. *Food Chemistry* 111, 837-844.

- Lou, Q., Liu, Y., Qi, Y., Jiao, S., Tian, F., Jiang, L., and Wang, Y. (2014). Transcriptome sequencing and metabolite analysis reveals the role of delphinidin metabolism in flower colour in grape hyacinth. *J Exp Bot* 65, 3157-3164.
- Mari, A., Lyon, D., Fragner, L., Montoro, P., Piacente, S., Wienkoop, S., Egelhofer, V., and Weckwerth, W. (2013). Phytochemical composition of L. analyzed by an integrative GC-MS and LC-MS metabolomics platform. *Metabolomics* 9, 599-607.
- Martinez-Esteso, M.J., Selles-Marchart, S., Lijavetzky, D., Pedreno, M.A., and Bru-Martinez, R. (2011). A DIGE-based quantitative proteomic analysis of grape berry flesh development and ripening reveals key events in sugar and organic acid metabolism. *J Exp Bot* 62, 2521-2569.
- Mazza, G., Fukumoto, L., Delaquis, P., Girard, B., and Ewert, B. (1999). Anthocyanins, Phenolics, and Color of Cabernet Franc, Merlot, and Pinot Noir Wines from British Columbia. *J. Agric. Food Chem.* 47, 4009-4017.
- Monagas, M., Suárez, R., Gómez-Cordovés, C., and Bartolomé, B. (2005). Simultaneous determination of nonanthocyanin phenolic compounds in red wines by HPLC-DAD/ESI-MS. *Am. J. Enol. Vitic.* 56, 139-147.
- Moyano, E., Portero-Robles, I., Medina-Escobar, N., Valpuesta, V., Muñoz-Blanco, J., and Caballero, J.L. (1998). A Fruit-Specific Putative Dihydroflavonol 4-Reductase Gene Is Differentially Expressed in Strawberry during the Ripening Process. *Plant Physiol.* 117, 711-716.
- Muñoz-Robredo, P., Robledo, P., Manríquez, D., Molina, R., and Defilippi, B.G. (2011). Characterization of sugars and organic acids in commercial varieties of table grapes. *Chilean J. Agric. Res.* 71, 452-458.
- Negri, A.S., Prinsi, B., Rossoni, M., Failla, O., Scienza, A., Cocucci, M., and Espen, L. (2008). Proteome changes in the skin of the grape cultivar Barbera among different stages of ripening. *BMC Genomics* 9, 378.
- Noah, A.M., Niemenak, N., Sunderhaus, S., Haase, C., Omokolo, D.N., Winkelmann, T., and Braun, H.P. (2013). Comparative proteomic analysis of early somatic and zygotic embryogenesis in Theobroma cacao L. *J Proteomics* 78, 123-133.
- Palumbo, M.C., Zenoni, S., Fasoli, M., Massonnet, M., Farina, L., Castiglione, F., Pezzotti, M., and Paci, P. (2014). Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative switch genes that induce major transcriptome reprogramming during grapevine development. *Plant Cell* 26, 4617-4635.
- Pandey, P.M., and Farmahan, H.L. (1977). Changes in the rate of photosynthesis and respiration in leaves and berries of Vitis vinifera grapevines at various stages of berry development. *Vitis* 16, 106-111.
- Papini, P.C., Mazza, G., Gatti, M., and Bavaresco, L. (2010). Anthocyanin and aroma profiling of the 'Albarossa' grapevine crossbreed (Vitis vinifera L.) and its parent varieties 'Barbera' and 'Nebbiolo di Dronero'. *Vitis* 49, 121-127.
- Pontin, M.A., Piccoli, P.N., Francisco, R., Bottini, R., Martinez-Zapater, J.M., and Lijavetzky, D. (2010). Transcriptome changes in grapevine (Vitis vinifera L.) cv. Malbec leaves induced by ultraviolet-B radiation. *BMC Plant Biol* 10, 224.
- Razavi, S.M., Gholamin, S., Eskandari, A., Mohsenian, N., Ghorbanihaghjo, A., Delazar, A., Rashtchizadeh, N., Keshtkar-Jahromi, M., and Argani, H. (2013). Red grape seed extract improves lipid profiles and decreases oxidized low-density lipoprotein in patients with mild hyperlipidemia. *J Med Food* 16, 255-258.
- Robinson, J., Harding, J., and Vouillamoz, J. (2012). Wine grapes: A complete guide to 1,136 vine varieties, including their origins and flavours. 1280.
- Ruffner, H.P., Koblet, W., and Rast, D. (1975). Gluconeogenesis in ripening fruit of Vitis vinifera. *Vitis* 13, 319-328.
- Sánchez-Alonso, I., Jiménez-Escrig, A., Saura-Calixto, F., and Borderías, A.J. (2007). Effect of grape antioxidant dietary fibre on the prevention of lipid oxidation in minced fish: Evaluation by different methodologies. *Food Chemistry* 101, 372-378.

- Sánchez-Moreno, C., Larrauri, J.A., and Saura-Calixto, F. (1999). Free radical scavenging capacity and inhibition of lipid oxidation of wines, grape juices and related polyphenolic constituents. *Food Research International* 32, 407-412.
- Sáyago-Ayerdi, S.G., Brenes, A., and Goñi, I. (2009). Effect of grape antioxidant dietary fiber on the lipid oxidation of raw and cooked chicken hamburgers. *LWT - Food Science and Technology* 42, 971-976.
- Schmidtke, L.M., Rudnitskaya, A., Saliba, A.J., Blackman, J.W., Scollary, G.R., Clark, A.C., Rutledge, D.N., Delgadillo, I., and Legin, A. (2010). Sensory, chemical, and electronic tongue assessment of micro-oxygenated wines and oak chip maceration: assessing the commonality of analytical techniques. *J Agric Food Chem* 58, 5026-5033.
- Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.N., Lindon, J.C., Marriott, P., Nicholls, A.W., Reilly, M.D., Thaden, J.J., and Viant, M.R. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3, 211-221.
- Sun, X., and Weckwerth, W. (2012). COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* 8, 81-93.
- Tebib, K., Besançon, P., and Rouanet, J. (1994). Dietary grape seed tannins affect lipoproteins, lipoprotein lipases and tissue lipids in rats fed hypercholesterolemic diets. *J Nutr.* 124, 2451-2457.
- Thevenin, J., Pollet, B., Letarnec, B., Saulnier, L., Gissot, L., Maia-Grondard, A., Lapierre, C., and Jouanin, L. (2011). The simultaneous repression of CCR and CAD, two enzymes of the lignin biosynthetic pathway, results in sterility and dwarfism in *Arabidopsis thaliana*. *Mol Plant* 4, 70-82.
- Tian, J., Han, Z.Y., Zhang, J., Hu, Y., Song, T., and Yao, Y. (2015). The Balance of Expression of Dihydroflavonol 4-reductase and Flavonol Synthase Regulates Flavonoid Biosynthesis and Red Foliage Coloration in Crabapples. *Sci Rep* 5, 12228.
- Valledor, L., Furuhashi, T., Recuenco-Munoz, L., Wienkoop, S., and Weckwerth, W. (2014). System-level network analysis of nitrogen starvation and recovery in *Chlamydomonas reinhardtii* reveals potential new targets for increased lipid accumulation. *Biotechnol Biofuels* 7, 171.
- Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J.T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J.A., Sterck, L., Vandepoele, K., Grando, S.M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S.K., Fontana, P., Gutin, A., Van De Peer, Y., Salamini, F., and Viola, R. (2007). A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. *PLoS ONE* 2, e1326.
- Vizcaino, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R.J., Kraus, H.-J., Albar, J.P., Martinez-Bartolome, S., Apweiler, R., Omenn, G.S., Martens, L., Jones, A.R., and Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotech* 32, 223-226.
- Wallis, C.M., and Chen, J. (2012). Grapevine phenolic compounds in xylem sap and tissues are significantly altered during infection by *Xylella fastidiosa*. *Phytopathology* 102, 816-826.
- Wang, H., Fan, W., Li, H., Yang, J., Huang, J., and Zhang, P. (2013). Functional characterization of Dihydroflavonol-4-reductase in anthocyanin biosynthesis of purple sweet potato underlies the direct evidence of anthocyanins function against abiotic stresses. *PLoS One* 8, e78484.
- Wang, L., Zhu, Y., Wang, P., Fan, Q., Wu, Y., Peng, Q.Z., Xia, G.X., and Wu, J.H. (2016). Functional Characterization of a Dihydroflavanol 4-Reductase from the Fiber of Upland Cotton (*Gossypium hirsutum*). *Molecules* 21.

- Weckwerth, W., Wenzel, K., and Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4, 78-83.
- Wu, B.H., Cao, Y.G., Guan, L., Xin, H.P., Li, J.H., and Li, S.H. (2014). Genome-wide transcriptional profiles of the berry skin of two red grape cultivars (*Vitis vinifera*) in which anthocyanin synthesis is sunlight-dependent or -independent. *PLoS One* 9, e105959.
- Wu, B.H., Liu, H.F., Guan, L., Fan, P.G., and Li, S.H. (2011). Carbohydrate metabolism in grape cultivars that differ in sucrose accumulation. *Vitis* 50, 51-57.
- Zamboni, A., Di Carli, M., Guzzo, F., Stocchero, M., Zenoni, S., Ferrarini, A., Tononi, P., Toffali, K., Desiderio, A., Lilley, K.S., Pe, M.E., Benvenuto, E., Delledonne, M., and Pezzotti, M. (2010). Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. *Plant Physiol* 154, 1439-1459.
- Zhang, P., Wen, P.F., Wan, S.B., Wang, W., Pan, Q.H., Zhang, J.C., and Huang, W.D. (2008). Molecular cloning of dihydroflavonol 4-reductase gene from grape berry and preparation of an anti-DFR polyclonal antibody. *Vitis* 47, 141-145.
- Zhang, Z., Li, B.Y., Li, X.L., Cheng, M., Yu, F., Lu, W.D., Cai, Q., Wang, J.F., Zhou, R.H., Gao, H.Q., and Shen, L. (2013). Proteomic analysis of kidney and protective effects of grape seed procyanidin B2 in db/db mice indicate MFG-E8 as a key molecule in the development of diabetic nephropathy. *Biochim Biophys Acta* 1832, 805-816.
- Zhao, Y., Li, X., Zeng, X., Huang, S., Hou, S., and Lai, X. (2014). Characterization of phenolic constituents in *Lithocarpus polystachyus*. *Analytical Methods* 6, 1359.
- Zybailov, B., Mosley, A.L., Sardiu, M.E., Coleman, M.K., Florens, L., and Washburn, M.P. (2006). Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *Journal of Proteome Research* 5, 2339-2347.

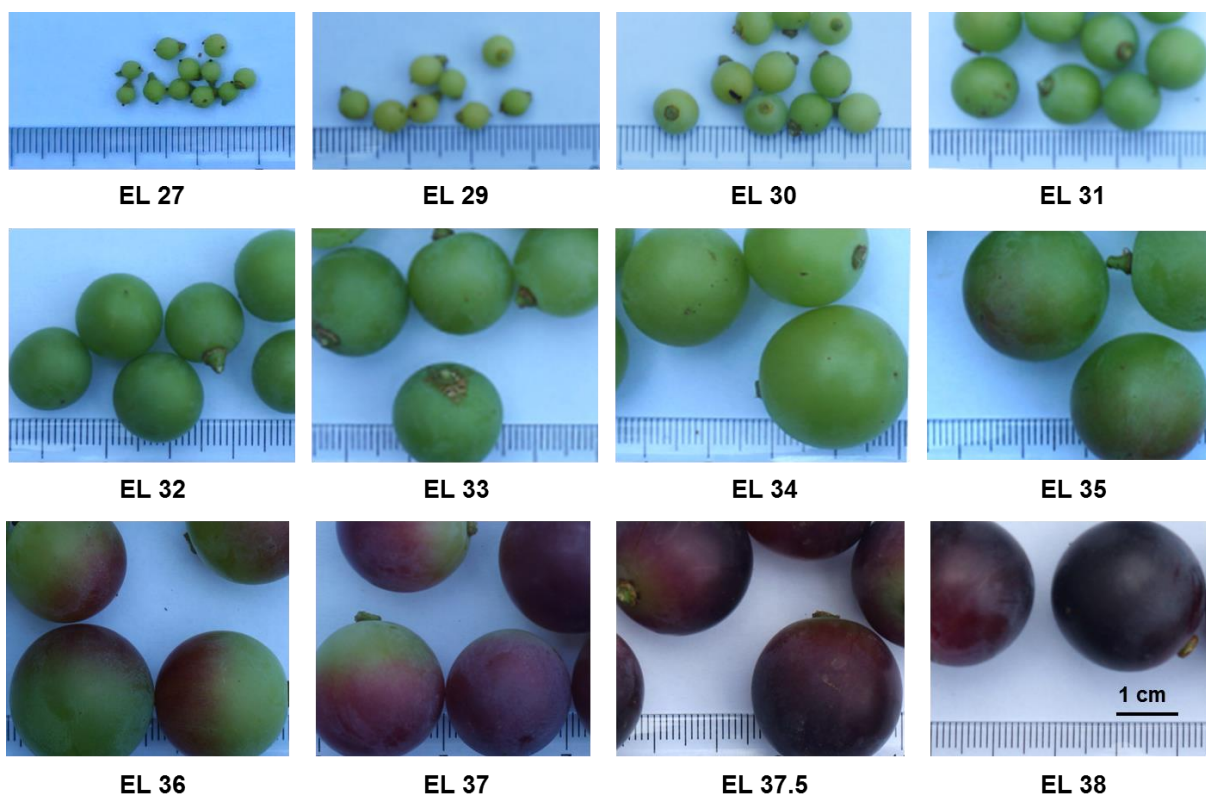
876

877 **Figures and Figure Legends**

878

879

880



881

882 **Figure 1. Grape berries harvested at 12 developmental stages according to modified E-L**
883 **system.**

884

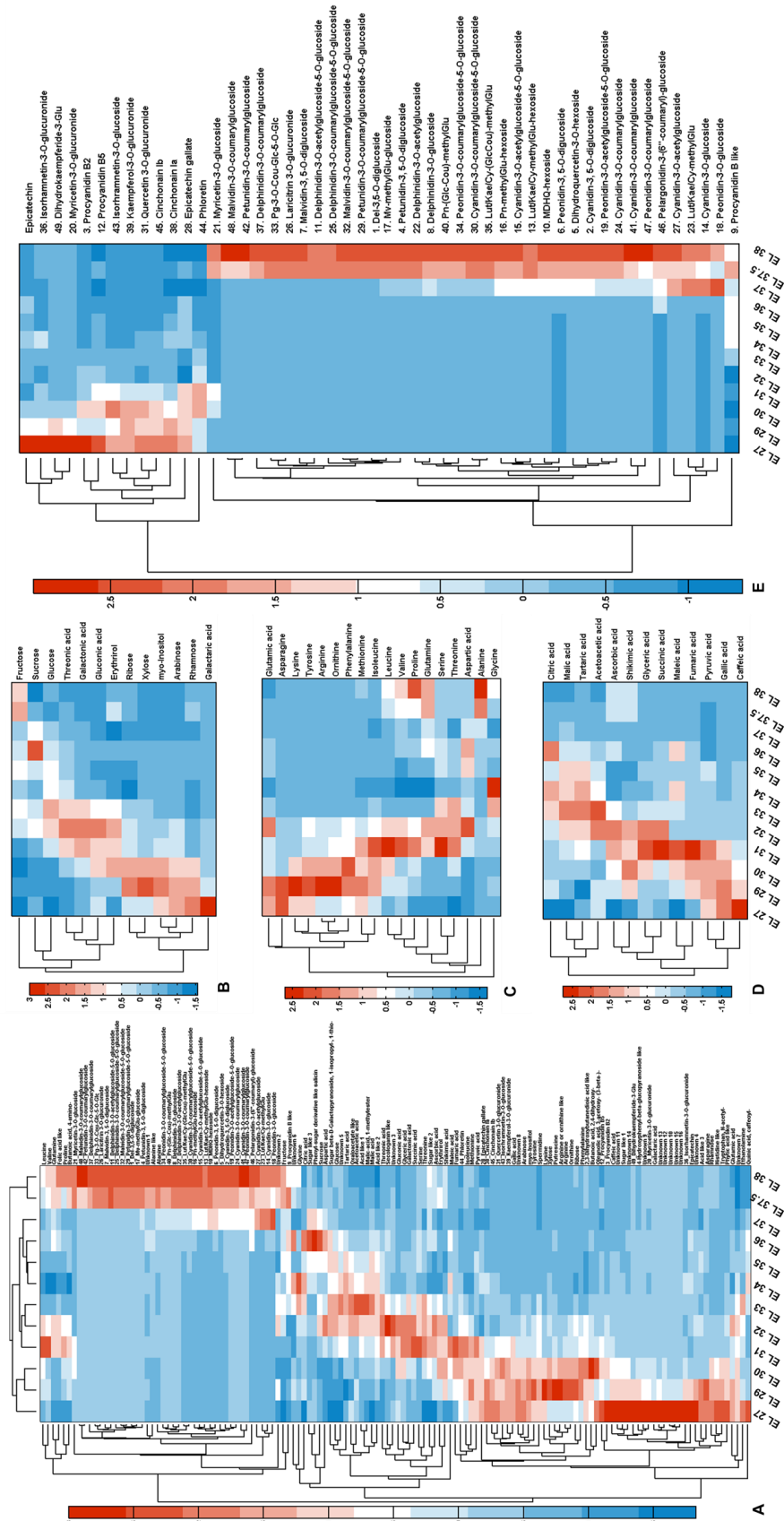


Figure 2. Metabolome dynamics of developing grape berry. (A) Overview of the metabolite dynamics with bi-hierarchical-clustering heat map. (B, C, and D) present the dynamics of sugars (including sugar alcohol and sugar acids), amino acids and organic acids.

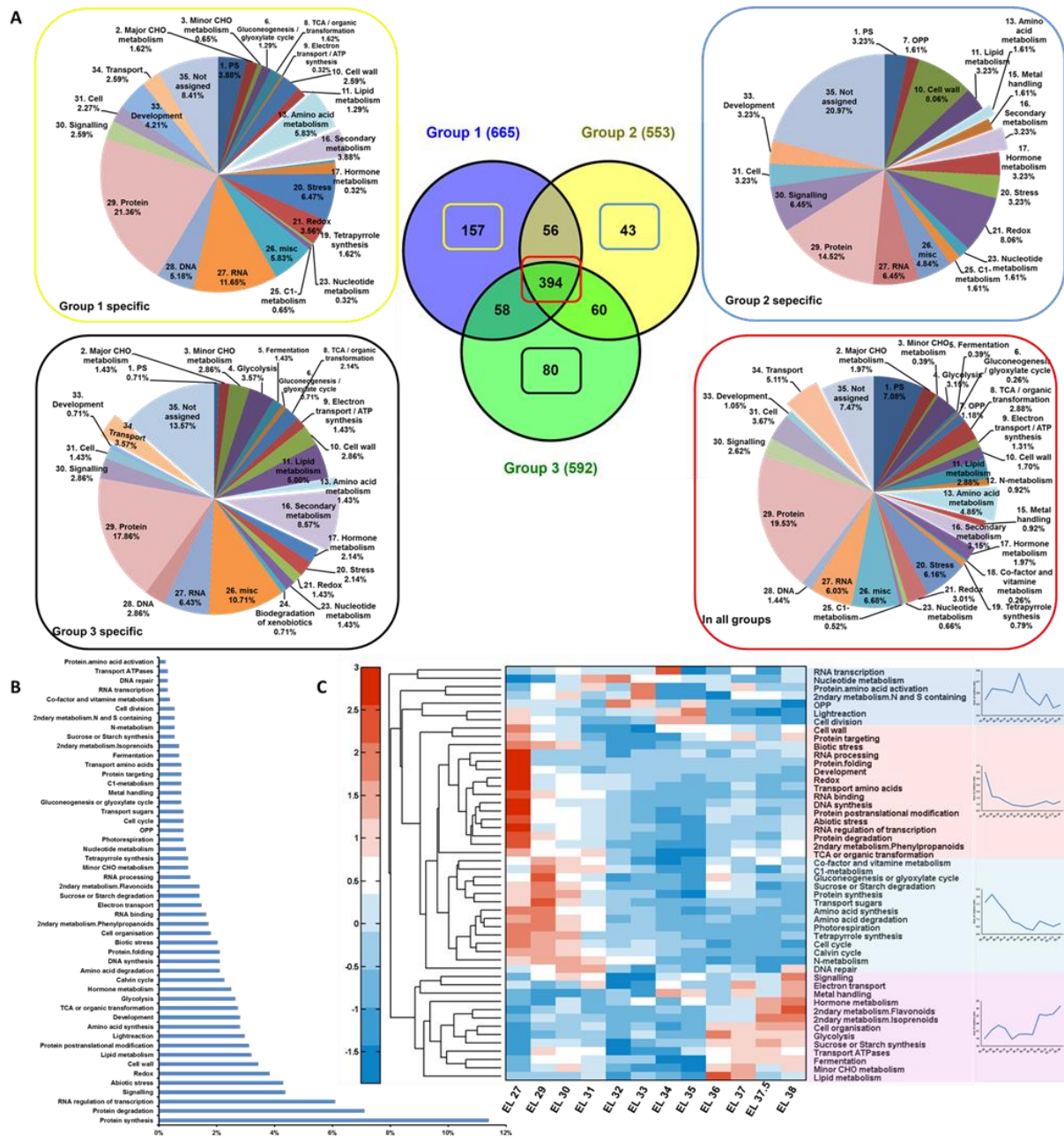


Figure 3. Proteomic analysis of developing grape berries. (A) Protein distribution throughout the developmental process. Samples of 12 developing stages were sorted to 3 groups with group 1 including samples at stage EL 27, 29, 30, 31; group 2, EL32, 33, 34, 35 and group 3, EL 36, 37, 37.5, 38. The proteins that specific to group 1, 2, 3 and those common to all groups were functionally summarized with pie charts in yellow, blue, black and red box respectively. (B) Protein frequencies. (C) Protein dynamics. Four distinctive changing patterns were summarized with line charts.

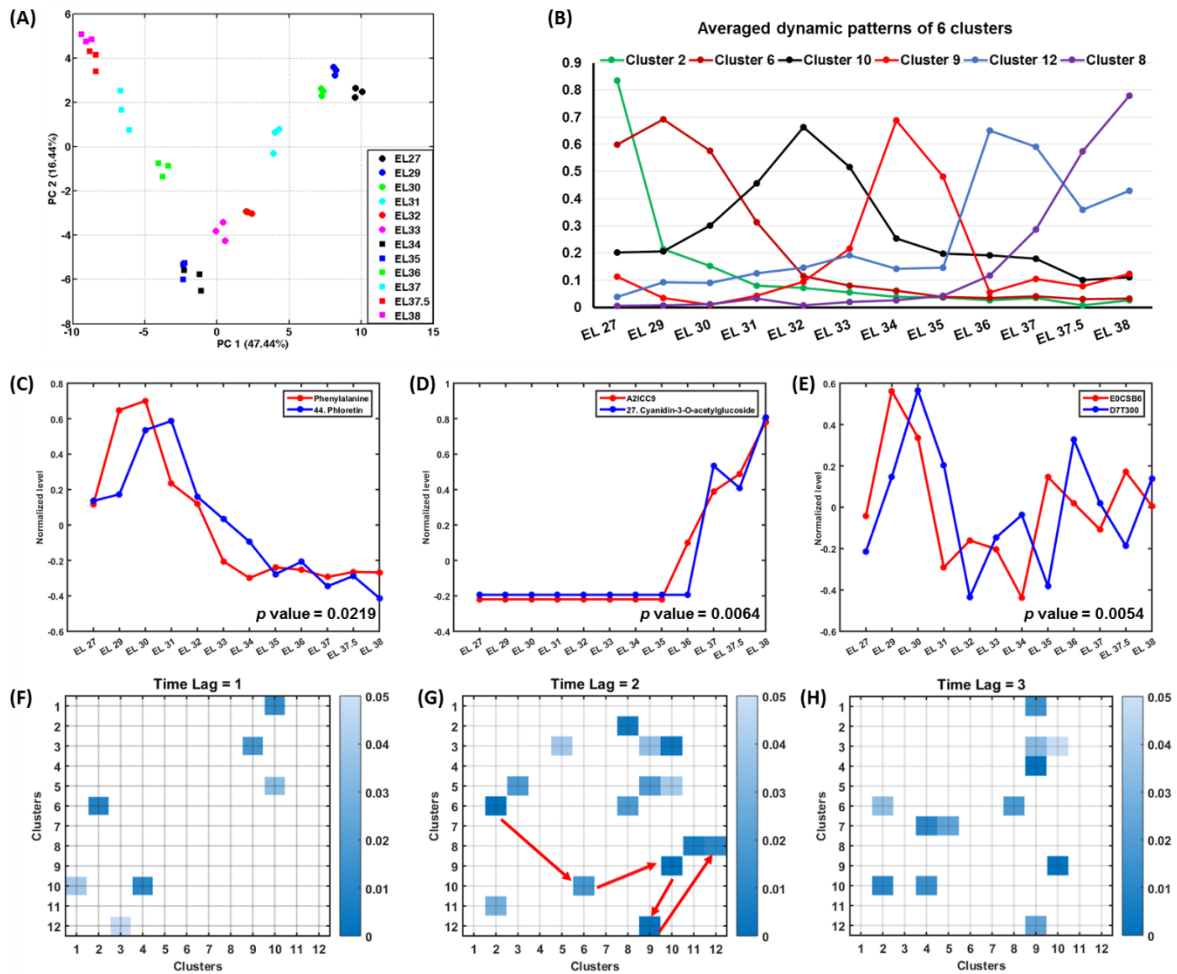


Figure 4. Multivariate statistical analysis of the integrated metabolome and proteome dataset.

A. PCA plot of the integrated dataset shows a trajectory during grape berry development.

B. The averaged dynamic patterns of 6 clusters from k-means clustering analysis.

C, D and E. Examples of Granger causalities showed directed interactions of metabolite-metabolite, metabolite-protein and protein-protein.

F, G and H present Granger causalities between clusters that resulted from k-means clustering analysis with time lag set as 1, 2 and 3, respectively.

911

912 **Figure 5. Visualization of metabolite and protein dynamics on their biosynthetic pathways.**

913 Metabolites are written in black letters with blue line charts indicating their changing patterns
914 whereas proteins are written in red letters with red line charts. Relative abundance of
915 metabolites and proteins were averaged over three biological replicates. Bars represent standard
916 errors. Abbreviation: Susy, sucrose synthase; UGPase, UDP-glucose pyrophosphorylase; PFP,
917 pyrophosphate-fructose 6-phosphate 1-phosphotransferase; FBPase, fructose 1, 6-
918 biphosphatase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; PK, pyruvate
919 kinase; PDC, pyruvate dehydrogenase complex; IDH, isocitrate dehydrogenase; OGDC,
920 oxoglutarate dehydrogenase complex; SCS, succinyl coenzyme A synthetase; SDH, succinate
921 dehydrogenase; MDH, malate dehydrogenase; AspAT, aspartate aminotransferase; AS,
922 asparagine synthetase; ASADH, aspartate-semialdehyde dehydrogenase; MetH, methionine
923 synthase; MAT, methionine adenosyltransferase; PHGDH, phosphoglycerate dehydrogenase;
924 PSAT, Phosphoserine transaminase; SHMT, serine hydroxymethyltransferase; OASTL, O-
925 acetylserine (thiol)-lyase; GLDH, Glutamate dehydrogenase; GS, Glutamine synthetase; ASS,
926 argininosuccinate synthase; ALAT, alanine aminotransferase; AGT, alanine-glyoxylate
927 transaminase; CS, chorismate synthase; PAL, phenylalanine ammonia lyase; C4H, cinnamate
928 4-hydroxylase; 4-coumarate-CoA ligase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl-
929 alcohol dehydrogenase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-
930 hydroxylase; F3'H, flavonoid 3'-hydroxylase; F3'5'H, flavanoid 3',5'-hydroxylase; FLS,
931 flavonol synthase; DFR, dihydroflavanol 4-reductase; ANS, anthocyanidin synthase; ANR,
932 anthocyanidin reductase; 3-GT, anthocyanidin 3-O-glucosyltransferase.

933

934

935

936

938 **Table**

939 **Table 1 Flavonoids annotated from LC-MS data grape berry.**

Candidate number	Deduced structure	RT (min)	[M + H] ⁺ (m/z)	Mass accuracy (ppm)	Export formula	MS/MS (m/z)	Reference	Annotation level
1	Del-3,5-O-diGlc	9.9	627.15536	-0.21	C27H31O17	465, 303	(Liang et al., 2008;He et al., 2010)	2
2	Cy-3, 5-O-diGlc	11.5	611.16031	-0.4	C27H31O16	449, 287	(Liang et al., 2008;He et al., 2010)	2
3	Procyanidin B2	12.07	579.14927	-0.43	C30H27O12	427, 409, 291, 247	(Monagas et al., 2005)	1
4	Pt-3, 5-O-diGlc	12.5	641.17117	-0.06	C28H33O17	479, 317	(Liang et al., 2008;He et al., 2010)	2
5	DHQ-3-O-hexoside	12.5	467.11823	-0.17	C21H23O12	449, 305	(Flamini et al., 2015)	2
6	Pn-3, 5-O-diGlc	14.25	625.17598	-0.33	C28H33O16	463, 301	(Liang et al., 2008;He et al., 2010)	2
7	Mv-3, 5-O-diGlc	14.7	655.18679	-0.09	C29H35O17	493, 331	(Liang et al., 2008;He et al., 2010)	2
8	Dp-3-O-Glc	14.9	465.10259	-0.16	C21H21O12	303	(Liang et al., 2008)	2
9	Procyanidin B like	15.58	579.14934	-0.37	C30H27O12	427, 289, 291, 409	(Monagas et al., 2005)	3
10	MDHQ-hexoside	15.91	481.13399	-0.06	C22H25O12	463, 319, 301	(Abu-Reidah et al., 2015)	3
11	Dp-3-O-acetylGlc-5-O-Glc	16.3	669.16632	0.18	C29H33O18	507, 303, 465	(He et al., 2010)	2
12	Procyanidin B5	16.61	579.1494	-0.31	C30H27O12	427, 409, 291, 247	(Monagas et al., 2005)	2
13	Lut/Kae/Cy-methylGlu-hexoside	16.7	639.15567	0.09	C28H31O17	477, 287, 449		3
14	Cy-3-O-Glc	16.88	449.10768	-0.16	C21H21O11	287	(Liang et al., 2008)	2
15	Cy-3-O-acetylGlc-5-O-Glc	17.73	653.17121	-0.02	C29H33O17	491, 287, 449	(Liang et al., 2008;He et al., 2010)	2
16	Pn-methylGlu-hexoside	18.46	653.17117	-0.06	C29H33O17	491, 301, 463		3

17	Mv-methylGlu-glucoside	18.46	683.18171	-0.08	C30H35O18	521, 331, 317, 493	3
18	Pn-3-O-Glc	19.65	463.12328	-0.21	C22H23O11	301 (Liang et al., 2008)	2
19	Pn-3-O-acetylGlc-5-O-Glc	20.32	667.18708	0.2	C30H35O17	505, 301, 463 (He et al., 2012)	2
20	Myr-3-O-Glu	22.69	495.07682	-0.11	C21H19O14	319 (Flamini et al., 2015)	2
21	Myr-3-O-Glc	22.85	481.09749	-0.18	C21H21O13	319 (Flamini et al., 2015)	2
22	Dp-3-O-acetylGlc	23	507.11316	-0.16	C23H23O13	303 (Liang et al., 2008)	2
23	Lut/Kae/Cy-methylGlu Lut/Kae/Cy-Cou-hexoside or	23.21	477.10265	-0.21	C22H21O12	287	3
24	Lut/Kae/Cy-Caff-Rha	23.74	595.14454	-0.07	C30H27O13	287 (Liang et al., 2008; He et al., 2010)	3
25	Dp-3-O-CouGlc-5-O-Glc	24.13	773.1922	-0.16	C36H37O19	611, 303, 465 (Liang et al., 2008; He et al., 2010)	2
26	Lar-3-O-Glu	24.8	509.09238	-0.2	C22H21O14	319 (Flamini et al., 2015)	2
27	Cy-3-O-acetylGlc	25.04	491.11828	-0.12	C23H23O12	287 (Liang et al., 2008)	2
28	Epicatechin gallate	25.17	443.09711	-0.16	C22H19O10	273, 151, 291 (Delcambre and Saucier, 2012)	2
29	Pt-3-O-CouGlc-5-O-Glc	25.76	787.20786	-0.15	C37H39O19	625, 317, 479 (He et al., 2010)	2
30	Cy-3-O-CouGlc-5-O-Glc	26.1	757.19691	-0.54	C36H37O18	595, 287, 449 (Liang et al., 2008; He et al., 2010)	2
31	Que-3-O-Glu	26.4	479.08192	-0.1	C21H19O13	303 (Flamini et al., 2015)	2
32	Mv-3-O-CouGlc-5-O-Glc	27.84	801.22316	-0.49	C38H41O19	639, 611, 493, 331, 303 (Liang et al., 2008; He et al., 2010)	2
33	Pg-3-O-CouGlc-5-O-Glc	28.11	741.20284	0.31	C36H37O17	579, 271, 433 (He et al., 2012)	2
34	Pn-3-O-CouGlc-5-O-Glc	28.37	771.21278	-0.31	C37H39O18	609, 301, 463 (Liang et al., 2008; He et al., 2010)	2
35	Lut/Kae/Cy-(GlcCou)-methylGlu	28.76	785.19222	-0.14	C37H37O19	287, 477, 595, 623	3
36	Isorhamnetin-3-O-Glu	28.96	493.09758	-0.08	C22H21O13	303 (Flamini et al., 2015)	2
37	Dp-3-O-CouGlc	29.27	611.13932	-0.21	C30H27O14	303 (He et al., 2010)	2
38	Cinchonain Ia	29.32	453.11789	-0.24	C24H21O9	343, 313, 301 (Chen et al., 2015)	2
39	Kae-3-O-Glu	29.42	463.08706	-0.04	C21H19O12	287 (Flamini et al., 2015)	2
40	Pn-(Glc-Cou)-methylGlu	30.29	799.20804	0.04	C38H39O19	609, 301, 491, 637	3

41	Cy-3-O-CouGlc	30.67	595.14423	-0.38	C30H27O13	287	(Liang et al., 2008;He et al., 2010)	2
42	Pt-3-O-CouGlc	30.98	625.15504	-0.14	C31H29O14	317	(Liang et al., 2008;He et al., 2010)	2
43	Isorhamnetin-3-O-Glc	31.2	479.11835	-0.05	C22H23O12	317	(Flamini et al., 2015)	2
44	Phloretin	31.38	275.09134	-0.23	C15H15O5	107, 169, 149, 127	(Zhao et al., 2014)	2
45	Cinchonain 1b	31.52	453.11784	-0.38	C24H21O9	343, 301, 313	(Chen et al., 2015)	2
46	Pg-3-O-Cou-Glc	31.91	579.1496	-0.1	C30H27O12	271	(Castillo-Munoz et al., 2009)	2
47	Pn-3-O-CouGlc	32.23	609.16016	-0.11	C31H29O13	301	(Liang et al., 2008;He et al., 2010)	2
48	Mv-3-O-CouGlc	32.23	639.17083	-0.01	C32H31O14	331	(Liang et al., 2008;He et al., 2010)	2
49	Dihydrokaempferide-3-Glu	35.83	479.11835	-0.06	C22H23O12	289		3

940

941 Abbreviation: Api, Apigenin; Caff, Caffeoyl; Cou, Coumaryl; Cy, Cyanidin; DHQ, Dihydroquercetin (taxifolin); Dp, Delphinidin; Glc, Glucose; Glu, Glucuronide; Kae,

942 Kaempferol; Lar, Laricitrin; Lut, Luteolin; MDHQ, methylidihydroquercetin; Mv, Malvidin; Myr, Myricetin; Pg, Pelargonidin; Pn, Peonidin; Pt, Petunidin; Que, Quercetin; Rha,
943 Rhamnoside; Tri, Tricetin.

944

945

946

947

Supplementary Material

Figure S1. Relative abundance of anthocyanins in mature berries (EL 38)

Figure S2. Hierarchical bi-clustering analysis of protein candidates

Figure S3. K-means clustering analysis of the integrated dataset.

Figure S4. Granger causality based network.

Table S1. Metabolite information

Table S2. Absolute quantification of 29 metabolite candidates

Table S3. Protein sequencing information

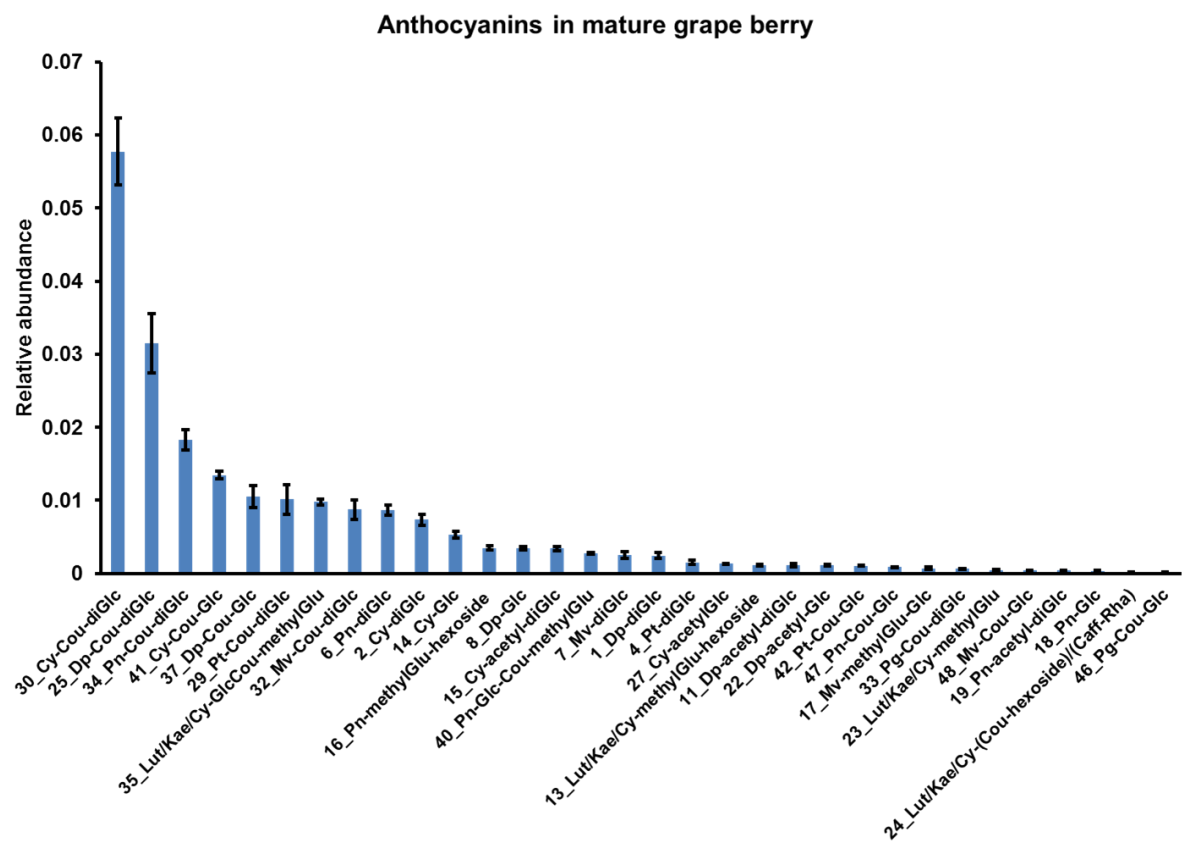
Table S4. NSAFs, ANOVA, BLAST results of protein candidates

Table S5. Loading scores

Table S6. K-means clustering analysis (12 clusters)

Note: Supplementary tables could be downloaded from email box:
thesis.supplements.gmail.com, passwords: csupplements.

963

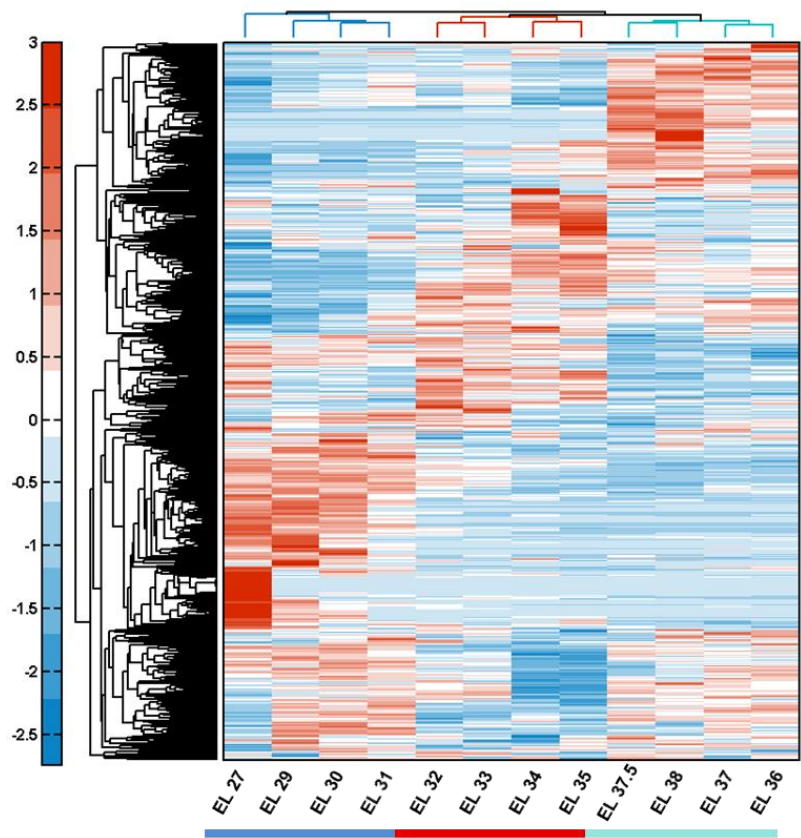


964

965 Figure S1. Relative abundance of anthocyanins in mature berries (EL 38)

966

967



968

969

970

Figure S2. Hierarchical bi-clustering analysis of protein candidates exist in all the 3 replicates of at least one stage.

971

972

973

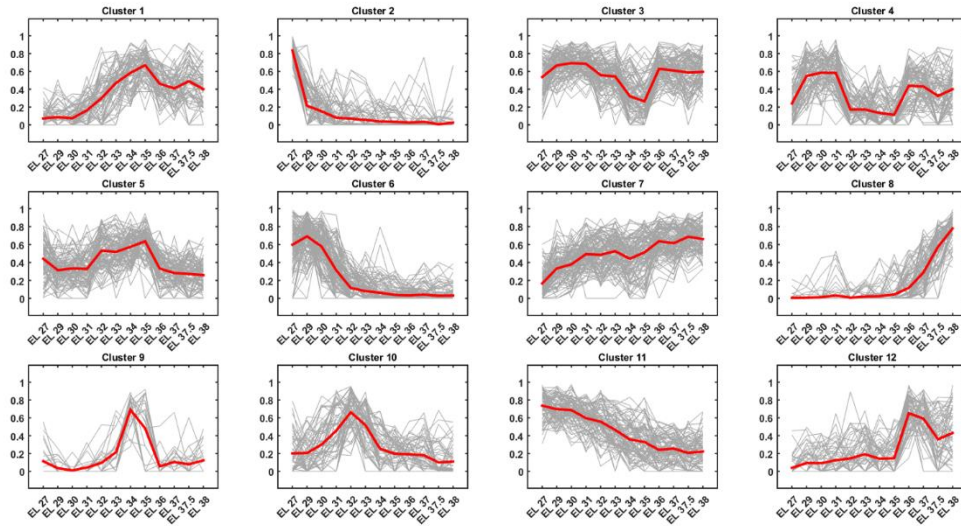


Figure S3. K-means clustering analysis of the integrated dataset. The bold red lines indicate the averaged dynamic patterns of the 12 clusters.

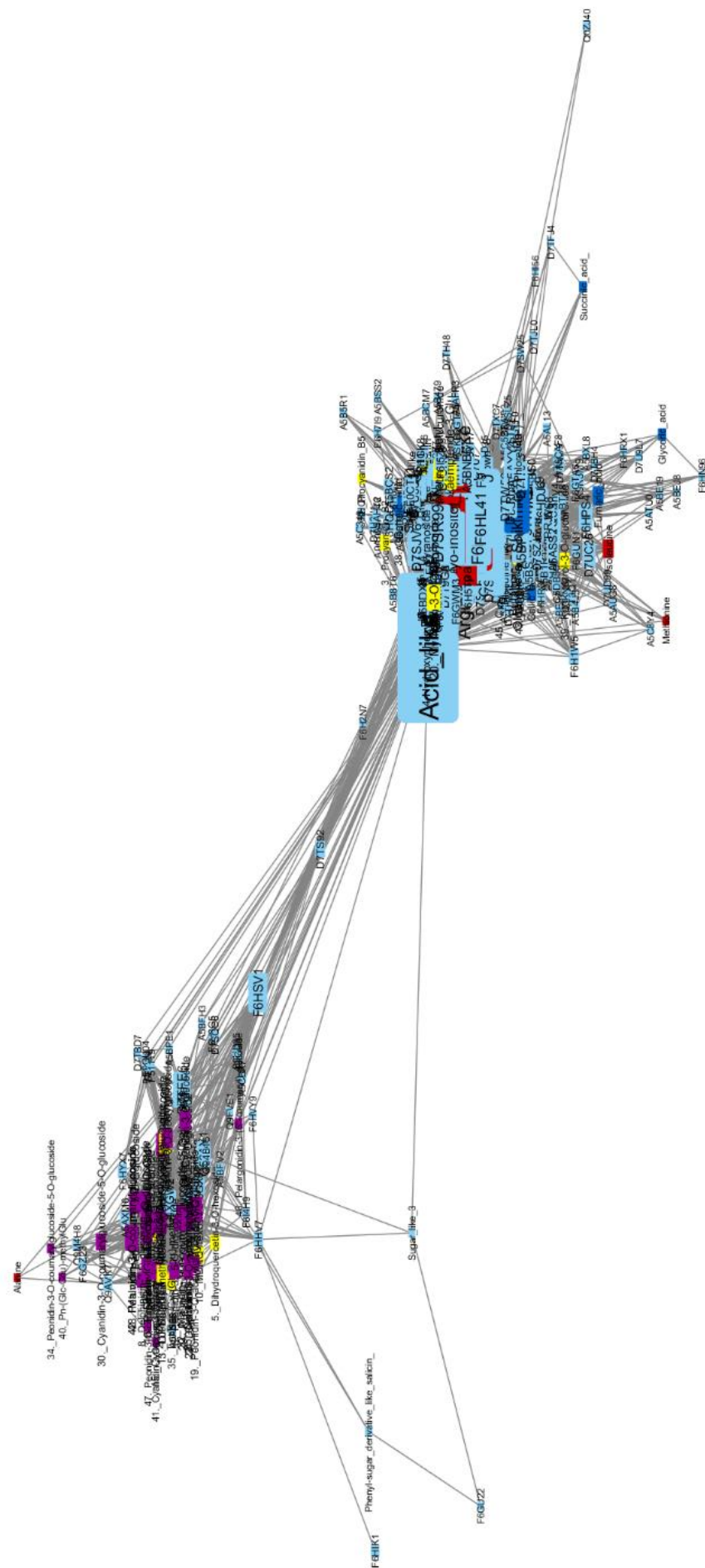


Figure S4. Granger causality based network. The font size and labeling size were set according to the node degrees. The amino acids were labeled with red, organic acids with dark blue, anthocyanins with purple and other flavonoids with red. The nodes with less than 5 degree were filtered out

III. Summary

Plants play vital roles in the natural web of life. Three plants are in the focus of this thesis including *Theobroma cacao*, *Nelumbo nucifera* and *Vitis vinifera*. These historic crops possess great cultural, medical and economic values. Cacao seeds are the source for chocolate production. Lotus (*N. nucifera*) seeds are widely consumed in Asia countries as snacks and medicine. Grape (*V. vinifera*) berries are either consumed freshly or made into juice or wine. These most valuable consuming parts, i.e. cacao seeds, lotus seeds and grape berries, are therefore the main study objects. Common to all is their high nutritional value in human diet with respect to secondary metabolites which are end products of complex developmental processes. The development and ripening process of seeds and fruits involves morphological as well as physiological and biochemical changes. Here, mass spectrometry (MS) based platforms were applied to obtain dynamic metabolome and proteome profiles during development. For increased metabolomic coverage with a focus on the interface of primary and secondary metabolism as well as natural product synthesis, gas chromatography coupled with time of flight MS (GC-TOF-MS) and liquid chromatography coupled with Orbitrap MS/MS (LC-Orbitrap-MS/MS) platforms were used. Furthermore, LC-Orbitrap-MS/MS analysis was used for shotgun proteomics. For the shotgun proteomics approach, the genome sequences of *T. cacao*, *N. nucifera* and *V. vinifera* were used. Multivariate statistical analysis such as principal component analysis (PCA), clustering analysis and Granger causality analysis were applied to assist data exploration.

In the first study, cacao seeds (*T. cacao* cv. Forastero) at four developmental stages (stage 1 to 4) were harvested from Costa Rica. The analysis of the metabolome and proteome data of developing cacao seeds revealed a complex consecutive coordination of primary and secondary metabolism. From embryo to mature beans, cacao seeds undergo cell division and expansion (stage 1 and 2), reserve accumulation (stage 3) and maturation phases (stage 4). At early developmental stages (stage 1 and 2), the transported sucrose was predominantly hydrolyzed to fructose and glucose leading to an increase in these hexoses. Further, glycolysis and the tricarboxylic acid (TCA) cycle showed enhanced perturbations by a metabolic modelling approach suggesting the liberation of energy and carbon skeletons for the synthesis of amino acids, lipids and secondary metabolites. At stage 3, sucrose metabolism switched to raffinose synthesis and resulted in the accumulation of raffinose that plays roles as energy storage reserves and osmoprotectants. Purine alkaloids such as theobromine and caffeine started to accumulate since stage 3 endowing cacao seed bitter taste and astringent mouthfeel. Most of the amino acids, on the contrary, declined at this stage due to the incorporation into synthesis of proteins and lipids. Flavonoids and N-phenylpropanoyl amino acids (NPPAAs) that both derive from phenylpropanoid pathway showed distinct changing patterns. All the

detected NPPAAs decreased during cacao seed development whereas most of the detected flavonoids accumulated in mature seeds. Stage-specific comparison of co-variance derived Jacobian entries also indicated the reprogramming of flavonoids and NPPAAs metabolism. Significant shift of proteins from phenylpropanoid metabolism to flavonoid synthesis from stage 1 to stage 2 indicated the preparation of precursors and enzymes for flavonoid synthesis at early developmental stages. Finally, in the mature seeds, the accumulated raffinose, flavonoids together with high abundance of stress related proteins synergistically protected the seeds against biotic and/or abiotic stresses.

Lotus seeds (*N. nucifera* cv. Baijiantian) are consumed either freshly or in a desiccated form. It is essential for lotus production to address the transition phase from the fresh consumption stage to the reserve filling stage. Our study focused on metabolic dynamics of developing lotus seed from 10 to 25 days after pollination (DAP) which correspond to the reserve accumulation phase. The volume of lotus seeds continuously expanded until 20 DAP. Soluble sugars such as glucose and fructose were most abundant around 15 DAP and decreased afterwards. Histological results showed a rapid accumulation of starch and proteins from 17 to 20 DAP indicating a metabolic switch to a reserve accumulation phase during this period. Therefore, lotus seeds at around 17 DAP with a relatively high sweetness are suitable for fresh consumption whereas those harvested after 20 DAP with high content of starch are better to be processed to snacks or used in cuisines. The dynamic profiles of metabolites and proteins involved in sugar metabolism, glycolysis pathway, TCA cycle and amino acid metabolism provide a reference data set for understanding the primary metabolism during lotus seed development.

The comparison of developmental processes of cacao seeds and lotus seeds showed similarities as well as differences. Like other dicotyledonous plants, cacao and lotus seeds went through similar developmental stages i.e. cell division and expansion stage, reserve accumulation stage and maturation stage. Their sugar metabolism switched from energetic mode to energy storage mode from cell division and expansion phase to reserve accumulation phase. Both of them accumulated proteins during their ripening. However, their preferred reserves are different. Cacao seeds predominantly accumulated lipids and purine alkaloids whereas lotus seeds mainly accumulated starch.

In the last study, proteome and metabolome profiles were obtained from grape berries (*V. vinifera* cv. Early Campbell) at 12 developmental stages from fruit set until maturation. PCA of the integrated metabolomic and proteomic data matrix showed a pronounced developmental trajectory with a turning point around veraison. The organic acids involved in central carbon metabolism increased before veraison and decreased afterwards. During ripening, sugars, especially fructose together with anthocyanins strongly accumulated endowing grape berry

sweet taste and reddish color. The similarity of anthocyanin profile of Early Campbell with other Concord derived varieties emphasized the strong influence of the genetic background to metabolism. The content of other flavonoid subfamilies such as proanthocyanins, flavan-3-ols, flavonols, flavanonols and their glycosides decreased due to the competition of precursors with anthocyanins. For the first time a novel type of statistical analysis called “Granger causality analysis” was applied for interpreting directed and non-linear correlations between metabolites and proteins during the time course of grape berry development. The results revealed significant Granger causations between flavonoids with their precursors and enzymes indicating time-lagged sequential coordination of precursors, enzymes and products. A comprehensive Granger causality network showed three main clusters with several specific proteins, amino acids (ornithine, arginine, phenylalanine and several others) as well as organic acids (e.g. shikimic acid) as hubs highlighting their central position in grape berry metabolism during development. The consecutive k-means clustering analysis with Granger causality analysis presented a general shift in metabolism from amino acids, sugars and flavonoids to organic acids accumulation and finally to lipid and anthocyanin accumulation from grape berry fruit set up to maturation.

High throughput mass spectrometry based platforms facilitated the exploration of genome-wide metabolic dynamics during seed and fruit development. These results not only present the developmental processes of cacao seed, lotus seed and grape berry on both metabolite and protein level but also provide reference data sets for further study. Multivariate statistical analysis assisted in the interpretation of biological information and revealed complex metabolic interaction networks and novel metabolic relationships from genome-scale datasets during developmental processes. Better understanding of these developmental processes will be helpful in crop growth monitoring, breeding and quality control.

IV. Zusammenfassung

Der Fokus dieser Arbeit liegt auf drei Nutzpflanzen, welche seit jeher großen kulturellen, medizinischen und ökonomischen Wert besitzen: *Theobroma cacao*, *Nelumbo nucifera* and *Vitis vinifera*. Kakaobohnen (*T. cacao*) sind der Grundstoff für Schokoladeproduktion. Lotussamen (*N. nucifera*) werden in Asien gerne als Snacks verzehrt, aber spielen auch als Arzneimittel eine Rolle. Trauben (*V. vinifera*) werden entweder frisch verzehrt oder zu Saft und Wein weiterverarbeitet. Deshalb stellen diese, vom Menschen benutzten Teile der Pflanzen die Studienobjekte der vorliegenden Arbeit dar. Ihnen ist gemeinsam, dass sie hohe Konzentrationen an Sekundärmetaboliten enthalten, welche einen großen Beitrag zu einer gesunden Ernährung beitragen. Diese Stoffe werden in den Samen und Früchten während des Entwicklungsprozesses gebildet, welcher komplexe morphologische, physiologische und biochemische Änderungen mit sich bringt. In dieser Studie wurden massenspektrometrische (MS) Methoden angewandt um dynamische Profile von Metabolom und Proteom im Laufe der Fruchtentwicklung aufzuzeichnen. Um eine möglichst effiziente und umfangreiche Erhebung der Primär- und Sekundärmetabolite zu gewährleisten wurde ein an eine gaschromatografische Auftrennung gekoppeltes „Time of Flight“ Massenspektrometer (TOF), sowie eine an eine flüssigkeitschromatografische Auftrennung gekoppelte Orbitrap MS/MS (LC-Orbitrap-MS/MS) eingesetzt. Das Orbitrap MS/MS System wurde auch für die shotgun-Proteomik Analyse verwendet. Um die Ergebnisse dieser shotgun-Proteomik interpretieren zu können wurden die Genomsequenzen der jeweiligen Pflanzen zu Hilfe gezogen. Multivariate statistische Analysen, wie Hauptkomponentenanalyse, Cluster Analyse, sowie die Berechnung der Granger-Kausalität wurden zur Datenauswertung herangezogen.

Für die erste Studie wurden Kakaobohnen (*T. cacao* cv. Forastero) in vier verschiedenen Entwicklungsstadien (Stadium 1-4) in Costa Rica geerntet und verarbeitet. Die Analyse des Metaboloms und des Proteoms offenbarte eine komplexe Koordination von Primär- und Sekundärmetabolismus. Vom Embryo bis zur Reife durchlaufen Kakaobohnen mehrere Stadien: Zellteilung und Streckung (Stadium 1 und 2), Einlagerung von Speicherstoffen (Stadium 3) und die finale Reifungsphase (Stadium 4). In den frühen Entwicklungsstadien (Stadium 1 und 2) wurde der Transportzucker Saccharose aus dem vegetativen Teil der Pflanze in Fruktose und Glukose gespalten, was zu einem Anstieg der Level dieser Hexosen in den Früchten führte. Weiters zeigten Glykolyse und der Zitratzyklus in einem Modellierungsansatz des Metabolismus verstärkte Perturbationen, was durch die Freisetzung von Energie und Kohlenstoffbausteinen für die Synthese von Aminosäuren, Lipiden und Sekundärstoffen zu erklären war. In der dritten Phase wechselte der Saccharose Stoffwechsel zu Raffinose Synthese, was wiederum zur Akkumulation von Raffinose führte, welche eine wichtige Rolle als Energiespeicher und kompatibles Solut spielt. Purinalkaloide, wie Theobromin und Koffein,

welche für den charakteristischen, bitteren Geschmack und das adstringierende Mundgefühl von Kakaobohnen verantwortlich sind, wurden ab dem dritten Stadium angereichert. Die Konzentrationen der meisten Aminosäuren, auf der anderen Seite, sanken ab diesem Stadium, da sie für den Aufbau von Proteinen und Lipiden verwendet wurden. Flavonoide und N-Phenylpropenoyl- Aminosäuren (NPPAAs), welche aus dem Phenylpropanoid Metabolismus stammen, wiesen charakteristische Änderungen in ihren Konzentrationen auf. Die Konzentrationen aller detektierten NPAAs sank während der Entwicklung der Kakaobohnen, während die meisten Flavonoide in den reifen Samen akkumulierten. Der Vergleich der aus der Kovarianz der Metabolitlevels abgeleiteten Einträge der Jakob-Matrizen wies auch auf eine Neuprogrammierung des Flavonoid- und Phenylpropanoid Metabolismus hin. Eine signifikante Verlagerung der Proteine von Phenylpropanoid Metabolismus zu Flavonoidsynthese von Stadium 1 zu Stadium 2 ließ auf die Vorbereitung des Enzymapparates für die Akkumulation von Flavonoiden in den frühen Stadien schließen. In den reifen Samen dienen die akkumulierten Stoffe wie Raffinose und Flavonoide, zusammen mit Stress-assoziierten Proteinen der Verteidigung gegen biotische und abiotische Stressoren.

Lotussamen (*N. nucifera* cv. Baijianlian) werden entweder frisch oder in getrockneter Form verzehrt. Es ist bei der Ernte von Lotus essentiell den richtigen Zeitpunkt zu wählen, da die Samen, wenn sie zu spät geerntet werden stark an Süße verlieren. Der Schwerpunkt dieser Studie lag darauf, Dynamiken im Metabolismus in Entwicklungsstadien von Lotussamen von 10 bis 25 Tagen nach der Befruchtung (DAP – Days after Pollination) zu untersuchen. Diese Phase entspricht jener Zeit, in der Reservestoffe im Samen eingelagert werden. Das Volumen der Samen stieg kontinuierlich bis 20 DAP und sank danach. Die Konzentration löslicher Zucker wie Glukose und Fruktose erreichte das Maximum um 15 DAP und sank danach wieder ab. Histologische Befunde zeigten eine schnelle Akkumulation von Stärke und Proteinen von 17- 20 DAP, was einen Umschlagpunkt im Metabolismus zur Akkumulation von Reservestoffen in diesem Zeitraum andeutete. Daraus folgt, dass Lotussamen um 17 DAP aufgrund ihrer relativ hohen Süße jenen, die nach 20 DAP geerntet wurden, zum frischen Verzehr vorzuziehen sind. Die nach 20 DAP geernteten Samen jedoch sind durch den hohen Stärkegehalt besser dafür geeignet um kulinarisch weiterverarbeitet zu werden. Die dynamischen Profile der Metabolite und Proteine, welche in Zuckerstoffwechsel, Glykolyse, Zitratzyklus und Aminosäurestoffwechsel involviert sind stellen einen Referenzdatensatz dar, der dabei behilflich sein wird den Primärstoffwechsel in der Entwicklung von Lotussamen besser zu verstehen.

Der Vergleich der Entwicklungsprozesse von Kakaobohnen und Lotussamen zeigte sowohl Ähnlichkeiten, als auch Unterschiede. So wie andere zweikeimblättrige Pflanzen durchlaufen Kakao- und Lotussamen ähnliche Entwicklungsstadien: Zellteilung und Expansion,

Akkumulation von Speicherstoffen und eine Reifungsphase. Die Aufgabe des Zuckerstoffwechsels verschob sich vom Bereitstellen von Energie zum Anlegen einer Energiereserve. Beide Studienobjekte akkumulierten Proteine im Reifungsprozess, jedoch die Zusammensetzung der gespeicherten Metabolite unterschied sich. In Kakaobohnen wurden vorwiegend Lipide und Purinalkaloide als Reservestoffe eingelagert, während Lotussamen auf Stärke setzten.

In der letzten Studie wurden die Metabolom- und Proteomprofile von 12 Entwicklungsstadien von Weintrauben (*V. Vinifera* cv. Early Campbell) verglichen. Eine Hauptkomponentenanalyse der kombinierten Datenmatrix aus Metabolit und Proteinmessungen zeigte eine klare Entwicklungstrajektorie mit einem Wendepunkt um die *Véraison*. Die Konzentrationen der organischen Säuren des zentralen Kohlenstoff-Stoffwechsel stiegen bis zum Punkt der *Véraison* und sanken danach wieder ab. Während des Reifeprozesses wurden Zucker und Anthocyanine in großen Mengen akkumuliert, was den Trauben zu ihrem süßen Geschmack und ihrer roten Farbe verhalf. Die Ähnlichkeiten im Anthocyaninprofil im Lauf der Entwicklung der Sorte Early Campbell mit anderen von der Sorte Concord abgeleiteten Weintraubensorten zeigte den starken Einfluss des genetischen Hintergrunds auf den Stoffwechsel. Die Menge an anderen Flavonoid- Unterfamilien, wie Proanthocyanine, Flavan-3-ole, Flavonole, Flavanole und deren Glycoside verringerte sich aufgrund des Wettbewerbs um Vorstufen und Grundbausteine mit Anthocyaninen. Zum ersten Mal wurde in diesem Zusammenhang eine „Granger Kausalitäts Analyse“ angewendet um direkte und nichtlineare Zusammenhänge zwischen Metaboliten und Proteinen im Laufe der Entwicklung der Trauben zu analysieren. Die Ergebnisse zeigten signifikante Zusammenhänge zwischen Flavonoiden, deren Vorstufen und den Enzymen, welche in der Flavonoidsynthese involviert sind. Diese Zusammenhänge waren über die Entwicklungszeit koordiniert und sequentiell angeordnet. Ein ausführliches Granger Kausalitätsnetzwerk zeigte drei Cluster mit mehreren spezifischen Proteinen, Aminosäuren (Ornithine, Arginine, Phenylalanine und einige weitere), sowie organische Säuren (z.B. Shikimisäure) als Knotenpunkte auf, was ihre zentrale Position im Metabolismus der Trauben hervorhob. Diese Ergebnisse zeigten, zusammen mit aufeinanderfolgenden k-means Clusterings, im Lauf der Entwicklung eine generelle Verschiebung im Stoffwechsel von Aminosäuren, Zuckern und Flavonoiden zur Einlagerung von organischen Säuren und letztendlich der Akkumulation von Lipiden und Anthocyanin.

Moderne, massenspektrometriebasierte Plattformen erleichterten das Verständnis von umfassenden metabolischen Dynamiken in der Entwicklung von Samen und Früchten. Diese Resultate repräsentieren nicht nur die besprochenen Entwicklungsprozesse, sondern stellen auch Referenzdatensätze für folgende Studien dar. Multivariate statistische Analysen ermöglichten die Interpretation von biologischen Informationen und deckten komplexe

Stoffwechselinteraktionsnetzwerke und bisher unbekannte Zusammenhänge auf. Ein besseres Verständnis dieser Entwicklungsprozesse wird bei der Kontrolle des Wachstums von Nutzpflanzen, sowie deren Zucht eine große Rolle spielen.