JOURNAL OF PROTEOMICS 75 (2011) 284–305

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Review

Green systems biology — From single genomes, proteomes and metabolomes to ecosystems research and biotechnology

Wolfram Weckwerth

Department of Molecular Systems Biology (MOSYS; http://www.univie.ac.at/mosys/), University of Vienna, Althanstrasse 14, 1090 Vienna, Austria

ARTICLE INFO

Article history:
Received 6 April 2011
Accepted 10 July 2011
Available online 23 July 2011

Keywords:
CIMMYT
INRRI
Natural variation
Biodiversity
Biofuels
Financial market
Land grabbing
Marker-assisted selection (MAS)
Genome-assisted breeding (GAB)
Ecosystems
Ecology
Ecophysiology
Genotype
Phenotype
Phenotypic plasticity
Plant systems biology
Modeling
Genome annotation
Green revolution
Next generation sequencing (NGS)
Genome-wide associations (GWA)
Single nucleotide polymorphisms (SNP)

ABSTRACT

Plants have shaped our human life form from the outset. With the emerging recognition of world population feeding, global climate change and limited energy resources with fossil fuels, the relevance of plant biology and biotechnology is becoming dramatically important. One key issue is to improve plant productivity and abiotic/biotic stress resistance in agriculture due to restricted land area and increasing environmental pressures. Another aspect is the development of CO_2-neutral plant resources for fiber/biomass and biofuels: a transition from first generation plants like sugar cane, maize and other important nutritional crops to second and third generation energy crops such as *Miscanthus* and trees for lignocellulose and algae for biomass and feed, hydrogen and lipid production. At the same time we have to conserve and protect natural diversity and species richness as a foundation of our life on earth. Here, biodiversity banks are discussed as a foundation of current and future plant breeding research. Consequently, it can be anticipated that plant biology and ecology will have more indispensable future roles in all socio-economic aspects of our life than ever before. We therefore need an in-depth understanding of the physiology of single plant species for practical applications as well as the translation of this knowledge into complex natural as well as anthropogenic ecosystems. Latest developments in biological and bioanalytical research will lead into a paradigm shift towards trying to understand organisms at a systems level and in their ecosystemic context: (i) shotgun and next-generation genome sequencing, gene reconstruction and annotation, (ii) genome-scale molecular analysis using OMICS technologies and (iii) computer-assisted analysis, modeling and interpretation of biological data. Systems biology combines these molecular data, genetic evolution, environmental cues and species interaction with the understanding, modeling and prediction of active biochemical networks up to whole species populations. This process relies on the development of new technologies for the analysis of molecular data, especially genomics, metabolomics and proteomics data. The ambitious aim of these non-targeted 'omic' technologies is to extend our understanding beyond the analysis of separated parts of the system, in contrast to traditional reductionistic hypothesis-driven approaches. The consequent integration of genotyping, pheno/morphotyping and the analysis of the molecular phenotype using metabolomics, proteomics and transcriptomics will reveal a novel understanding of plant metabolism and its interaction with the environment. The analysis of single model systems – plants, fungi, animals and bacteria – will finally emerge in the analysis of populations of plants and other organisms and their

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

doi:10.1016/j.jprot.2011.07.010

adaptation to the ecological niche. In parallel, this novel understanding of ecophysiology will translate into knowledge-based approaches in crop plant biotechnology and marker- or genome-assisted breeding approaches. In this review the foundations of green systems biology are described and applications in ecosystems research are presented. Knowledge exchange of ecosystems research and green biotechnology merging into green systems biology is anticipated based on the principles of natural variation, biodiversity and the genotype–phenotype environment relationship as the fundamental drivers of ecology and evolution.

Contents

1. Introduction: from early plant breeding and the green revolution to green systems biology

Plant breeding has a very long history. Plants have always served as a resource for human feeding. Systematic usage goes back more than 50,000 years [1]. Accordingly, Paleolithic and Neolithic proto-farmers soon learned to improve their crops by selection and breeding [1]. The basic principles of plant breeding are simple: the requirements are genetic variation in a population – natural variation (see also chapter 5)–and the means to identify and to select the most suitable variants and traits. In the 18th century breeders started to systematically produce new genetic variation by hybridization. Ever since these beginnings, (see Chapter 12.1) the general principles of exploiting genetic variation and selection for plant breeding have not changed [1]. Besides hybridization, mutagenesis and many other technologies, natural variation was always an essential element of the implementation of novel genetic variation and producing new plant varieties.

One of the most impressive examples was initiated by Norman Borlaug in 1945. He started the “green revolution” as a result of a consequent funding initiative on the part of public research institutes and the foundation of the CIMMYT [1]. After 20 years of systematic trait selection breeding based on natural varieties of wheat, Mexico converted from an

A

Land plants

Completed Large-Scale Sequencing Projects

Arabidopsis thaliana (thale cress)
5 chromosomes: 1, 2, 3, 4, 5, plastid, mitochondrion
Glycine max (soybean)
20 chromosomes: A1, A2, B1, B2, C1, C2, D1a, D1b, D2, E, F, G, H, I, J, K, L, M, N, O, plastid, mitochondrion
Medicago truncatula (barrel medic)
8 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8 plastid, mitochondrion
Oryza sativa (rice)
12 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, plastid , mitochondrion, mitochondrial plasmid B1, mitochondrial plasmid B2
Populus trichocarpa (black cottonwood)
19 chromosomes: I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, plastid, mitochondrion
Sorghum bicolor
10 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, plastid, mitochondrion
Vitis vinifera (wine grape)
19 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, plastid, mitochondrion
Zea mays (corn)
10 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, plastid, mitochondrion, mitochondrial 1.9 kb plasmid

In-progress Large-Scale Sequencing Projects - funded, genome sequence expected in GenBank

Brachypodium distachyon
5 chromosomes: 1, 2, 3, 4, 5, plastid, mitochondrion
Carica papaya (papaya)
9 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, plastid, mitochondrion
Lotus japonicus (lotus)
6 chromosomes: 1, 2, 3, 4, 5, 6, plastid, mitochondrion
Manihot esculenta (cassava)
20 chromosomes: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, plastid, mitochondrion
Solanum lycopersicum (tomato)
12 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, plastid, mitochondrion
Solanum tuberosum (potato)
12 chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, plastid, mitochondrion

B

Green Algae

Abbreviations: GB - GenBank Accessions; PM - PubMed; R - RefSeq Accessions; G - Entrez Gene; T

11 Eukaryotic Genome Sequencing Projects Selected: Complete - 3, Assembly - 2, In Progress - 6

	Organism Information						Sequence Information				
GPID	Organism	Group	Subgroup	TaxID	Genome Size (Mb)	# Chr	Status	Method	Depth	Release Date	Center/Consort
12260	Chlamydomonas reinhardtii	Plants	Green Algae	3055	100	17	Assembly	WGS	12.8X	02/14/2004	DOE Joint Genome Institute
45853	Chlorella sp. NC64A	Plants	Green Algae	310507	46.2		In Progress	WGS			DOE Joint Genome Institute [more]
18715	Chlorella vulgaris	Plants	Green Algae	3077			In Progress	WGS			DOE Joint Genome Institute
32657	Coccomyxa sp. C-169	Plants	Green Algae	574566			In Progress	WGS			DOE Joint Genome Institute [more]
22771	Dunaliella salina CCAP 19/18	Plants	Green Algae	3046	130		In Progress	WGS			DOE Joint Genome Institute [more]
15678	Micromonas pusilla CCMP1545	Plants	Green Algae	564608	15		Assembly	WGS		04/07/2009	Micromonas Genome Consortium [more]
15676	Micromonas sp. RCC299	Plants	Green Algae	296587	21.09	17	Complete	WGS		04/10/2009	Micromonas genome consortium [more]
13044	Ostreococcus lucimarinus CCE9901	Plants	Green Algae	436017	13.25	21	Complete	WGS		04/10/2007	DOE Joint Genome Institute [more]
30933	Ostreococcus sp. RCC809	Plants	Green Algae	385169			In Progress	WGS			DOE Joint Genome Institute [more]
12912	Ostreococcus tauri OTH95	Plants	Green Algae	70448	12.5	20	Complete	WGS & Clone-based	7X	04/30/2005	Laboratoire Arago, France
13109	Volvox carteri f. nagariensis	Plants	Green Algae	3068	120		In Progress				DOE Joint Genome Institute

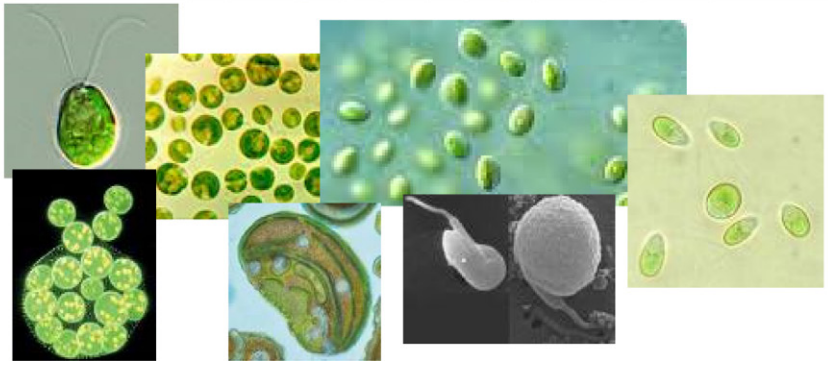

Fig. 1 – Progress of plant genome sequencing. A Genome sequences of plant model species in ecology and crop biotechnology. B Genome sequencing projects with algae.

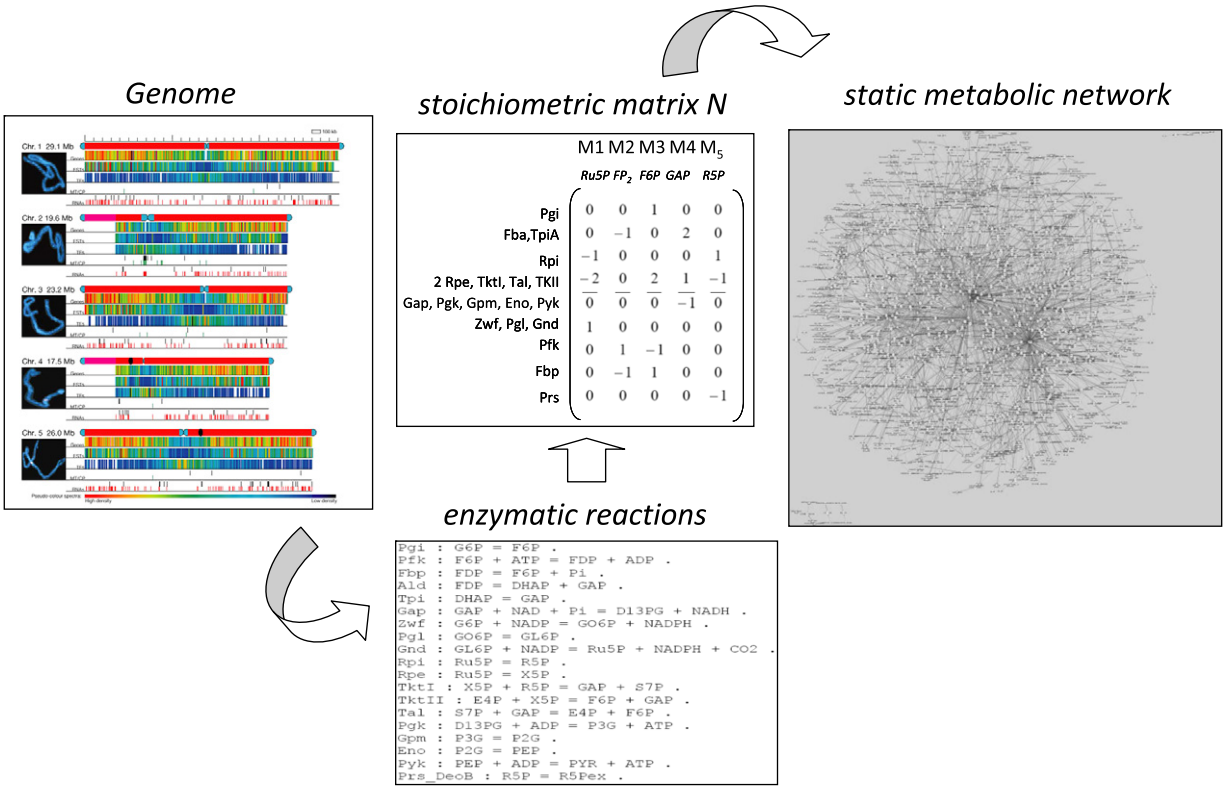

Fig. 2 – Genome-wide reconstruction of the regulatory and metabolic network in a sequenced organism. A key element is the so-called stoichiometric matrix N. This kind of information is genotype-specific, however, and only provides static information. The details of this approach and the relations to the dynamic phenotype (see below) are discussed in detail in [2].

importer to an exporter of wheat in the sixties (see also Chapter 13.1).

In the late 80s the first transgenic plants were created, today ~20% of crops worldwide are GMOs. Since the mid-80s the area of "AGBIOTECH" was monopolized by the private sector [1], leading to an imbalance between public and private research in plant breeding and the profit-oriented development of agricultural science, in contrast to a sustainability-driven development known from the "green revolution" and the public research institutes CIMMYT and IRRI and many more institutions consolidated by the CGIAR (reviewed in [1]) (see also Chapters 12.2 and 13).

The development of complete new technologies in biological research over the last 10 years – the foundations of green systems biology (see Chapter 2) – might lead to a refinement of plant ecology and evolution as well as classical breeding and biotechnological approaches thus addressing the following areas:

i. Productivity and stress adaptation of nutritional and energy crop plants, exploitation of natural variation, population dynamics and a better understanding of the genotype-phenotype relationship
ii. Exploitation of genome-sequenced plant model systems and their impact on improving plant productivity and stress adaptation; some model systems are *Arabidopsis thaliana*, *Brassica napus* for dicotyledons, *Chlamydomonas reinhardti*, *Chlorella* sp, *Botryococcus* sp., *Synechocystis* and others for algae and cyanobacteria, rice and maize for monocotyledons, poplar for trees and many more
iii. Genome-scale investigation of natural variations in their corresponding ecosystems to understand adaptation mechanisms and to provide fundamental knowledge for genetic variation, also for trait selection and genome/marker-assisted breeding approaches
iv. Addressing global climate change by CO_2-neutral biomass production (biomass crops) and renewable energy resources
v. Addressing biofuels, especially the transition from first-generation biofuels (corn, sugar cane, rape seed etc.) to second-generation (lignocellulose; Miscanthus, Poplar etc.) to third-generation biofuels (algae)
vi. Increase in public funding for agriculture to address global problems (see China's investments 2010; http://business.globaltimes.cn/china-economy/2010-03/510100.html) and national and international project-driven intensified cooperative knowledge transfer and exchange between the private and public sector for sustainability

In the next sections, individual aspects of this comprehensive task list are discussed, reviewed in the literature and commented upon to provide a solid basis for the future role of green systems biology.

2. Foundations of green systems biology

Three major developments have revolutionized biology [2]: (i) genome sequencing, (ii) the OMICS revolution and (iii) computer-assisted theoretical and modeling biology including the rapid development of the Internet into a knowledge platform and scientific database. First, genome sequencing has led to a repertoire of plant genome sequences which still has to be explored in its depth, starting with *Arabidopsis thaliana* as a first plant model system. Many achievements since the first release of the *Arabidopsis* genome sequence have justified all the efforts that try to understand a non-crop plant thoroughly [3–9]. Since then, many other plants, algae, cyanobacteria and photosynthetic active species have been sequenced or are in the process of being sequenced.

There is an exponential growth of genome sequences and this will increase even more due to novel sequencing technologies called next generation sequencing (NGS) (see following sections) [2,10].

Based on genome raw sequences derived through classical sequencing and shotgun genomics, genome assembly is a computer-based approach. After the assembly of a full genome, the next step is functional annotation. Predicted genes are searched for homology against databases of characterized genes and proteins. It is obvious that this initial functional annotation is not capable of producing a complete functional interpretation of the whole genome and a prediction of the molecular phenotype [2]. Consequently, the molecular phenotype needs to be measured for the functional interpretation of the genotype. These demands coincide with another technical development in biology called the "OMICS revolution". In summary, technologies such as transcriptomics, proteomics and metabolomics were developed which aim to analyze molecular data of living systems on a genome scale [9,11,12]. This leads to genome-scale, dynamic molecular data in combination with a genomic template. The ultimate goal is to derive a mathematical model of metabolism that is driven by genome data and predicts the phenotype and ecophysiology of the plant correctly [2].

Therefore, systems biology can be summarized as integrating experimental data, genome-scale reconstruction of metabolic networks and the derivation of mathematical models that are able to predict the molecular phenotype of the plant in its natural environment.

If we were able to develop several models of individual plant metabolisms based on this integrative approach we would be able to give much better functional interpretations of newly derived genome sequences and to identify new gene functions.

In the following sections I will shortly summarize the basic instruments for the integrative approach taken by green systems biology. I will present international efforts to consolidate these exceptionally complex research fields. In the last part I will show applications in ecosystem research, biotechnological research and conclude with connecting these different disciplines based on the fundamental assumptions of green systems biology.

3. Next generation sequencing and plant genomes

In the last 10 years improved sequencing methods have been developed, called next generation sequencing (NGS) (for overview see [10]). The demand for rapid and cost-effective sequencing technologies and a consequent funding policy for method development has led to the development of several alternative approaches in the use of genomic template libraries, number of reads, read length, genome coverage, the scale of the application and many other parameters. More importantly, NGS platforms have substantially lowered reagent costs and dramatically increased the throughput. As a result of these developments, the limitations of DNA sequencing have shifted from hardware to the software aspects. The strongest drawbacks of any of these technologies are short read lengths compared to Sanger sequencing (454/Roche: ~400 bases; Illumina/ABI-SOLiD: ~60 bases; Sanger sequencing ~1 kb, [13–16]) as well as different error characteristics. As a result, the assembly of genome sequences from these short reads is difficult, demands high computer power, novel algorithms and partial complementation and verification with high-quality sequencing strategies such as third-generation, long-read technology or Sanger sequencing [17–19].

After or during genome sequence assembly, gene prediction and functional annotation are the concomitant steps. *Ab initio* gene prediction with molecular data constraints is increasingly favored [20,21]. Here, especially NGS transcriptomics data can be used for gene prediction and functional annotation. Longer contig and singleton sequences are assembled from short reads and analyzed for homology with sequences in public databases using BLAST algorithms. Assembled contigs and singletons are subsequently translated into peptides and annotated with a biological function using a homology search against various public databases [19].

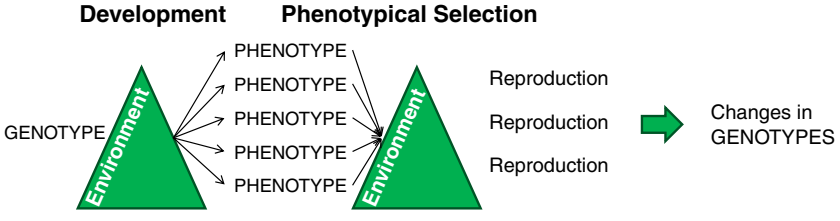

Fig. 3 – The genotype–phenotype relationship. A single genotype can produce several phenotypes depending on the environment. These phenotypes are subject to environment-dependent selection which leads to different rates of reproduction and changes in genotypes.

Due to these developments the number of sequenced genomes increased exponentially in general and also with regard to plant systems. Fig. 1 lists only a few examples with completed sequence projects and projects in progress, ranging from *Arabidopsis thaliana* with 5 chromosomes up to *Glycine max* with 20 chromosomes. It is foreseeable that in the near future any plant system of interest will be sequenced and the complete genome sequence will be available.

Consequently, the next level of investigation is genome interpretation [2]. Transcriptomics, proteomics and metabolomics data can be exploited for gene prediction and functional gene annotation in fully sequenced organisms [22–27] (see also Section 7.2). Major studies in plant model systems such as *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* have demonstrated the applications of proteo- and metaproteogenomics [22–25]. Here, very large proteomics datasets covering up to 60% or more of the predicted proteome are matched against genomics databases, especially 6 frame translations, to discover novel peptides which are not predicted by the assembled and functionally annotated genome sequence due to splice variants or completely missing annotations.

Knowledge generation in these plant model systems can be transferred into other plant systems. Newly sequenced plant systems can be interpreted by a homology search against other plant model systems. International activities and consortia have had a great impact on systematic functional elucidation of genome-scale gene function, especially MASC for *Arabidopsis*. Recently, a new initiative was founded exactly for this translational research in plant proteomics (INPPO). These activities are reviewed in chapter 9.

It is foreseeable that the classical approach of investigating a plant model system and translates this assembled knowledge to other non-sequenced plant systems will rather be substituted by the sequencing of any system of interest and the direct analysis of the genotype–phenotype relationship (see Chapter 8).

4. The dynamic genotype–phenotype relationship

Once the genome is sequenced and assembled it is possible to search for gene functions. Predicted genes are searched for sequence similarity in other organisms. In Fig. 2 the strategy for the genome-scale metabolic reconstruction is shown (more details can be found in [2]). Based on the detection of orthologeous genes in other organisms, many genes can be characterized only with reference to their homology. Enzymatic reactions can be postulated depending on these postulated gene functions. Educts and products participate in an enzymatic reaction. Pathways are structured so that the product of the former enzymatic

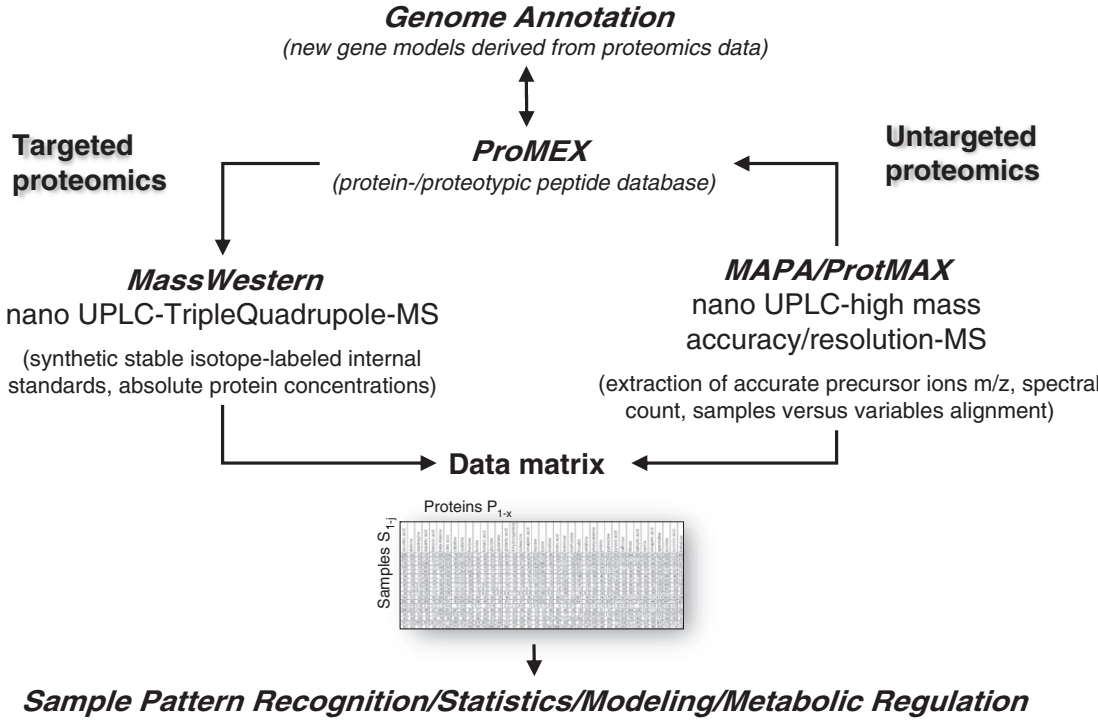

Fig. 4 – Overview of a proteomic toolbox for systems biology and genome annotation [25]. A central proteome/peptide spectral database (ProMEX, www.promexdb.org, [68]) serves as a basis for the selection of proteotypic peptides that are suitable for the targeted Mass Western analysis of complex proteome samples [61]. The MAPA method (MAPA = Mass Accuracy Precursor Alignment, [58]) allows for the detection and quantification of new proteins. Quantitative proteomics data will be aligned in a data matrix for statistical data mining.

reaction is the educt of the next enzymatic reaction, for instance in the Calvin Cycle. Thus, the list of reactions can be mapped to existing knowledge of pathways. Fragmentary pathways can be filled with reactions if the corresponding gene is not annotated in the genome sequence. Based on this reaction list, a stoichiometric matrix N can be built that is also known from chemical reaction lists. A metabolic network can be postulated for any organism (Fig. 2) on the basis of this stoichiometric matrix. Nowadays, the whole workflow can be automated [28]. Two recent studies have postulated genome-scale metabolic networks of *Arabidopsis thaliana* [29,30]. In another study the first metabolic draft network of *Chlamydomonas reinhardtii* was built on the basis of the newly sequenced genome of this unicellular green algae [24]. Some generic properties can be derived [29] from these reconstructions. Furthermore, using flux balance analysis (FBA) dynamic properties of this generic network can be postulated [31]. However, these initial draft metabolic networks are continuously improved using new knowledge about pathways and regulation. An example here is the metabolic reconstruction of yeast with many subsequent studies [32].

The databases of the metabolic reconstruction of plant species are continuously growing and will provide the basis for comparative plant genomics (http://www.plantcyc.org/).

The metabolic network can be predicted from the genome (see Fig. 2). As the genome is static information, the predicted metabolic network is static as well and therefore represents a model of all possible metabolic processes. Hence not all genes are constitutively expressed at the same time but differentially switched on and off. The metabolism is highly dynamic in the phenotype and cannot be directly derived from the genotype. Consequently, this static genomic information needs to be complemented with genome-wide molecular data to reveal the dynamic genotype–phenotype relationship [2] (see below and Chapter 8).

The dynamic genotype–phenotype relationship determines all functions of ecology and evolution. In Fig. 3 this relationship is simplified. A distinct genotype is interacting with its environment. This interaction can produce a variety of phenotypes which results in different success rates of reproduction and vice versa alterations in genotypes. Due to very recent developments in bioanalytical chemistry and plant biotechnology this intimate relationship of the genotype and the phenotype can now be investigated in much more detail [2]. SNP arrays or SNP-NGS are able to characterize genotypes directly or indirectly using genome-wide association studies (GWA) [3,33,34]. Phenotyping platforms can analyze growth, flowering, seed development, senescence and other parameters in an automated fashion. At the same time transcriptomics, proteomics and metabolomics (see [6,7,8]) will reveal the molecular dynamics of the phenotype.

The following chapters describe the framework for the measurement of the dynamic molecular phenotype and how it can be connected to the static genotype information. Based on the integration of genotype data, especially in conjunction with SNP measurements, a systematic investigation of this intimate relationship is possible by means of dynamic transcriptomic, proteomic and metabolomic data. Recently, a systematic approach was proposed explicitly on the basis of a genotype–phenotype-equation [2].

Ecology and evolution are intimately bound to processes of genetic variation and the genotype–phenotype relationship. Here, natural variation is the key driver and we are beginning to understand this relationship, especially with respect to phenotypic adaptation to the environment. Therefore, results from this research field have revealed important molecular processes which have given an insight into ecophysiological mechanisms and on the other hand will provide means in biotechnology and breeding for improving plant resistance against biotic and abiotic stresses. Eventually, due to the rapid development of bioanalytical technologies, we have completely new platforms for the investigation of natural variation and the genotype–phenotype relationship available. This is described in the next sections.

5. Natural variation

Recognition of the natural variation of plants is entirely bound to ecological and evolutionary research and is maybe one of the oldest investigative fields in human history. Natural variation – *e.g.* colors of ornamental flowers and thus also traits for breeding – was recognized and explored in early times by Mendel's investigations and the rediscovery of Mendel's laws and defined the genotype–phenotype relationship [35–38]. A classical phenotype is described with morphological and anatomical parameters [37]. However, any phenotype is of course a result of molecular changes based on genotype variation and leading into phenotype-environment variation (see Fig. 3). This is especially important for the functional interpretation of a genome. Accordingly, a gene's function should ideally be defined in the context of the system's state as a response to the environment [11]. The situation is very complex because a single genotype displays an array of different phenotypes depending on the environment: this is also called phenotypic plasticity [2,39,40]. The systematic investigation of genotypes and their variations with the integration of molecular data began lately [41]. Recent studies systematically investigated the relationship of natural variation, QTL mapping and metabolism [42–45]. Studies with genome-wide association using SNP-based microarrays enable the rapid mapping of traits [46]. In more recent studies, custom-made Affymetrix genotyping arrays containing up to 250,000 SNPs are used [47]. In combination with morphological and partly molecular phenotyping approaches, these SNP arrays reveal many common alleles with correlations to the phenotype. The molecular mechanisms are purely hypothetical, however [33]. Thus, the next logical step is to integrate these approaches with molecular profiling using NGS-RNA sequencing, epigenomics, proteomics and metabolomics. With the onset of these technologies and NGS, a dramatic increase in these research fields can be expected [2,39,48,49]. These bioanalytical platforms are described below.

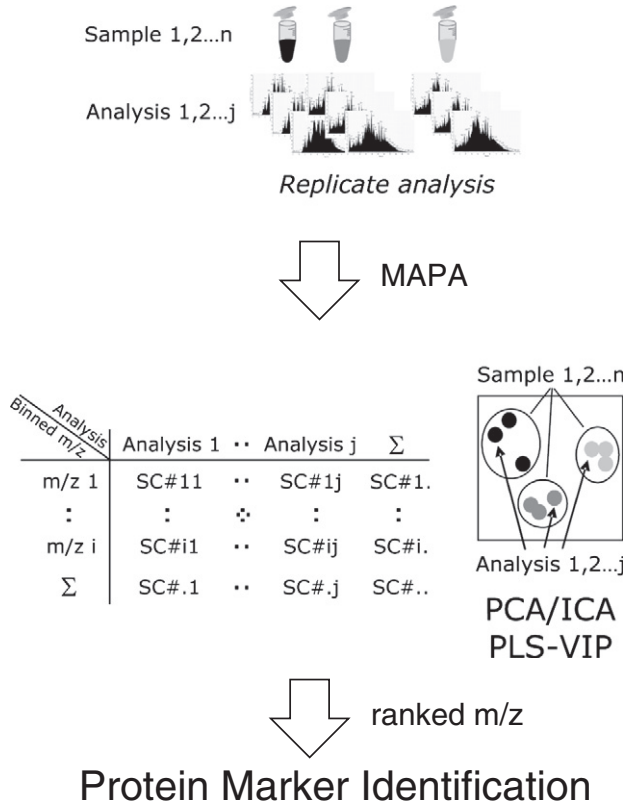

Fig. 5 – Schematic view of the MAPA process for rapid proteomic phenotyping and identification of phenotype-specific protein marker (for further details see [58] and Chapter 12.1).

6. OMICS technologies — the dynamic phenotype

6.1. *Transcriptomics*

The typical analysis of the dynamic transcriptome is usually performed with microarray technology and is one of the pioneering genome-scale, hypothesis-free screening methods. Several large-scale studies have revealed differential gene expression under different conditions and almost every gene in *Arabidopsis thaliana* is already characterized based on RNA-expression data under specific conditions (http://www.arabidopsis.org/). Nowadays, NGS provides an alternative technology for RNA sequencing [50,51]. However, this technology is still in development and for eukaryotic cells still very expensive because several fold genome coverage has to be measured to obtain statistically significant data. For further information about RNA-seq and epigenomics, reference is made to recent publications [52,53].

6.2. *A proteomics toolbox for green systems biology*

6.2.1. *Non-targeted versus targeted proteomics — MAPA versus Mass Western*

Genome sequencing and systems biology revolutionized life sciences. Proteomics emerged as a fundamental technique of this novel research area. In the following, a proteomic toolbox adapted to systems biology needs will be introduced. To capture the dynamic of a biological system, its components need to be identified and quantified. Proteomics is confronted

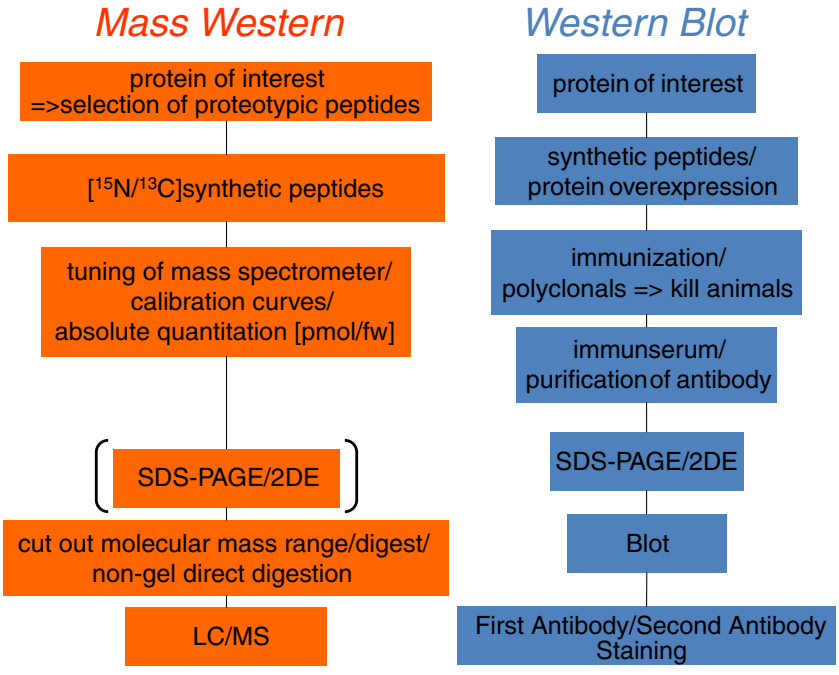

Fig. 6 – Mass Western strategy. Proteotypic peptides are initially selected from a proteome analysis. Those peptides serve as a model for synthetic, internal, stable isotope standards and are introduced for absolute quantification using a triple quadrupole mass analyzer [62]. The typical strategy of a Western blot using specific antibodies is shown as a comparison.

with a task that appears unsolvable. Both in plant proteomics and also in human proteomics, the following numbers are anticipated: assuming approx. 20,000–30,000 annotated genes of a genome, after consideration of splice variances and post-translational modifications such as phosphorylation or glycosylation, several hundreds or thousands of possible protein species per annotated gene may be reached. It is however not assumed that all possible protein isoforms are active at the same time. Nevertheless, existing technologies are confronted with enormous challenges due to the high number and dynamic concentration range of all proteins of a single steady state. Based on this consideration, in gaining a holistic overview of the dynamics of a continuously transient biological system it is important to analyze many different steady states, time series, diverse genotypes and their phenotypic plasticity. In modern biology, a simple comparison between states A versus B is no longer adequate to perform functional modeling. Instead, we need high sample throughput technologies that are able to identify as many proteins of the system as possible. Thus, there is a need for fundamental method development and improvement.

Nevertheless, there are clear strategies for increasing sample throughput as well as the number of protein detections. One strategy is based on the unbiased or untargeted identification and quantification of as many proteins as possible. In the following, this method is called "untargeted" protein analysis and is based on recent developments in proteomics technology called "shotgun proteomics". This is a high sample-throughput method in which complex protein samples from tissue or cells are directly cut into small peptides and subsequently analyzed via liquid chromatography coupled to mass spectrometry [54,55]. Peptides are identified via their fragment fingerprint against genomic/predicted proteomic databases [56]. Protein identification is then based on the reconstruction from these identified peptides and thus called shotgun proteomics. Compared to classical methods in proteomics, shotgun proteomics is characterized by a very high protein identification rate and thus qualified to establish huge qualitative and quantitative proteome catalogues for model organisms. Another strategy uses those proteome catalogues for the design of "proteotypic" peptides of a protein for a "targeted" analysis (see Fig. 4). These techniques will be introduced in the following paragraph.

6.2.2. Rapid proteomic phenotyping using MAPA (Mass Accuracy Precursor Alignment) and ProtMAX

Genomic databases and their corresponding computer-predicted proteomic databases became essential for proteome science [57]. They form the basis for protein identification of shotgun proteomics analyses. To enable confident protein

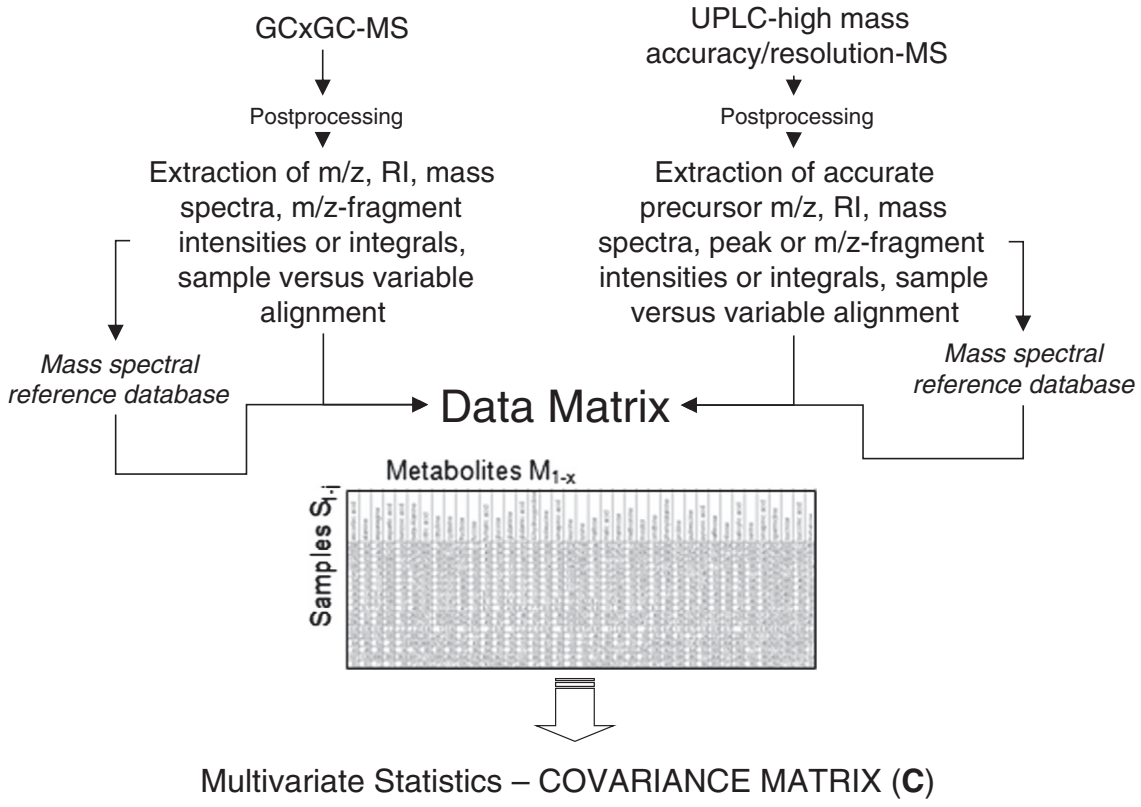

Fig. 7 – Metabolomic platform combining GC/MS and LC/MS techniques to cope with metabolomic complexity in biological systems [40]. GC/MS is one of the current "gold standards" with respect to comprehensiveness, sample throughput and identification rates [95]. Two-dimensional GCxGC–ToF-MS further increases the resolution of ultracomplex metabolome samples [80]. High mass accuracy/high resolution mass spectrometry emerges in parallel with high-resolution ultra-performance mass spectrometry (UPLC) and increases the detection capacities' orders of magnitude. Data can be combined in one data matrix to reveal the covariance matrix for the detection of physiological biomarkers, metabolite correlation networks and network topologies (see text for further details and [11,40,82,84,94].

identifications, ideally their complete sequence information needs to be available or otherwise a "de novo" interpretation of mass spectrometric raw data is necessary. The untargeted analysis can now be divided into two possible strategies, the database-dependent and independent interpretation [59]. The general goal is the identification of as many proteins as possible out of a complex protein mixture. With the database-dependent approach, only unambiguously identified sequences are considered. The database-independent MAPA approach is based upon an algorithm called PROTMAX [58] that groups all peptide precursor ions with the same mass to charge ratio (m/z) derived from mass spectrometric raw data in a data matrix without an initial database search (see Fig. 5). At the same time, the frequency of observed m/z fragments according to the concentration of peptides are counted (spectral count) [59,60] and added to the data matrix. This data matrix can be analyzed statistically to rank peptide precursor ions according to their phenotype-dependent impact (see Chapter 12.1 and Fig. 12). Interesting candidates are then identified via a database search or de novo interpretation. This also allows for the identification of rarely detectable forms of protein modifications and polymorphisms [58]. In summary, although the untargeted analysis only allows for relative quantification, it enables a more holistic overview of all detectable proteins within a system. Moreover, this strategy also delivers basic data for the targeted analysis — the Mass Western (see next chapter and [61]).

6.2.3. *Mass Western and ProMEX*

The best-known method for a targeted analysis of specific proteins out of a complex sample is based on antibodies (*e.g.* Western Blot, see Fig. 6). Besides time-consuming and extensive production of protein-specific antibodies, it is very difficult to distinguish between proteins of high homology [61]. Furthermore, absolute quantification of exact concentrations is not feasible. We and others developed a strategy based on mass spectrometry, which enables protein isoforms or whole pathways to be distinguished and absolutely quantified out of complex samples using stable isotope labeled synthetic peptides [25,62–67]. Due to its similarity with the Western Blot it is also called "Mass Western" (see Fig. 6) [61]. Furthermore, the Mass Western allows for a high sample throughput and the possibility of analyzing many proteins (~100) within a single analysis [25,62]. Proteome data measured in different cell states can be stored in a proteome database (ProMEX, http://promex.pph.univie.ac.at/promex/; [68]). They comprise proteotypic

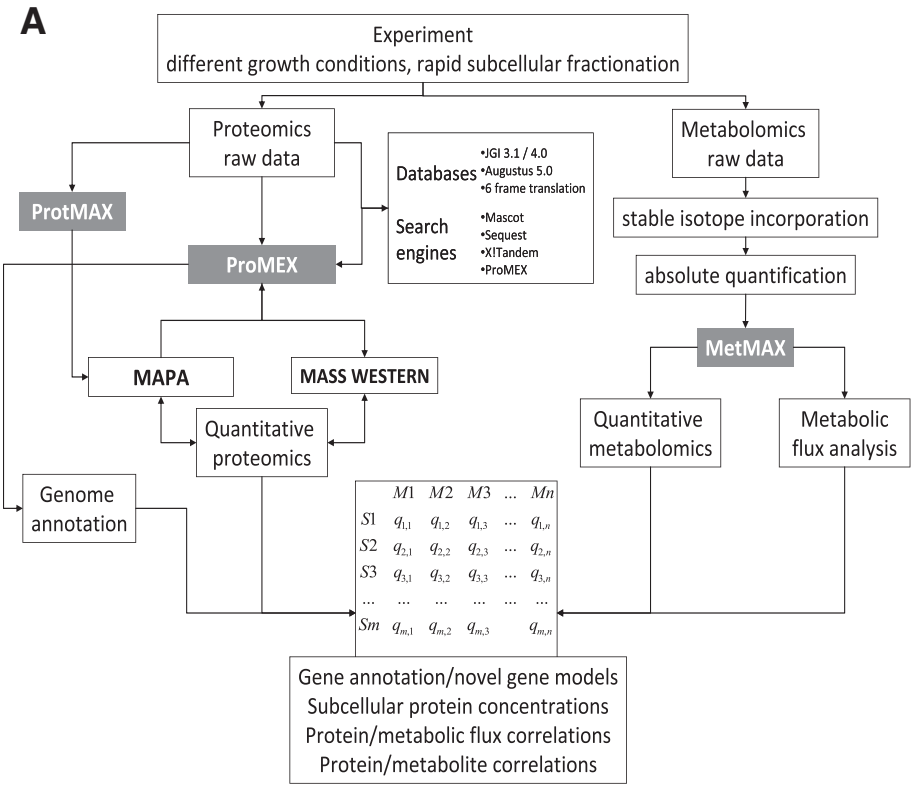

Fig. 8 – A. Analytical platform for the metaproteogenomics approach. For details of this platform see Wienkoop et al. [25]. B. Mapman visualization (http://mapman.mpimp-golm.mpg.de/) of detected proteins and metabolites of a systems biology-based *Chlamydomonas reinhardtii* study. Different metabolic pathways are shown such as the TCA cycle, Calvin cycle, amino acid synthesis and breakdown, etc. Blue boxes indicate all detected proteins. Red boxes indicate proteins that are targeted using the Mass Western strategy (see text). White dots correspond to experimentally identified metabolites. Not all predicted gene models are supported by proteome analyses. At the same time, proteomic data enable the suggestion of new gene models, not found via computational prediction from the raw genome sequence alone. All the data are stored in the proteomics database ProMEX (www.promexdb.org)[68].

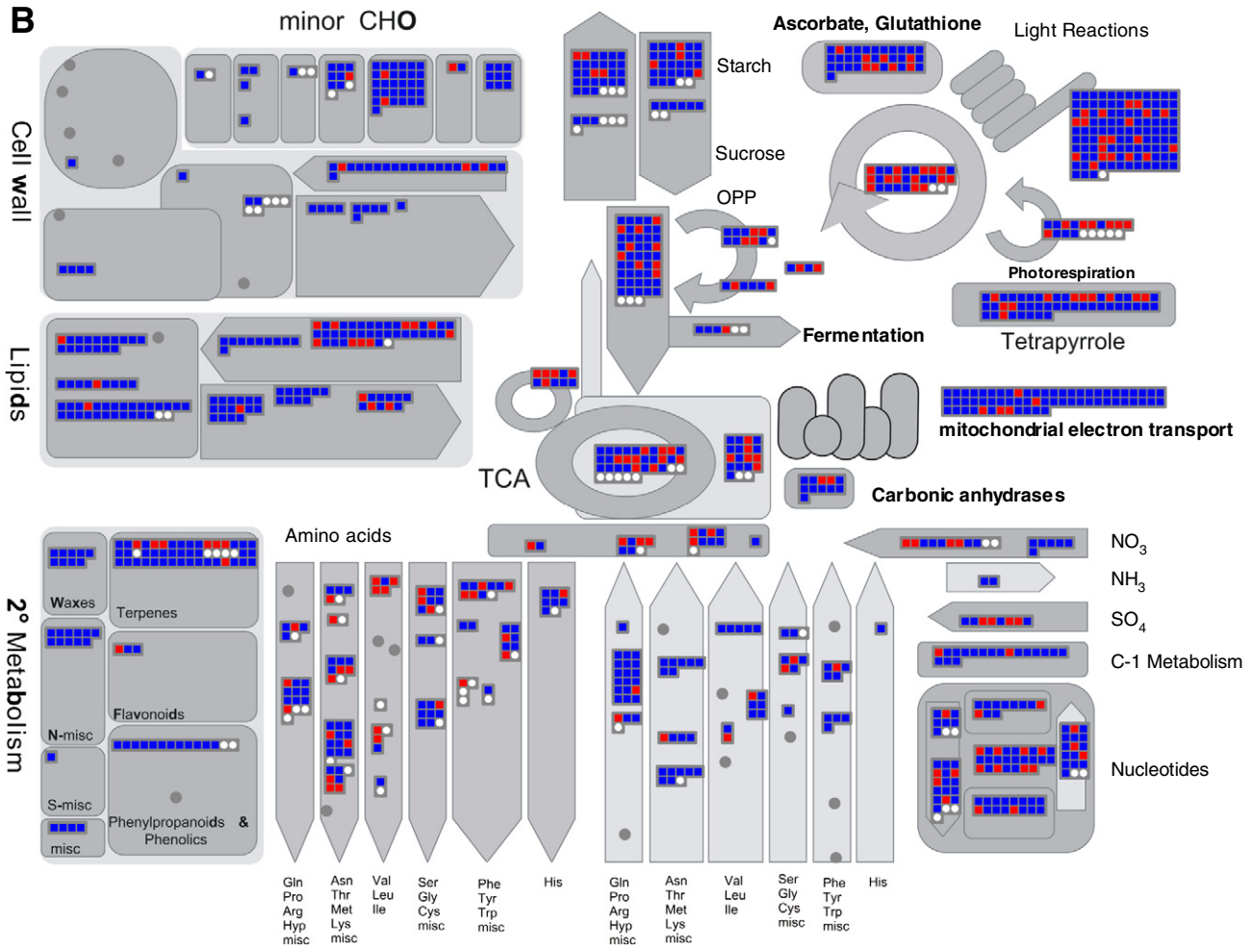

Fig. 8 (*continued*).

peptide libraries as a basis of the Mass Western method design (Fig. 6). It allows for the detection of entire metabolic pathways such as the Calvin Cycle, glycolysis, citric acid cycle etc. in parallel [25]. These data are important for the mathematical reconstruction and prediction of metabolic pathways and their regulatory mechanisms and hence are relevant for the investigation of the genotype-phenotype relationship (see Chapters 4 and 8).

6.3. *Metabolomics*

Bioanalytical methods in metabolomics science provide the most direct tools for the quantitative measurement of the metabolism in an organism. Overviews of available technologies can be found in recent reviews and books [11,69–71]. In view of the physico-chemical diversity of small biological molecules in a biological organism, the challenge remains to develop one or more protocols to gather the whole "metabolome". The general estimation of size and dynamic range of a species-specific metabolome is at a preliminary stage. In the plant kingdom the structural diversity is enormous and reveals new compounds on a daily basis. Estimates exceed 5 million putative structures. Many different techniques are needed for the analysis of all these different chemical structures. Therefore a combination of techniques has to be used and the data have to be combined [11,69]. Mass spectrometry is one of the technologies which has rapidly developed and also revolutionized the field. Different hyphenated technologies are presented in [2]. Each different technique provides other features. It can therefore be expected that by combining different technologies the coverage of a metabolome will be substantially increased. Metabolic fingerprinting techniques using NMR or IR spectroscopy, for instance, achieve a high sample throughput and a global view on *in vivo* dynamics of metabolic networks [69,72–76]. One of the gold standard techniques in terms of sample throughput, comprehensiveness and accuracy in metabolite identification is gas chromatography coupled to mass spectrometry [77].

A very recent development is the use of two-dimensional gas chromatography coupled to fast acquisition rate mass spectrometry (GCxGC–MS). The online coupling of two GC columns with different functionality, for instance a first long hydrophobic and a second short polar column, increases the separation efficiency of a complex metabolomic sample and improves spectral quality after deconvolution. However, the deconvolution process from such extended, two-dimensional raw chromatograms is very complicated. Moreover, metabolite identification and data alignment acts as a bottleneck. Recently,

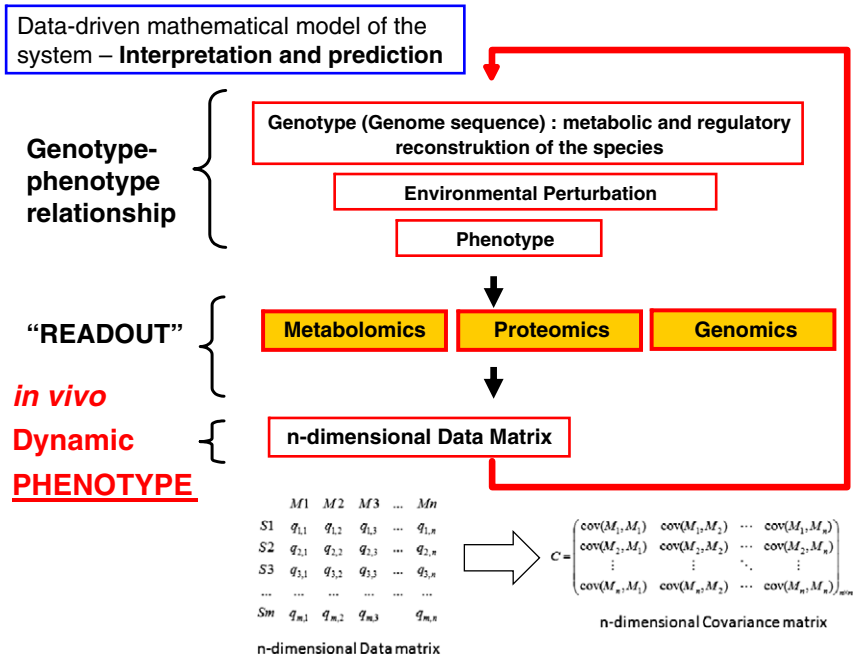

Fig. 9 – A framework for the systematic investigation of the genotype–phenotype relationship and integration with genomic reconstruction and modeling approaches. Data are integrated into a data matrix which can be analyzed with multivariate statistics. The result – the covariance matrix – is closely bound to the mathematical model of the system (see text, Figs. 10 and 11 and [2]).

we presented a complete strategy to perform a convenient data extraction and alignment using GCxGC–MS technology [78]. The introduction of a second retention index which can be used to increase the confidence in metabolite identification is especially important [80]. One of the most promising platforms for metabolomics is the combination of GC–MS and LC–MS (see Fig. 7) [39]. Due to their specific technology, both technologies provide a complementary view of the metabolome [39] — central metabolites such as amino acids, sugars, organic acids, free fatty acids, etc. by GC–MS, higher molecular masses, *e.g.* secondary metabolites, co-factors and/or sugar-phosphates profiles by LC–MS [2,39]. In Fig. 7 such a platform is shown, combining GCxGC/MS and LC/MS for metabolome analysis. However, the reader should be aware that most of the metabolomics platforms still need further method validation and daily quality checks. This is an essential requirement to guarantee meaningful biological applications. Furthermore, databases, experimental standards and data exchangeability between labs is an urgent issue for further developments in metabolomics [2,79].

7. Data mining and metaproteogenomics

7.1. *Metabolomics, proteomics, transcriptomics and sample pattern recognition in systems biology*

Systems biology aims to attain a holistic overview of all regulatory processes and reactions (phenotypic plasticity) of a biological system in response to environmental perturbations. The resolution of these processes improves with the amount of data available. Consequently, the integration of protein, metabolite and transcript data enhances the resolution [80–82]. A workflow for this data integration approach was recently proposed ([83], see also Chapter 8). In particular, with untargeted protein analysis but also by integrating Mass Western analyses (see Figs. 4 and 6), huge amounts of data are generated. Statistical and bioinformatic methods are therefore necessary for comprehensive data mining and the extraction of biologically relevant information. One of the most important methods for data mining and data visualization is a pattern recognition strategy based on supervised and unsupervised multivariate statistics, *e.g.* principal components analysis (PCA) or independent components analysis (ICA) (see Chapters 8 and 12.1 and Figs. 9, 10 and 11) [84,85]. By means of pattern recognition, conclusions about biologically active regulatory processes and proteins can be drawn (see following chapters 8, 11 and 12) [83].

7.2. *Genome annotation: shotgun proteomics complements shotgun genomics*

The number of novel sequenced genomes and new genome projects – both prokaryotic, eukaryotic as well as the "metagenome" of communities of organisms [86,87] – is exponentially increasing. The high amount of data generated displays the enormous challenge for bioinformatics. A classical approach is the solely computer-assisted annotation of predicted Open Reading Frames (ORF). Newly developed

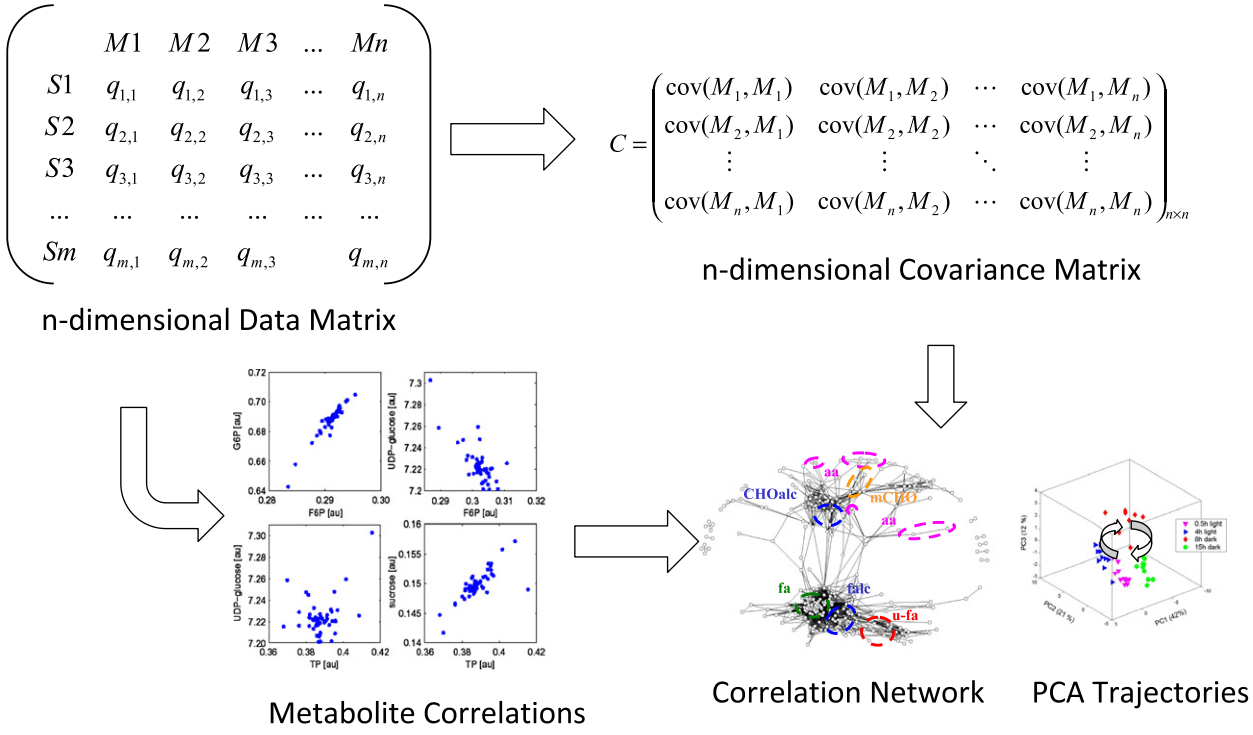

Fig. 10 – The direct linkage of the n-dimensional data matrix, the resulting covariance matrix as a result of multivariate statistics and correlation networks as well as principal components analysis and the resulting trajectories (for further details see [2]).

techniques consider experimental data such as transcriptomic, proteomic but also metabolomic data for improved genome annotation [24]. A high-throughput method for a high protein identification rate is shotgun proteomics described in Chapter 6.2. Similar to the shotgun genomics technology, proteins can be reconstructed from protein fragments: e.g. tryptic peptides. Shotgun proteomics is characterized by a very high protein identification rate and generates huge proteome catalogues for model organisms [22,26,88–90]. This way, predicted gene models can be confirmed by proteomic data. Furthermore, many proteomic data not found in the predicted gene models may point to new gene models that would not have been found using computer based in silico-analyses only [22,24,58,88].

7.3. *Metaproteogenomics*

Recently, we proposed a metaproteogenomics strategy for data integration and combination with genome reconstruction [24,25]. Fig. 8A shows the bioanalytical platform for this approach (for details of this approach see [25]). Fig. 8B shows a projection of metabolomic and proteomic data of all identified proteins within one proteome study into a functional genome annotation and subsequent metabolic reconstruction of the unicellular green algae *Chlamydomonas reinhardtii*, a recently sequenced model organism [91] for photosynthesis and CO_2-neutral biomass production also called the "green yeast" (see also Chapter 12.2). Fig. 8B represents the first metabolic draft of a genomic reconstruction of *Chlamydomonas reinhardtii*. Based on this projection, several Open Reading Frames (ORF) were identified with novel annotations and new pathways.

8. A combined bioanalytical platform for the measurement and modeling of the genotype–phenotype relationship

The combined analysis of genetic variation – the genotype – and the corresponding molecular phenotype and its physiology is one of the most pressing challenges in the next decades. Raw genome sequences will be present for almost any plant species of interest. However, recent molecular analysis combining metabolomics, metabolic flux, targeted and non-targeted analysis demonstrates the difficult nature of the dynamic phenotype — it is only predictable when based on the static genome sequence [25]. A framework for the systematic investigation of this dynamic interaction of a known genotype and the environmentally-controlled phenotype (Fig. 9) can be introduced here. Integration of genomic, proteomic and metabolomic data into data matrices will reveal correlations and covariance, respectively, between the molecular constituents [2,11,83] (see Fig. 9 and 10). This covariance matrix is a central component of the data integration and interpretation strategy [82]. At the same time, metabolic reconstruction and the mathematical description of the system are obtained with coupled ordinary differential equations (ODEs). This system of ODEs is directly connected to the data matrix and

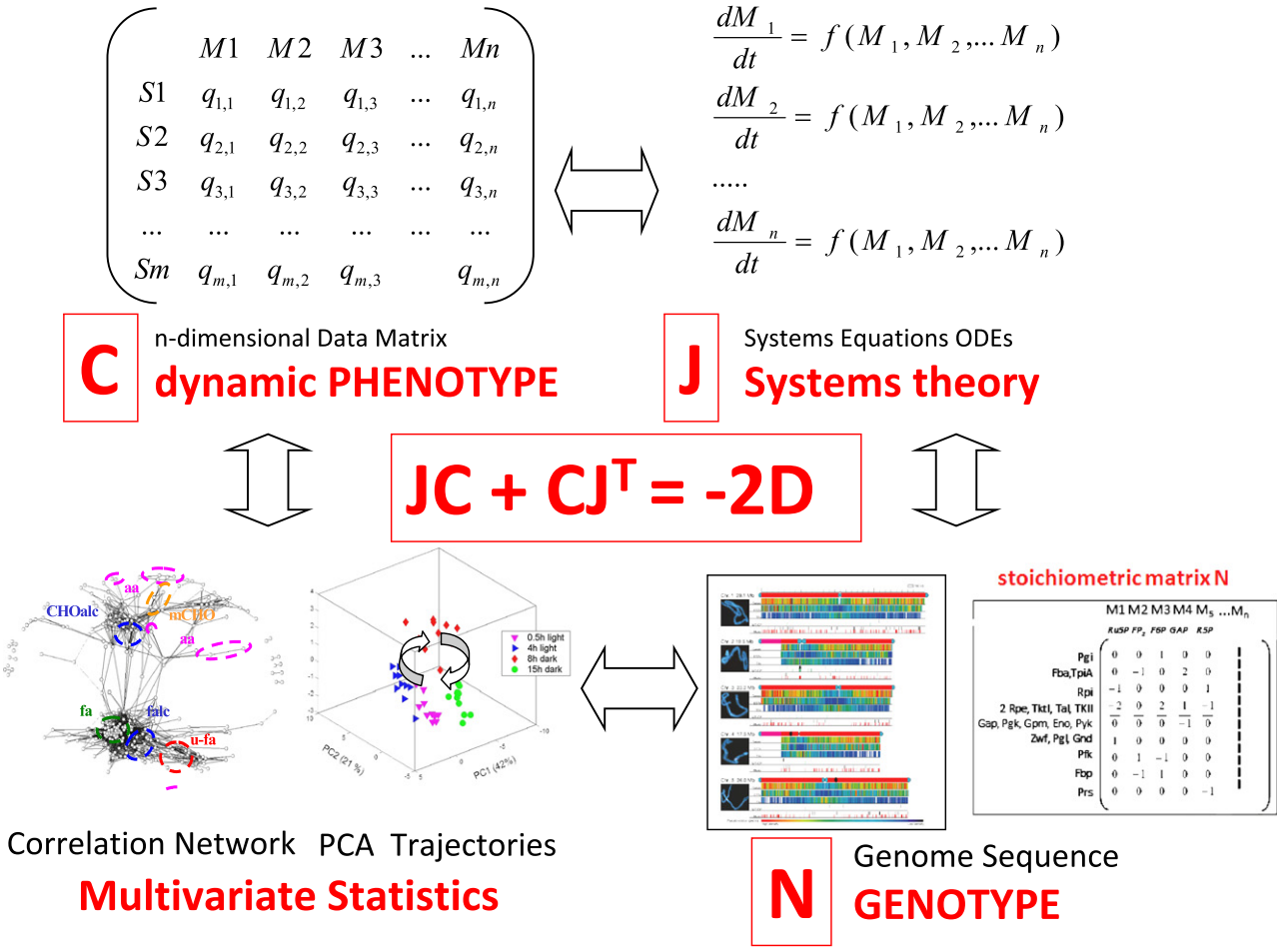

Fig. 11 – A genotype–phenotype equation ($JC + CJ^T = -2D$) links the genotype characterized by the stoichiometric matrix N and the systems equations (ODEs) resulting in the Jacobian J. The phenotype is characterized by the data matrix and the resulting covariance matrix C as a result of multivariate statistics. The equation also contains a diffusion matrix D by assuming stochastic fluctuations in metabolic networks. For detailed explanation see [2].

the resulting covariance matrix (Fig. 11). This relationship from the genotype and the dynamic phenotype can be described with a generic equation (Fig. 11) (for further details see [2]).

We have used this workflow for several studies [2,25,80,82,83,92,93] and in future work will explore the genotype–phenotype equation in more detail.

9. International public activities

The integration of genomic, proteomic, metabolomic, environmental as well as morphological and anatomical data is by nature too complex to be achieved by single laboratories. In recent years several international collaborations and initiatives have been founded that support the open source consolidation of techniques, data, databases, functional interpretation of plant genomes and other activities. One of the largest communities is the Multinational Arabidopsis Steering Committee (MASC) which is a later branch of the highly successful North American Arabidopsis Steering Committee (NAASC). In recent years MASC has co-founded several subcommittees (see http://www.arabidopsis.org/portals/masc/Subcommittees.jsp) including Bioinformatics, ORFeomics, Metabolomics, Natural variation, Phenomics, Proteomics and Systems biology. This diversity of combined activities for a single plant model system is unique and will accelerate the achievements and collaborations of international laboratories [94–96]. An important instrument was developed in this consortium — the gene function tracking thermometer (http://www.arabidopsis.org/portals/masc/2009_MASC_Report.pdf) which reveals how many *Arabidopsis thaliana* genes are functionally characterized or under investigation using a variety of methods such as RNA expression levels analyzed with microarrays (more than 26,893 genes analyzed), insertion mutants, RNAi constructs and others.

Recently, the MASCP (proteomics subcommittee; http://www.masc-proteomics.org/mascp/index.php/Main_Page) established the GATOR portal. This web portal allows the

Arabidopsis researcher to search for any AGI code or lists of AGI codes in all proteomic databases assembled by MASCP [95]. Similar activities are planned within the Metabolomics subcommittee (MASCM; http://www.masc-metabolomics.org).

Very recently, the importance of integrating knowledge of plant proteomics and especially the translation into practical applications was recognized by a group of researchers. The result is an international initiative called the International Plant Proteomics Organization (INPPO; http://www.inppo.com/) with the goal of consolidating techniques, data and databases and functional proteomics analysis for various plant species, especially crop plants, and finally address global agricultural problems as discussed in this review [97]. Members of MASCP are on board and will provide their knowledge of a single model system – *Arabidopsis thaliana* [94] – in this important activity. INPPO is an excellent example of a non-profit, open-source initiative. Interestingly, it is merely based on the initiative of scientists without any funding, comparable to the early stages of MASC and the corresponding subcommittees or other communities. In later stages these organizations are able to find funding because of the acknowledged importance of these integrative open-source initiatives. This might be a very interesting strategy for the public sector to initiate important areas of research which are not well funded yet or where it is necessary to produce political interest.

10. Knowledge transfer from model to applied systems: translational biology

As discussed above, biological research, especially molecular biology, is experiencing dramatic improvements. The throughput and combination of NGS, genomics, proteomics and metabolomics technologies will dramatically improve our current view of the molecular principles behind living systems. For the last 10 years or more, this kind of research was possible for selected model organisms such as *Arabidopsis thaliana*, yeast and others. This will also change and maybe comparative genomics will be able to predefine functional genome annotation. However, the functional annotation of all *Arabidopsis* genes is still in progress. This contrasts with expectations at the beginning of 2000 and the hope that the full genome sequence will enable a complete understanding of the principles of life. The opposite is true; with respect to the complexity of molecular interaction networks it becomes clear that almost all goals postulated for the recent decade will not be achieved [9], which is of course not a failure of the researchers in these areas but a general epochal underestimation of the complexity of the molecular phenotype [2].

The last decade has taught us that a strong effort in translational biology is necessary. All gene functions learned from model systems need to be systematically compared with other organisms [98]. At the moment it is unclear how well this

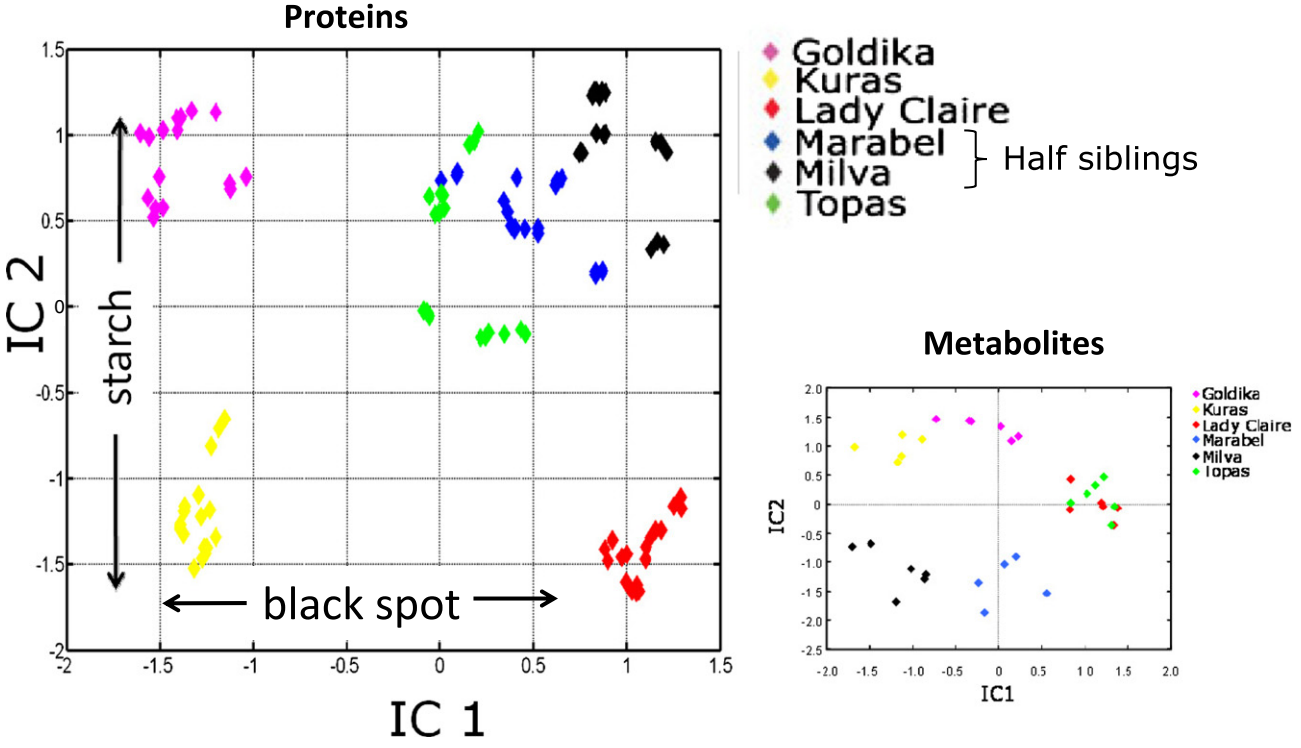

Fig. 12 – Sample pattern recognition for a MAPA protein data matrix of six potato cultivars. The clear separation of genotypes is visible. The separation in the principal component analysis is due to the differences in protein abundance. Specific protein markers are assigned to the corresponding potato cultivars. In the smaller plot on the right side, an ICA plot of metabolite data is shown. It is remarkable that the pattern is similar for some cultivars but also shows some dissimilarities to the protein data. These differences are discussed in the later Chapter 12.1.

kind of knowledge can be translated into other systems, or the extent to which different plant families such as Arabidopsis and spinach or soybean are similar. However, there is sufficient knowledge already available to initiate this comparative approach (http://www.plantcyc.org/).

The following sections include brief examples of how the platforms discussed above are applied.

11. Applications in ecology and ecosystems

The application of NGS, RNA-seq, microarrays, proteomics and metabolomics is still in its infancy in ecological research. This might be due to the complexity of integrating molecular data with ecological research, however, the next logical step is to combine traditional ecological questions with systems biology technologies. In metagenomics, NGS is applied in rather accessible systems such as microbial communities [99]. Environmental metabolomics, *e.g.*, is applied in toxicity testing in ecological risk assessment [100]. In plant ecology and plant communities, however, these technologies are at a very early stage.

Ecosystems dynamics are driven by the diversity of their species. Natural variation is a key principle for diversity. Thus, the molecular investigation of the genotype–phenotype relationship will give an insight into the dynamic behavior of ecosystems. Here, model systems such as *Arabidopsis thaliana* (L.) Heynh. play major roles. In the early 20th century, Friedrich Laibach already emphasized the role of *Arabidopsis* as a model plant for genetic and developmental studies [101,102]. He was especially interested in the large variation of phenotypes and physiological traits such as flowering time and seed dormancy of different *Arabidopsis* ecotypes [8]. Thus, he was a pioneer in the investigation of natural variation. With respect to the developments of modern research in plant physiology and ecophysiology (see Chapter 5, natural variation) it is remarkable that there is so much truth in his prophecy and that the consolidated efforts of a small *Arabidopsis* researcher community are so important [8]. A group of about 25 people in 1976 provided the nucleus for the approx. 25,000 *Arabidopsis* researchers worldwide today and for developing *Arabidopsis thaliana* into one of the most important plant model species [8] (see also Chapter 9).

Together with the predictions of Friedrich Laibach, Arabidopsis is an ideal model system for the investigation of ecotypes and natural variation, which is exemplified by a plethora of publications in this area [3,43,48,103,104].

In a recent study by Chevalier et al. eight *Arabidopsis* ecotypes were analyzed with two-dimensional gel electrophoresis [105]. It was possible to classify the ecotypes by the identification of different protein markers and functional differences between the ecotypes.

In a pioneering study Keurentjes et al. investigated the effects of different *Arabidopsis thaliana* accessions on metabolism, represented by untargeted profiling of methanolic extracts with LC/MS technology. Several correlations of proposed metabolites and QTLs, especially in the glucosinolate pathway, a major pathway in brassicacea, were detected in accordance with former studies [44]. In a more recent study, Chan et al. used a similar approach to correlate metabolite profiles stemming from GC/MS analysis with SNP-array data of different *A. thaliana* accessions from former studies [48]. Interestingly, in both studies Keurentjes et al. and Chan et al. have worked on different fractions of the metabolome as LC/MS data rather reveal secondary metabolites and GC/MS data provide metabolites of the central metabolism such as sugars, amino acids, organic acids etc. (see Chapter 6.3. Metabolomics and for more details [2,39]. Therefore it can be expected and was already demonstrated that the combination of GC/MS and LC/MS platforms will reveal a much better picture [2,39,70].

All these studies have dealt with lab-based experiments to provide as controlled conditions as possible. The next challenging demand is to analyze the plant in its natural environment. Classical biodiversity studies investigate grassland communities with different compositions of grasses, herbs and legumes [106,107]. The JENA experiment is the largest long-term European project of this kind (http://www.the-jena-experiment.de/; [108]). In conjunction with this Jena experiment, our lab performed metabolite profiling of different plant species in their quasi-natural environment and investigated their phenotypic plasticity and the effects of biodiversity on their metabolism. We combined GC/MS and LC/MS approaches to cover a large fraction of the metabolome, central metabolism and secondary metabolism as discussed above. A pronounced diversity gradient was observable in the metabolite profiles as well as individual responses of the different plant species. These different responses can be summarized as metabolic signatures that are characteristic for each individual plant species [39].

Proteomic analyses tend to be missing in plant ecological studies in the natural environment. Recently, we analyzed 12 different potato cultivars growing in fields (see below) [58]. For this approach we had to implement new methods to cope with high sample numbers and a reasonable workflow for data mining and developed the MAPA approach (see Chapter 6.2.2). Shotgun proteomics combined with rapid data mining strategies has the potential to complement transcriptomics and metabolomics data with respect to sample throughput [83]. However, proteome coverage, detection of posttranslational regulation and sensitivity against low abundant proteins are still limited in almost all proteomic approaches (see also Chapter 6.2).

12. Applications in biotechnology

12.1. *Marker-assisted selection (MAS)*

All these technological platforms described above enable the genome-wide molecular analysis of different genotypes. This integrated high throughput analysis of metabolites, proteins and transcripts allows the definition of biochemical phenotypes and their relationship to the corresponding genotype and to environmental conditions [83]. The integration of metabolite and protein profiling has already been demonstrated to significantly improve pattern recognition and the selection and interpretation of multiple physiological and biomarkers for plant systems and different plant genotypes under different environmental conditions such as day–night

rhythms or cold stress [80,82]. Integration of metabolite and transcript data was also demonstrated to reveal the relationship between mRNA expression and dynamics of secondary metabolism [81]. The exploitation of these technologies in plant biotechnology and QTL-based marker-assisted breeding approaches [109–113] is an obvious development.

Most of the studies are focused on DNA markers. In recent studies the successful application of these technologies was also demonstrated for proteomics and metabolomics. De Vienne and colleagues introduced for the first time the terminus Proteome Quantity Loci (PQL) and systematically combined proteome analysis with genetics and QTL mapping in maize [111,112]. Schauer and colleagues demonstrated the application of metabolomics for the characterization of interspecific introgression lines (ILs) of tomatoes and correlated these data to QTLs related to yield [114].

In a recent project for potato breeding funded by the German Federal Ministry of Education and Science (BMBF; http://www.bmbf.de/) the integration of different molecular levels was a major aim. Potato breeding is complicated by heterozygous and autotetraploid genetics. Typically nine or more years of selection work are needed to define successful candidates for official trials with the federal variety authorities. Marker-assisted selection could accelerate this process for the identification of useful traits in the early years of the selection process. It is anticipated that new technologies such as genomics, proteomics and metabolomics will yield such marker systems, however, these technologies have hardly reached the stage of application for breeding in the private sector. In a potato breeding pool, a multitude of trait alleles of various origins are present. A similar multitude of diagnostic marker assays will therefore be required for marker-assisted selection [115]. In the context of this study we have developed a procedure for shotgun proteomics in combination with novel data mining algorithms called MAPA (mass accuracy precursor alignment, see also Chapter 6.2.2) which enables a high throughput strategy for forward screening of potential protein markers in this complex system [58]. The principle of the method is shown in Fig. 5. In Fig. 12 the ICA plot for six different commercial cultivars measured at different places in Germany in the corresponding fields is shown. The cultivars are easily distinguished. Based on this sample pattern, protein markers can be assigned to the different traits of each potato cultivar. These proteins are potential markers for the development of a MAS strategy [58]. However, the robustness of these markers must be analyzed with higher statistical power from a higher number of samples. Thus, this kind of high sample throughput capacity of the MAPA strategy has great importance for the future.

In a recent study, we have shown that MAPA is capable of distinguishing many different isoforms of protein families and assigning their differential abundance to specific cultivars [90]. We speculate that this process is related to different developmental properties of the different cultivars. Thus, these proteins are potential physiological markers — this is work in progress. Altogether, with all the analyses we have assembled the largest tuber proteome catalogues available [90] and all the proteins are stored in PROMEX (see Chapter 6.2.3) [69].

There is another remarkable feature in these HTP molecular data. Fig. 12 shows the ICA plot of metabolites [58]. Both data sets – the metabolomics data and the proteomics data – show a good cultivar discrimination, however, the sample pattern can be interpreted differently depending on the properties of the different cultivars. Thus, the metabolite data carry different information to the protein data. For instance, in the metabolite data we observed a pronounced dynamic of sugar metabolism in potato tuber tissue, as expected. In contrast, the protein data are more characteristic for developmental processes in the potato tuber tissue. Integration of these data leads consequently to optimized pattern recognition processes and improved interpretation of the molecular data with respect to the molecular phenotype which was indeed observed in several previous studies [80,82]. It is proposed by us that we are able to reveal synergetic effects in the molecular data. So far, all our integrative analyses point to these properties [83]. The basis for this assumption is the genotype–phenotype equation presented in Chapter 8 and in a recent review [2]. This equation directly connects the covariance structure of the data – in other words the pattern recognition using multivariate statistics – with the underlying genotype.

12.2. *Biofuels*

Biofuels have the capacity to substitute fossil fuels and to normalize the global natural balance of CO_2 consumption and emission on earth. Biofuels are produced from renewable, CO_2-neutral resources such as plants, photoautotrophic microbes or algae. [116–118]. Misconceptions of biofuel production have shown, however, that the transition from fossil fuel consumption to renewable energy resources is a long and painstaking process and comprises all aspects from scientific to socio-economic complications [119]. Therefore it is of the utmost importance that the transition from first-generation biofuels (corn, sugar cane, rape seed etc.) to second-generation (lignocelluloses-based production of bioethanol from *Miscanthus* and trees, poplar etc.) to third-generation biofuels (photoautotrophic microbes, microalgae) is addressed as soon as possible with profound support from

Fig. 13 – Poplar shoots in *in vitro* culture. Left without endophytes, right with endophytes. The shoots have been grown for two weeks in a nitrogen-free medium.

international funding agencies. These problems are indeed as pressing as biomedical applications of systems biology.

Due to the developments discussed above, green systems biology will contribute at all levels to the useful applications of biofuel production. Applied to energy crops such as grasses and trees and photoautotrophic microbes as well as microalgae, biofuel production can be investigated and enhanced at all levels from modern breeding approaches up to genetic engineering solving biomass recalcitrance [120–123].

Chlamydomonas reinhardtii is one of the most accessible model systems for photoautotrophic growth and hydrogen and lipid production. The unicellular green algae was recently sequenced and in many laboratories now serves as the model of choice for physiological, ecophysiological and economical investigations of biofuel production [91,124].

We recently set up a comprehensive analytical platform to investigate growth and lipid production of *Chlamydomonas* [25]. This platform comprises the described targeted and non-targeted proteomics analyses of Chapter 6.2.1, the metabolomics platform described in Chapter 6.3 and metabolic flux analysis using metabolic labeling with subsequent GC/MS analysis [25,78]. In Chapter 7.3 the combination of these technologies is exemplified.

Based on recent growth experiments and the application of the analytical platform, remarkable plasticity of the metabolism of *Chlamydomonas* was observed [25]. Especially pronounced are effects of the carbon concentrating mechanism (CCM). Here, up to 12 isoform of carbonic anhydrases are postulated from the genome. We were able to measure 5 isoforms with the Mass Western approach (Chapter 6.2.3) and revealed enormous differences in the concentration patterns (attomol/1000 cells) of these isoforms. A mitochondrial isoform (CAH4) showed a very high dynamic range and is very active under CO_2-limiting conditions [25].

These observations coincide with recent studies in *Chlamydomonas* and *Arabidopsis* and point to a significant role of carbonic anhydrases in CO_2-sensing pathways in higher plants as well as in microalgae [125–127]. We will investigate these processes in more detail in future to link processes of lipid production to CCM.

Populus trichocarpa is a tree model system for energy crops [120]. Poplar belongs to the family of Salicaceae and is a fast-growing tree. As an energy crop it is used worldwide in short rotation farming [128] (http://www.probstdorfer.at/index.php?url=energieholz.htm).

Recently, we investigated the growth-promoting effects of an endophyte on poplar cuttings in *in vitro* culture [129]. This endophyte – *Paenibacillus* sp. – stimulates root formation in poplar cuttings [130].

Inoculated plants showed dramatically changed metabolite profiles, indicating that the interaction of the endophyte with the plant indeed alters the physiology of the plant [129]. Especially, the nitrogen metabolism is influenced which indicates better uptake of nitrate from the medium or other effects which are initiated by the endophyte-plant interaction. In Fig. 13 endophyte-free and inoculated plants are shown growing on nitrogen-free medium. We have observed better survival statistics with the inoculated plants under these conditions.

These processes might have significant effects on short rotation farming. Here, short poplar cuttings from different high-yield poplar clones are simply planted into soil. Root formation of the cuttings is of course a decisive step with respect to growth in this short rotation farming but also uptake of nutrition. We will investigate the underlying principles of the intimate plant-endophyte interaction in more detail in the future.

These are only two examples of a rapidly growing research field of CO_2-neutral biofuel production. Besides cost effectiveness, core research questions for all these developments should include the rapid substitution of nutritional crops (rapeseed, sugar cane, sugar beet, maize, wheat etc.) used for biofuel production by energy crops (grasses, trees, others) and algae which are already naturally designed for highly efficient biomass production as well as highly efficient CO_2 fixation. This process, which is as pressing as any question in biomedical research — or may be even more relevant for our future life on earth, will address the following socio-economic problems:

(i) Food market price is directly influenced by using food crops for biofuel production, especially a problem for developing countries
(ii) Global climate changes, use of algae for biofuels would diminish the competition for arable land, competition for arable land otherwise leads to a rapid deforestation and soil erosion processes
(iii) "Land grabbing" in developing countries by industrialized countries, irrational agricultural use of arable land [131]
(iv) Financial market situation, after IMMO bubble now the next "Land Grab and Food Market" bubble.
(v) healthy balance between public and private sector in AGBIOTECH [1]

13. Natural variation, biodiversity banks and plant breeding — conservation is the key

In the early times of the green revolution there was only one single trait: yield [132]. Nowadays, the potential for selection of specific traits has dramatically changed. Due to genome-scale molecular analysis and elucidation of gene function, in the future a modern breeder will be able to select a plethora of single traits or their combinations. Natural variation provides the richest source of traits such as disease resistances, insect resistances, drought tolerance, natural compounds, nutritional quality etc. This natural richness and diversity needs to be protected because in the continuing global environment of efficient monocultural plant production, we will only be able to cope with any imbalance in the future if we preserve natural genetic variation. The logical step is conservation of biodiversity to protect our environment and translate these processes into agricultural biotechnology.

13.1. *CIMMYT, INRRI and other public plant breeding institutions*

The CIMMYT (Centro Internacional de Mejoramineto de Maiz y Trigo) for maize and wheat breeding was founded by Norman Borlaug in the early 40s as a joint program of the US Rockefeller

Foundation and the Mexican government. This institute is a truly international scientific and public effort and one of the most outstanding worldwide centers of crop improvement. Here, the green revolution started with rigorous and systematic plant breeding programs for wheat and other crops [1,133]. Another important institute – the International Rice Research Institute (IRR) – was founded in the 60s also by the Ford and Rockefeller Foundation and the government of the Philippines and focuses exclusively on the improvement of rice varieties [1]. These institutions and many other public institutions worldwide have systematically established "biodiversity banks" consisting of native and improved food crop varieties. For instance, the CIMMYT in Mexico manages the most diverse maize and wheat collections: 140,000 unique samples of Triticeae seeds from more than 100 countries and 28,000 samples of seeds in the maize bank (http://www.cimmyt.org/). Each variety conserved in these biodiversity banks has a slightly different genetic makeup, consisting of different combinations of gene variants which provide the building blocks for breeding new and improved cultivars. The bank collection has been used on many occasions to obtain genes for resistance to diseases and pests of both crops, as well as for other traits of value.

The institute states (http://www.cimmyt.org/):

> "The collections are being conserved for the long-term benefit of humanity, free from any intellectual property restrictions. CIMMYT observes the terms of the International Treaty on Plant Genetic Resources for Food and Agriculture, signed by more than 100 countries since 2004. Each year the Center ships several tons of seed, in the form of small packets of samples of more than 5000 genotypes, in response to requests from over 100 researchers in dozens of countries worldwide."

Many other public institutions worldwide are taking care of seed banks and biodiversity banks (not only food crops) (for overview see Consultative Group on International Agriculture Research (CIGIAR); http://www.cgiar.org/ and [1].

These public efforts may provide one of the most natural "treasures" mankind has established so far.

13.2. *MAS and GAB: marker-assisted selection and genomic-assisted breeding and prediction*

These resources of biodiversity provide the plant geneticist and breeder with a compelling genetic variety. In combination with the developments of green systems biology such as NGS, RNA-seq, proteomics and metabolomics, the basic investigation of these natural variations will reveal complete novel workflows for the breeder. Traditional plant breeding programs rely mainly on phenotypes being evaluated in several environments; selection and recombination are based solely on the resulting data. Marker-assisted selection (MAS) uses molecular markers in linkage disequilibrium (LD) with QTL. Genomic selection (GS) is a new approach for improving quantitative traits in large plant breeding populations that uses whole-genome molecular markers (high density markers and high-throughput genotyping). Genomic prediction combines marker data with phenotypic and pedigree data in an attempt to increase the accuracy of the prediction of breeding and genotypic values.

New modern breeding tools are now available for successful integration of such polygenic traits during the breeding process in order to select better varieties. The newly identified varieties may be more sustainable and much easier to handle in seed and trade systems.

It is foreseeable that all these opportunities will revolutionize plant breeding.

14. Conclusion

Summarizing the broad view presented here, it is remarkable that the traditional and modern approaches in plant physiology, systems biology and plant biotechnology and breeding are so intimately linked. It is easily arguable from the facts above that the preservation of biodiversity is of the utmost importance for our future agricultural and socio-economical approaches, improving plant productivity and diversity for world feeding and renewable energy resources. Biodiversity is natural genetic variation — a research field which forced scientists like Gregor Mendel or Charles Darwin to reveal the secrets of inheritance, ecology and evolution. Even with today's completely new technologies such as NGS, RNA-seq, epigenetics, metabolomics and shotgun proteomics, the old questions remain unanswered and are at the same time the key to almost all applications in plant biotechnology:

How does the genotype determine the environmentally triggered phenotype?

Any functional studies on the exponentially growing volume of genome sequences of plant species or any other species will finally merge into this question.

Green systems biology provides the means to investigate this genotype–phenotype relationship for the first time in a fundamental way, combining genome-scale molecular measurements, phenotyping and computer-assisted modeling approaches.

Acknowledgements

I thank Stefanie Wienkoop and Wolfgang Hoehenwarter for many years of fruitful collaborations. I thank Anke Bellaire for many helpful discussions and useful comments. I thank the University of Vienna and the Faculty of Life Sciences for their great support.

REFERENCES

[1] Murphy DJ. Plant Breeding and Biotechnology. Cambridge: Cambridge University Press; 2007.
[2] Weckwerth W. Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. Anal Bioanal Chem 2011;400:1967–78.
[3] Platt A, Horton M, Huang YS, Li Y, et al. The scale of population structure in *Arabidopsis thaliana*. PLoS Genet 2010;6:e1000843.
[4] Somerville C, Dangl L. Genomics — Plant biology in 2010. Science 2000;290:2077–8.
[5] Last RL, Jones AD, Shachar-Hill Y. Towards the plant metabolome and beyond. Nat Rev Mol Cell Biol 2007;8:167–74.

[6] Meyerowitz EM, Bowman JL, Brockman LL, Drews GN, et al. A genetic and molecular model for flower development in *Arabidopsis thaliana*. Dev Suppl 1991;1:157–67.
[7] Meyerowitz EM. Arabidopsis, a useful weed. Cell 1989;56: 263–9.
[8] Somerville C, Koornneef M. A fortunate choice: the history of Arabidopsis as a model plant. Nat Rev Genet 2002;3:883–9.
[9] Somerville C, Dangl. Genomics. Plant biology in 2010. Science 2000;290:2077–8.
[10] Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet 2010;11:31–46.
[11] Weckwerth W. Metabolomics in systems biology. Annu Rev Plant Biol 2003;54:669–89.
[12] Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems biology. Annu Rev Genomics Hum Genet 2001;2: 343–72.
[13] 454. Roche 454 GSFLX. http://www.454.com/.
[14] SOLiD. http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/index.htm.
[15] Illumina. http://www.illumina.com/.
[16] Helicos. http://www.helicosbio.com/.
[17] Nagarajan N, Pop M. Sequencing and genome assembly using next-generation technologies. Methods Mol Biol 2010;673:1–17.
[18] Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods 2011;8:61–5.
[19] Cantacessi C, Jex AR, Hall RS, Young ND, et al. A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. Nucleic Acids Res 2010;38:e171.
[20] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 2005;33:W465–7.
[21] Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet 2010;11:476–86.
[22] Castellana NE, Payne SH, Shen ZX, Stanke M, et al. Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 2008;105:21034–8.
[23] Baerenfaller K, Grossmann J, Grobei MA, Hull R, et al. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 2008;320:938–41.
[24] May P, Wienkoop S, Kempa S, Usadel B, et al. Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. Genetics 2008;179:157–66.
[25] Wienkoop S, Weiss J, May P, Kempa S, et al. Targeted proteomics for *Chlamydomonas reinhardtii* combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses. Mol Biosyst 2010;6:1018–31.
[26] Brunner E, Ahrens CH, Mohanty S, Baetschmann H, et al. A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol 2007;25:576–83.
[27] Jungblut PR, Muller EC, Mattow J, Kaufmann SHE. Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. Infect Immun 2001;69: 5905–7.
[28] Henry CS, DeJongh M, Best AA, Frybarger PM, et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 2010;28: 977–82.
[29] Poolman MG, Miguet L, Sweetlove LJ, Fell DA. A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiol 2009;151:1570–81.
[30] Dal'Molin CGD, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK. AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. Plant Physiol 2010;152:579–89.
[31] Boyle NR, Morgan JA. Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. BMC Syst Biol 2009;3.
[32] Herrgard MJ, Swainston N, Dobson P, Dunn WB, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 2008;26:1155–60.
[33] Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 2010;465:627–31.
[34] Huang XH, Wei XH, Sang T, Zhao QA, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 2010;42:961–76.
[35] Johannsen W. The genotype conception of heredity. Am Nat 1911;XLV:129–59.
[36] Turesson G. The genotypical response of the plant species to the habitat. Hereditas 1922;III.
[37] Mendel G. Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn. Band 1865;IV:3–47.
[38] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc 1918;53: 399–433.
[39] Scherling C, Roscher C, Giavalisco P, Schulze ED, Weckwerth W. Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. PLoS One 2010;5:e12569.
[40] Pigliucci M, Hayden K. Phenotypic plasticity is the major determinant of changes in phenotypic integration in Arabidopsis. New Phytol 2001;152:419–30.
[41] Mitchell-Olds T, Pedersen D. The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. Genetics 1998;149:739–47.
[42] Keurentjes JJ. Genetical metabolomics: closing in on phenotypes. Curr Opin Plant Biol 2009;12:223–30.
[43] Keurentjes JJ, Fu J, de Vos CH, Lommen A, et al. The genetics of plant metabolism. Nat Genet 2006;38:842–9.
[44] Kliebenstein D, Pedersen D, Barker B, Mitchell-Olds T. Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in *Arabidopsis thaliana*. Genetics 2002;161:325–32.
[45] Kliebenstein DJ, Kroymann J, Brown P, Figuth A, et al. Genetic control of natural variation in Arabidopsis glucosinolate accumulation. Plant Physiol 2001;126:811–25.
[46] Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, et al. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet 1999;23:203–7.
[47] Kim S, Plagnol V, Hu TT, Toomajian C, et al. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 2007;39:1151–5.
[48] Chan EK, Rowe HC, Hansen BG, Kliebenstein DJ. The complex genetic architecture of the metabolome. PLoS Genet 2010;6: e1001198.
[49] Kliebenstein DJ. Systems biology uncovers the foundation of natural genetic diversity. Plant Physiol 2010;152:480–6.
[50] Brautigam A, Gowik U. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol (Stuttg) 2010;12: 831–41.
[51] Wang L, Li P, Brutnell TP. Exploring plant transcriptomes using ultra high-throughput sequencing. Brief Funct Genomics 2010;9:118–28.
[52] Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 2011;12:87–98.
[53] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008;133:523–36.
[54] Wolters DA, Washburn MP, Yates JR. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 2001;73:5683–90.
[55] Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 2001;19:242–7.

[56] Eng JK, Mccormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectrom 1994;5:976–89.
[57] Yates JR. Mass spectrometry — from genomics to proteomics. Trends Genet 2000;16:5–8.
[58] Hoehenwarter W, van Dongen JT, Wienkoop S, Steinfath M, et al. A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. Proteomics 2008;8:4214–25.
[59] Liu H, Sadygov RG, Yates III JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004;76:4193–201.
[60] Wienkoop S, Larrainzar E, Niemann M, Gonzalez E, et al. Stable isotope-free quantitative shotgun proteomics combined with sample pattern recognition for rapid diagnostics — a case study in *Medicago truncatula* nodules. J Sep Sci 2006;29:2793–801.
[61] Lehmann U, Wienkoop S, Tschoep H, Weckwerth W. If the antibody fails—a mass Western approach. Plant J 2008;55: 1039–46.
[62] Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell 2009;138:795–806.
[63] Wienkoop S, Weckwerth W. Relative and absolute quantitative shotgun proteomics: targeting low-abundance proteins in *Arabidopsis thaliana*. J Exp Bot 2006;57:1529–35.
[64] Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A 2003;100:6940–5.
[65] Wienkoop S, Larrainzar E, Glinski M, Gonzalez EM, et al. Absolute quantification of *Medicago truncatula* sucrose synthase isoforms and N-metabolism enzymes in symbiotic root nodules and the detection of novel nodule phosphoproteins by mass spectrometry. J Exp Bot 2008;59: 3307–15.
[66] Desiderio DM, Kai M. Preparation of stable isotope-incorporated peptide internal standards for field desorption mass-spectrometry quantification of peptides in biologic tissue. Biomed Mass Spectrom 1983;10:471–9.
[67] Desiderio DM, Kai M. Field desorption mass-spectral measurement of enkephalins in canine brain with O-18 peptide internal standards. Int J Mass Spectrom Ion Process 1983;48:261–4.
[68] Hummel J, Niemann M, Wienkoop S, Schulze W, et al. ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. BMC Bioinformatics 2007;8: 216.
[69] Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 2004;22:245–52.
[70] Weckwerth W. Metabolomics: an integral technique in systems biology. Bioanalysis 2010;2:829–36.
[71] Weckwerth W. Metabolomics: methods and protocols. Methods Mol Biol 2007;358:1–312.
[72] Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999;29:1181–9.
[73] Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 2002;1:153–61.
[74] Castrillo JO, Oliver SG. Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics. J Biochem Mol Biol 2004;37:93–106.
[75] Kell DB. Metabolomics and systems biology: making sense of the soup. Curr Opin Microbiol 2004;7:296–307.
[76] Dunn WB, Ellis DI. Metabolomics: current analytical platforms and methodologies. Trac-Trend Anal Chem 2005;24:285–94.
[77] Fiehn O. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. Trac-Trend Anal Chem 2008;27:261–9.
[78] Kempa S, Hummel J, Schwemmer T, Pietzke M, et al. An automated GCxGC–TOF–MS protocol for batch-wise extraction and alignment of mass isotopomer matrixes from differential C-13-labelling experiments: a case study for photoautotrophic–mixotrophic grown *Chlamydomonas reinhardtii* cells. J Basic Microbiol 2009;49:82–91.
[79] Sansone SA, Fan T, Goodacre R, Griffin JL, et al. The metabolomics standards initiative. Nat Biotechnol 2007;25:846–8.
[80] Morgenthal K, Wienkoop S, Scholz M, Selbig J, Weckwerth W. Correlative GC–TOF–MS based metabolite profiling and LC–MS based protein profiling reveal time-related systemic regulation of metabolite–protein networks and improve pattern recognition for multiple biomarker selection. Metabolomics 2005;1:109–21.
[81] Tohge T, Nishiyama Y, Hirai MY, Yano M, et al. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J 2005;42:218–35.
[82] Wienkoop S, Morgenthal K, Wolschin F, Scholz M, et al. 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 2008;7: 1725–36.
[83] Weckwerth W. Integration of metabolomics and proteomics in molecular plant physiology — coping with the complexity by data-dimensionality reduction. Physiol Plant 2008;132: 176–89.
[84] Scholz M, Selbig J. Visualization and analysis of molecular data. Methods Mol Biol 2007;358:87–104.
[85] Steuer R, Morgenthal K, Weckwerth W, Selbig J. A gentle guide to the analysis of metabolomic data. Methods Mol Biol 2007;358:105–26.
[86] Butlin RK. Population genomics and speciation. Genetica 2010;138:409–18.
[87] Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet 2005;6:805–14.
[88] Baerenfaller K, Grossmann J, Grobei MA, Hull R, et al. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 2008;320:938–41.
[89] Wienkoop S, Glinski M, Tanaka N, Tolstikov V, et al. Linking protein fractionation with multidimensional monolithic RP peptide chromatography/mass spectrometry enhances protein identification from complex mixtures even in the presence of abundant proteins. Rapid Commun Mass Spectrom 2004;18:643–50.
[90] Hoehenwarter W, Larhlimi A, Hummel J, Egelhofer V, et al. MAPA distinguishes genotype-specific variability of highly similar regulatory protein isoforms in potato tuber. J Proteome Res 2011;10:2979–91.
[91] Merchant SS, Prochnik SE, Vallon O, Harris EH, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 2007;318:245–50.
[92] Weckwerth W, Loureiro ME, Wenzel K, Fiehn O. Differential metabolic networks unravel the effects of silent plant phenotypes. Proc Natl Acad Sci U S A 2004;101:7809–14.
[93] Weckwerth W, Wenzel K, Fiehn O. Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. Proteomics 2004;4:78–83.

[94] Weckwerth W, Baginsky S, van Wijk K, Heazlewood JL, Millar H. The Multinational Arabidopsis Steering Subcommittee for Proteomics assembles the largest proteome database resource for plant systems biology. J Proteome Res 2008;7: 4209–10.
[95] Joshi HJ, Hirsch-Hoffmann M, Baerenfaller K, Gruissem W, et al. MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. Plant Physiol 2011;155:259–70.
[96] Bastow R, Beynon J, Estelle M, Friesner J, et al. An international bioinformatics infrastructure to underpin the Arabidopsis community. Plant Cell 2010;22:2530–6.
[97] Agrawal GK, Job D, Zivy M, Agrawal VP, et al. Time to articulate a vision for the future of plant proteomics — a global perspective: an initiative for establishing the International Plant Proteomics Organization (INPPO). Proteomics 2011;11:1559–68.
[98] Cox J, MAH R, James P, Jorrin-Novo JV, et al. Facing challenges in Proteomics today and in the coming decade: Report of Roundtable Discussions at the 4th EuPA Scientific Meeting, Portugal, Estoril 2010. J Proteomics in press.
[99] Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol 2008;6:693–9.
[100] Viant MR. Recent developments in environmental metabolomics. Mol Biosyst 2008;4:980–6.
[101] Laibach F. Zur Frage nach der Individualität der Chromosomen im Plfanzenreich. Beih Bot Zentralbl 1907;22: 191–210.
[102] Laibach F. *Arabidopsis thaliana* (L.) Heynh. als Object für genetische und entwicklungsphysiologische Untersuchungen. Bot Archiv 1943;44:439–55.
[103] Pigliucci M, Byrd N. Genetics and evolution of phenotypic plasticity to nutrient stress in Arabidopsis: drift, constraints or selection? Biol J Linn Soc 1998;64:17–40.
[104] Schmid KJ, Sorensen TR, Stracke R, Torjek O, et al. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res 2003;13:1250–7.
[105] Chevalier F, Martin O, Rofidal V, Devauchelle AD, et al. Proteomic investigation of natural variation between Arabidopsis ecotypes. Proteomics 2004;4:1372–81.
[106] Tilman D, Reich PB, Knops J, Wedin D, et al. Diversity and productivity in a long-term grassland experiment. Science 2001;294:843–5.
[107] Tilman D, Reich PB, Knops JM. Biodiversity and ecosystem stability in a decade-long grassland experiment. Nature 2006;441:629–32.
[108] Roscher C, Schumacher J, Baade J, Wilcke W, et al. The role of biodiversity for element cycling and trophic interactions: an experimental approach in a grassland community. Basic Appl Ecol 2004;5:107–21.
[109] Fernie AR, Schauer N. Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 2009;25: 39–48.
[110] Varshney RK, Graner A, Sorrells ME. Genomics-assisted breeding for crop improvement. Trends Plant Sci 2005;10: 621–30.
[111] de Vienne D, Leonardi A, Damerval C, Zivy M. Genetics of proteome variation for QTL characterization: application to drought-stress responses in maize. J Exp Bot 1999;50:303–9.
[112] Zivy M, de Vienne D. Proteomics: a link between genomics, genetics and physiology. Plant Mol Biol 2000;44:575–80.
[113] Collard BC, Mackill DJ. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 2008;363: 557–72.
[114] Schauer N, Semel Y, Roessner U, Gur A, et al. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 2006;24:447–54.
[115] Gebhardt C, Bellin D, Henselewski H, Lehmann W, et al. Marker-assisted combination of major genes for pathogen resistance in potato. Theor Appl Genet 2006;112:1458–64.
[116] Schnoor JL. Highlighting biofuels research. Environ Sci Technol 2010;44:8796–7.
[117] Raven J, Grasses Thomas H. Curr Biol 2010;20:R837–9.
[118] Somerville C, Youngs H, Taylor C, Davis SC, Long SP. Feedstocks for lignocellulosic biofuels. Science 2010;329: 790–2.
[119] Kullander S. Food security: crops for people not for cars. Ambio 2010;39:249–56.
[120] Studer MH, Demartini JD, Davis MF, Sykes RW, et al. Lignin content in natural Populus variants affects sugar release. Proc Natl Acad Sci U S A 2011;108:6300–5.
[121] Singh A, Nigam PS, Murphy JD. Renewable fuels from algae: an answer to debatable land based fuels. Bioresour Technol 2011;102:10–6.
[122] Sakuragi H, Kuroda K, Ueda M. Molecular breeding of advanced microorganisms for biofuel production. J Biomed Biotechnol 2011;2011:416931.
[123] Himmel ME, Ding SY, Johnson DK, Adney WS, et al. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 2007;315:804–7.
[124] Rupprecht J. From systems biology to fuel—*Chlamydomonas reinhardtii* as a model for a systems biology approach to improve biohydrogen production. J Biotechnol 2009;142: 10–20.
[125] Moroney JV, Ma Y, Frey WD, Fusilier KA, et al. The carbonic anhydrase isoforms of Chlamydomonas reinhardtii: intracellular location, expression, and physiological roles. Photosynth Res 2011;109:133–49.
[126] Xue S, Hu H, Ries A, Merilo E, et al. Central functions of bicarbonate in S-type anion channel activation and OST1 protein kinase in CO(2) signal transduction in guard cell. EMBO J 2011;30:1645–58.
[127] Hu H, Boisson-Dernier A, Israelsson-Nordstrom M, Bohmer M, et al. Carbonic anhydrases are upstream regulators of CO2-controlled stomatal movements in guard cells. Nat Cell Biol 2010;12:87–93 sup pp 81–18.
[128] Tolbert V, Schiller A. Environmental Enhancement Using Short-Rotation Woody Crops Environmental Enhancement through Agriculture: Proceedings of a Conference, Boston, Massachusetts, November 15–17, 1995; 1995.
[129] Scherling C, Ulrich K, Ewald D, Weckwerth W. A metabolic signature of the beneficial interaction of the endophyte Paenibacillus sp isolate and in vitro-grown poplar plants revealed by metabolomics. Mol Plant Microbe Interact 2009;22:1032–7.
[130] Ulrich K, Stauber T, Ewald D. Paenibacillus — a predominant endophytic bacterium colonising tissue cultures of woody plants. Plant Cell Tiss Org 2008;93:347–51.
[131] White B, Dasgupta A. Agrofuels capitalism: a view from political economy. J Peasant Stud 2010;37:593–607.
[132] Borlaug N. Feeding a hungry world. Science 2007;318:359.
[133] Borlaug NE. Contributions of conventional plant breeding to food production. Science 1983;219:689–93.