MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis

Rainer Schwacke[1], Gabriel Y. Ponce-Soto[1], Kirsten Krause[2], Anthony M. Bolger[3], Borjana Arsova[1], Asis Hallab[1], Kristina Gruden[4], Mark Stitt[5], Marie E. Bolger[1,*] and Björn Usadel[1,3]

[1]Institute for Bio- and Geosciences (IBG-2: Plant Sciences), Forschungszentrum Jülich, Wilhelm Johnen Straße, Jülich, Germany

[2]Department of Arctic and Marine Biology, The Arctic University of Norway, Biology Building, 9037 Tromsø, Norway

[3]Institute for Botany and Molecular Genetics, BioEconomy Science Center, Worringer Weg, RWTH Aachen University, 52074 Aachen, Germany

[4]National Institute of Biology, Department of Biotechnology and Systems Biology, Večna Pot 111, 1000 Ljubljana, Slovenia

[5]Max Planck Institute for Molecular Plant Physiology, Department of Systems Regulation, 14476 Potsdam-Golm, Germany

*Correspondence: Marie E. Bolger (m.bolger@fz-juelich.de)

https://doi.org/10.1016/j.molp.2019.01.003

ABSTRACT

Genome sequences from over 200 plant species have already been published, with this number expected to increase rapidly due to advances in sequencing technologies. Once a new genome has been assembled and the genes identified, the functional annotation of their putative translational products, proteins, using ontologies is of key importance as it places the sequencing data in a biological context. Furthermore, to keep pace with rapid production of genome sequences, this functional annotation process must be fully automated. Here we present a redesigned and significantly enhanced MapMan4 framework, together with a revised version of the associated online Mercator annotation tool. Compared with the original MapMan, the new ontology has been expanded almost threefold and enforces stricter assignment rules. This framework was then incorporated into Mercator4, which has been upgraded to reflect current knowledge across the land plant group, providing protein annotations for all embryophytes with a comparably high quality. The annotation process has been optimized to allow a plant genome to be annotated in a matter of minutes. The output results continue to be compatible with the established MapMan desktop application.

Key words: Plant genomes, Functional annotation, MapMan, Mercator, Transcriptomes

INTRODUCTION

Plant sciences have seen a dramatic increase in the use of high-throughput omics platforms, driven by recent technological improvements. Among the quantitative omics technologies, transcriptomics can be seen as mature with continually decreasing costs as protocols are improved (Tzfadia et al., 2018), metabolomics is widely adopted (Alseekh and Fernie, 2018), and proteomics offers great potential (Vanderschuren et al., 2013). Application of omics technologies to epigenetics data is allowing great strides, for example in understanding the role of plant DNA methylation (Zhang et al., 2018). Since these technologies are complementary, their combined use is uncovering novel key players in plant metabolism, signaling and

regulation, and driving functional pathway elucidation. This is especially the case for coupled transcriptomics and metabolite data, for example, in secondary metabolism (Fernie and Tohge, 2017; Wisecaver et al., 2017) and for transcriptomics in combination with chromatin immunoprecipitation sequencing to study transcriptional regulation (e.g., Ezer et al., 2017). Genomics analyses are greatly aided by pathway and process databases that capture existing knowledge and allow visualizations of individual or combined datasets. Cross-species genomics analyses are also playing an increasing role

in research. For example, comparative genome analyses can be leveraged to offer a better insight into plant gene regulatory networks (Ferrari et al., 2018). Integration of datasets from different omics technologies and from different genotypes and species is necessary for analysis of pan-genomic datasets, but is challenging and requires the creation of contextual frameworks. One example is the MapMan family of software, which allows users to evaluate omics data based on the biological context (Jaiswal and Usadel, 2016).

The association between a genomic locus and its immediate products in transcriptomic and proteomic datasets is usually trivial. However, the coupling of metabolites with transcripts/proteins (Fernie and Stitt, 2012) or of different transcripts/proteins to each other requires prior knowledge. While relationships between proteins can be established using large-scale protein–protein interaction screens (Altmann et al., 2018), these do not cover all potential interactions even in the case of the model plant *Arabidopsis thaliana*. Thus, machine-readable pathway and process annotations have been used as a stopgap. They can also be used to infer relationships between transcripts and metabolites. The reverse is also true, as ontologies can be inferred from transcriptional coupling using a guilt-by-association approach (Di Salle et al., 2017).

The MapMan framework was developed specifically for plants with the design goal to facilitate the visualization of omics data on plant pathways (Thimm et al., 2004). MapMan uses a simple hierarchical tree structure of terms referred to as "bins," which describe biological contexts/concepts. Major biological processes (e.g., photosynthesis) are encompassed in top-level bins, and each child bin represents a more narrowly focused subprocess or component within the context of the parent bin. Assignment of proteins to the lowest-level (i.e., leaf) bins was preferred in order to make the annotation as precise as possible, although assignment to abstract higher-level bins was supported. Proteins were mostly assigned to a single bin, but for some proteins with functions in diverse biological processes it was necessary to correspondingly assign to multiple bins.

While initially focused on metabolic processes, the MapMan framework rapidly evolved to include regulatory processes such as transcription factors and signaling pathways as well as biotic and abiotic stress responses. The ontology was exploited as the foundation for the MapMan application, which allows quantitative omics data to be visualized on functional pathways (Thimm et al., 2004). It also allows users to investigate enriched pathways and to functionally explore differentially expressed genes and accumulated metabolites.

Although MapMan was originally developed for use with the model species *A. thaliana*, it was later adapted to other species by similarity transfer and manual curation (Ling et al., 2013). This proved to be infeasible as a long-term approach, due to the rapid increase in the number of species for which genomics data were available. The tool Mercator was therefore developed to allow automatic annotation of plant protein sequences with MapMan terms (Lohse et al., 2014). Mercator relied on sequence similarity and, when appropriate, protein domains from InterPro (Mulder and Apweiler, 2007) and CDD (Marchler-Bauer et al., 2013) that had been manually assigned to the bins. However,

this approach resulted in many annotations to abstract levels, partly due to the absence of appropriate bins for the new species.

In terms of describing proteins, the most commonly used framework is the Gene Ontology (GO), which is widely used across all life forms. The GO framework defines terms and their relationship to each other as a means to formalize protein description (Gene Ontology Consortium, 2014). The terms are partitioned in three specific categories (named "Biological Process," "Molecular Function," and "Cellular Component"). These GO terms are arranged as a directed acyclic graph (DAG), where a child term may have more than a single parent. Generally, a single protein can expect to be annotated with a multitude of GO terms originating from each category. While GO is beneficial for a rich annotation, it can pose difficulties when it is used to visualize omics data because the multiple annotations lead to a strong redundancy (Jantzen et al., 2011).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) ontology is a collection of databases covering many different aspects of biology (Kanehisa et al., 2017). Of these, the KEGG Orthology (KO) database is the closest equivalent to the MapMan framework, using a similar hierarchical structure of protein function terms. KO encompasses genes from both eukaryotes and prokaryotes, and while it was traditionally focused on metabolism, it has been expanded to include a wider range of biological processes. Other frameworks such as PlantCyc (Schläpfer et al., 2017) and Plant Reactome (Naithani et al., 2017) focus primarily on metabolic processes, and, while highly detailed within their area of focus, do not cover a broad range of biological processes and thus cannot easily be compared with MapMan.

We have now completely redesigned the MapMan framework and developed a more powerful Mercator pipeline to sustainably annotate the proteome of any land plant. Here we present the first stable release of the new MapMan4 framework and the improved online tool Mercator4, and showcase their application in determining gene loss in parasitic plants.

RESULTS

MapMan4: A Novel Biological Context-Based Framework

The MapMan4 ontology represents a comprehensive set of common biological processes and incorporates genetic information from a wide variety of plant species. The core design principles from the original MapMan (Thimm et al., 2004) have been retained, such as the simple tree structure with each top-level category representing a main biological concept with each sublevel becoming increasingly specialized. In MapMan4, proteins are only classified into leaf node categories (Figure 1, Lopez-Obando et al., 2015), thus ensuring that all assignments receive precise protein descriptions. In contrast to the original MapMan framework, assigning proteins to top-level or intermediate nodes is no longer possible. The total number of bin categories has been almost tripled, with currently 4147 leaf nodes and 1340 branch nodes, compared with 1550 leaf nodes and 341 branch nodes in MapMan v.3. This increase provides a finer granularity that enables users to perform more precise analyses at the biological level.
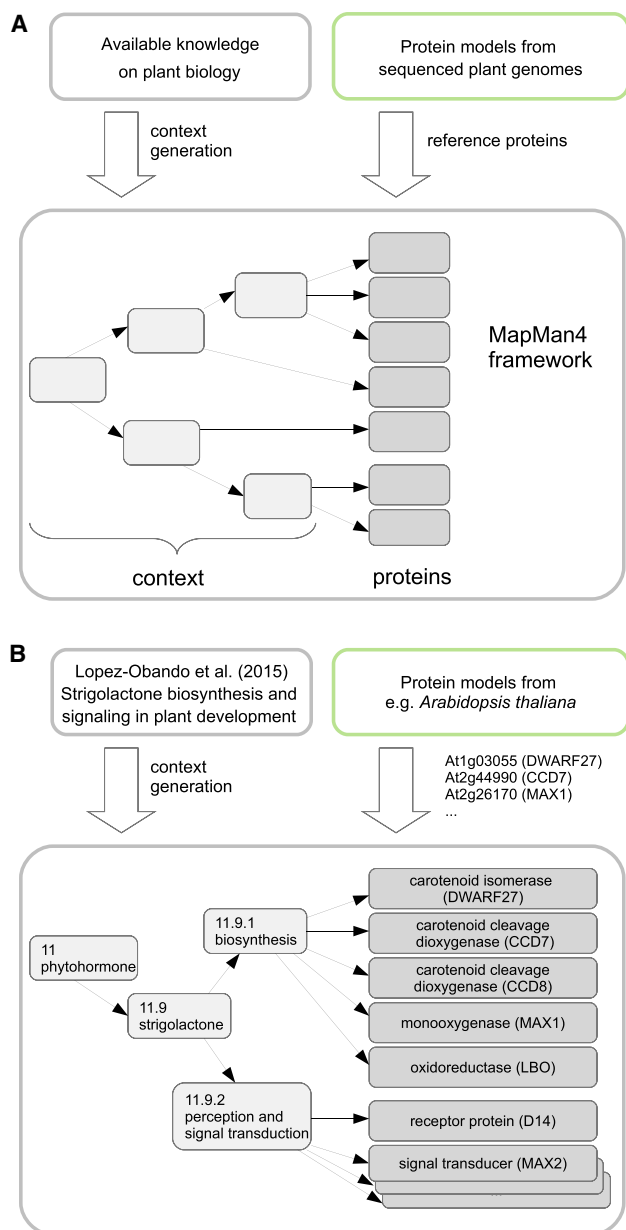
Figure 1. Scheme of the MapMan4 Framework.
(A) The MapMan4 hierarchical tree structure describes the biological context of proteins. The reference protein descriptions form the leaf nodes.
(B) Example generation of MapMan4 categories for the biosynthesis, perception, and signaling of the phytohormone strigolactone.

Currently, the MapMan4 ontology comprises 27 functional top-level categories representing a diverse range of biological processes (Table 1). In principle, these 27 top-level bins should contain only proteins that have a strong biological context, e.g., the well-defined function of a protein within a pathway. However, this strict approach results in the classification of around only one-third of plant proteins, due to limited plant biological knowledge.

Therefore, the criteria were broadened, in specific cases, to also accept proteins that had weaker biological contexts. One

example are transcription factors, which are simply classified using their canonical transcription factor family as context. MapMan4 currently distinguishes 91 transcription factor families. These families were designed primarily using PlantTFDB (Jin et al., 2017) and PlnTFDB (Pérez-Rodríguez et al., 2010) as a guide, but in some cases sequence comparison suggested additional subdivision. As an example, for the HD-ZIP family from PlantTFDB, this has been divided into *HD-ZIP I/II transcription factor*, *HD-ZIP III transcription factor*, and *HD-ZIP IV transcription factor* in MapMan4, following the structure from Ariel et al. (2007). Comparison of the MapMan4 transcription factor bins (BIN-15.7 and BIN-15.8) against the iTAK (Zheng et al., 2016) transcription factor classification for *A. thaliana* revealed 1688 common transcription factor genes, 256 genes in MapMan4 but not in iTAK, and 79 genes in iTAK but not in MapMan4, indicating substantial agreement between the annotations. Mutual Information between the specific classes assigned by iTAK and MapMan4 to these shared 1688 genes is 3.499, very close the maximum possible value, given the entropy of each classification (3.589 for iTAK, 3.665 for MapMan4), thus indicating almost complete agreement of the specific class of each transcription factor gene.

Other cases where limited functional context is available are the large enzyme families, which are currently gathered into Bin-50. This category includes proteins that are known to belong to enzyme families, but information pertaining to their specific function may not have been ascertained. This category follows the Enzyme Commission structure to the second level, and currently contains 50 categories applicable to plants. In compliance with the original MapMan v.3 framework, proteins that have not been classified are assigned to Bin-35. This bin is further subdivided into Bin-35.1 (not assigned.annotated) and Bin-35.2 (not assigned.not annotated), depending whether they can or cannot be assigned Swiss-Prot based annotations (for details see Methods).

Comparison with Kyoto Encyclopedia of Genes and Genomes

The KO uses a hierarchical structure of protein function terms similar to MapMan4 (see Introduction). However, MapMan4 focuses exclusively on the plant kingdom, and thus includes plant-specific processes at a finer level of granularity. The plant focus of MapMan4 is more apparent in some biological processes than others. In metabolism, there is a considerable similarity, with, e.g., hexokinases, which in KO are under the hierarchy "Metabolism.Carbohydrate metabolism.Glycolysis/Gluconeogenesis.HK; hexokinase" (ko0844), in MapMan4 as "Carbohydrate metabolism.sucrose metabolism.degradation.hexokinase" (BIN-3.1.4.3). In other areas, there is a substantially different organization, with important plant-specific processes, such as Cell wall and Nutrient uptake, which are top-level categories in MapMan4 distributed in unrelated parts of the KO hierarchy, based on, for example, the substrates involved.

The KO structure uses up to four levels, which results in very high branching factors, particularly between the third and fourth levels. In contrast, the MapMan4 framework currently uses up to eight levels, and can allocate hierarchy levels as needed to the biological process (e.g., Protein degradation.peptide

Top level bin name	Number of leaf bins	Plant species (release) –Number of bins containing at least one protein (% bins occupied) –Total number of proteins in this category (% of all proteins of that plant)			
		Arabidopsis thaliana (TAIR v10)	*Solanum lycopersicum* (ITAG v3.2)	*Triticum aestivum* (IWGSC v1)	*Zea mays* (AGP v4)
1 Photosynthesis	226	219 (97%)	183 (81%)	219 (97%)	193 (85%)
		288 (1.04%)	296 (0.83%)	814 (0.79%)	341 (0.86%)
2 Cellular respiration	136	132 (97%)	119 (88%)	133 (98%)	128 (94%)
		244 (0.88%)	234 (0.65%)	597 (0.58%)	306 (0.77%)
3 Carbohydrate metabolism	92	92 (100%)	92 (100%)	90 (98%)	90 (98%)
		232 (0.84%)	243 (0.68%)	740 (0.71%)	290 (0.73%)
4 Amino acid metabolism	135	131 (97%)	130 (96%)	131 (97%)	131 (97%)
		237 (0.86%)	242 (0.68%)	692 (0.67%)	331 (0.84%)
5 Lipid metabolism	173	171 (99%)	167 (97%)	165 (95%)	162 (94%)
		443 (1.60%)	485 (1.36%)	1495 (1.44%)	615 (1.56%)
6 Nucleotide metabolism	53	53 (100%)	52 (98%)	53 (100%)	53 (100%)
		103 (0.37%)	97 (0.27%)	270 (0.26%)	131 (0.33%)
7 Coenzyme metabolism	158	155 (98%)	154 (97%)	152 (96%)	150 (95%)
		221 (0.80%)	226 (0.63%)	643 (0.62%)	266 (0.67%)
8 Polyamine metabolism	12	11 (92%)	11 (92%)	11 (92%)	9 (75%)
		25 (0.09%)	25 (0.07%)	64 (0.06%)	22 (0.06%)
9 Secondary metabolism	93	86 (92%)	65 (70%)	64 (69%)	59 (63%)
		223 (0.81%)	180 (0.50%)	573 (0.55%)	189 (0.48%)
10 Redox homeostasis	47	47 (100%)	47 (100%)	46 (98%)	45 (96%)
		124 (0.45%)	137 (0.38%)	344 (0.33%)	164 (0.42%)
11 Phytohormones	140	138 (99%)	133 (95%)	126 (90%)	128 (91%)
		585 (2.12%)	597 (1.67%)	1489 (1.44%)	614 (1.55%)
12 Chromatin organization	113	113 (100%)	110 (97%)	109 (96%)	109 (96%)
		312 (1.13%)	357 (1.00%)	1305 (1.26%)	435 (1.10%)
13 Cell cycle	258	258 (100%)	252 (98%)	252 (98%)	247 (96%)
		448 (1.62%)	432 (1.21%)	1260 (1.22%)	558 (1.41%)
14 DNA damage response	67	67 (100%)	64 (96%)	67 (100%)	66 (99%)
		84 (0.30%)	83 (0.23%)	247 (0.24%)	104 (0.26%)
15 RNA biosynthesis	295	294 (100%)	288 (98%)	289 (98%)	285 (97%)
		2310 (8.36%)	2563 (7.17%)	7282 (7.03%)	3114 (7.89%)
16 RNA processing	328	326 (99%)	315 (96%)	314 (96%)	309 (94%)
		498 (1.80%)	515 (1.44%)	1404 (1.36%)	648 (1.64%)
17 Protein biosynthesis	328	327 (100%)	313 (95%)	325 (99%)	310 (95%)
		627 (2.27%)	626 (1.75%)	1668 (1.61%)	791 (2.00%)
18 Protein modification	299	294 (98%)	292 (98%)	296 (99%)	294 (98%)
		1485 (5.38%)	1465 (4.10%)	5324 (5.14%)	1742 (4.41%)
19 Protein degradation	187	186 (99%)	186 (99%)	186 (99%)	186 (99%)
		1044 (3.78%)	1089 (3.04%)	3405 (3.29%)	1307 (3.31%)
20 Cytoskeleton	107	102 (95%)	102 (95%)	101 (94%)	99 (93%)
		307 (1.11%)	281 (0.79%)	760 (0.73%)	368 (0.93%)

Table 1. Functional Assignment of *Arabidopsis thaliana*, *Solanum lycopersicum*, *Triticum aestivum* and *Zea mays* proteins Across MapMan4 Categories

Top level bin name	Number of leaf bins	Plant species (release) –Number of bins containing at least one protein (% bins occupied) –Total number of proteins in this category (% of all proteins of that plant)			
		Arabidopsis thaliana (TAIR v10)	*Solanum lycopersicum* (ITAG v3.2)	*Triticum aestivum* (IWGSC v1)	*Zea mays* (AGP v4)
21 Cell wall	126	122 (97%)	115 (91%)	116 (92%)	114 (90%)
		585 (2.12%)	540 (1.51%)	1648 (1.59%)	595 (1.51%)
22 Vesicle trafficking	212	212 (100%)	210 (99%)	209 (99%)	208 (98%)
		551 (1.99%)	538 (1.50%)	1361 (1.31%)	713 (1.81%)
23 Protein translocation	135	135 (100%)	132 (98%)	132 (98%)	128 (95%)
		198 (0.72%)	211 (0.59%)	606 (0.59%)	288 (0.73%)
24 Solute transport	174	171 (98%)	171 (98%)	171 (98%)	170 (98%)
		1137 (4.12%)	1268 (3.55%)	3936 (3.80%)	1433 (3.63%)
25 Nutrient uptake	52	46 (88%)	45 (87%)	43 (83%)	44 (85%)
		159 (0.58%)	134 (0.37%)	437 (0.42%)	153 (0.39%)
26 External stimuli response	111	97 (87%)	107 (96%)	99 (89%)	95 (86%)
		359 (1.30%)	313 (0.88%)	731 (0.71%)	331 (0.84%)
27 Multi-process regulation	38	38 (100%)	37 (97%)	37 (97%)	37 (97%)
		138 (0.50%)	145 (0.41%)	463 (0.45%)	209 (0.53%)
50 Enzyme classification	50	35 (70%)	36 (72%)	37 (74%)	40 (80%)
		1170 (4.24%)	1827 (5.11%)	6479 (6.26%)	1936 (4.90%)

Table 1. *Continued*

tagging.Ubiquitin (UBQ)-anchor addition (ubiquitylation)), the particular step in the process (UBQ-ligase E3 activities), the protein complex grouping (Cullin-based ubiquitylation complexes), the specific protein complex (SKP1-CUL1-FBX (SCF) E3 ligase complexes), and specific protein component within the complex (F-BOX substrate adaptor components.SKP component).

Comparison with Gene Ontology

The GO consists of a DAG of annotation terms, from three categories, Cellular Component, Molecular Function, and Biological Process (see Introduction). Furthermore, it is common in GO that a single protein will be annotated with multiple terms from each GO category, e.g., the *A. thaliana* Hexokinase 1 protein (HXK1, At4g29130) is annotated (on www.arabidopsis.org) with multiple Biological Processes including hexose catabolic process, glycolytic process, cellular glucose homeostasis, and likewise multiple Molecular Functions including hexokinase activity, ATP binding, and zinc ion binding. Since the various terms capture different aspects of the protein, the terms cannot easily be inferred from each other, and thus GO terms should be considered as a group rather than individually. As previously noted in the Introduction, the DAG-based structure and the assignment of multiple terms per protein makes the visualization and interpretation of the GO annotations more complex than those from MapMan4 or KEGG, where annotation with a single term is often possible.

Automatic Annotation of Plant Proteomes

The Mercator4 annotation process was assessed using the 57 available plant genomes from Ensembl Plants version 41 (Bolser et al., 2017) (Figure 2 and Supplemental Table 1). When considering all splice forms, the average protein classification rate was found to be 43.51% for dicots, 39.42% for monocots, and lower for other species (33.83%), reflecting the high diversity within the algae. Annotation rates were notably higher, with 64.65%, 58.5%, and 46.88% for dicots, monocots, and other species, respectively. Repeating this analysis using only the longest splice isoform of each gene resulted in a 1%–2% drop in the rate of bin assignment and annotation. These figures compare favorably with the annotation state of most sequenced plant species (Rhee and Mutwil, 2014).

For the model plant *A. thaliana*, the Mercator4 protein classification rate is currently at about 47% (Figure 2 and Supplemental Table 1). This compares favorably with a KEGG annotation using the KAAS pipeline (Moriya et al., 2007), which covered 32%. It is, however, lower than the rate of 64% achieved by the GO framework (having at least one GO term assigned to Molecular Function or Biological Process) (Bolger et al., 2018). However, as previously noted, assignment of a single GO term is not generally equivalent to a MapMan4 classification, and this annotation is the result of a large-scale community effort rather than an automated annotation pipeline.

Web Annotation Interface

To facilitate use of MapMan4 for any plant proteome or transcriptome, we have made available the corresponding Mercator4 pipeline for automated plant protein annotation (Figure 3, https://www.plabipd.de/portal/web/guest/mercator4). This web-based user interface allows the user to submit a text file (FASTA format) of either nucleotide or protein sequences. The user can provide an optional job name that will be used to name the result file. Multiple
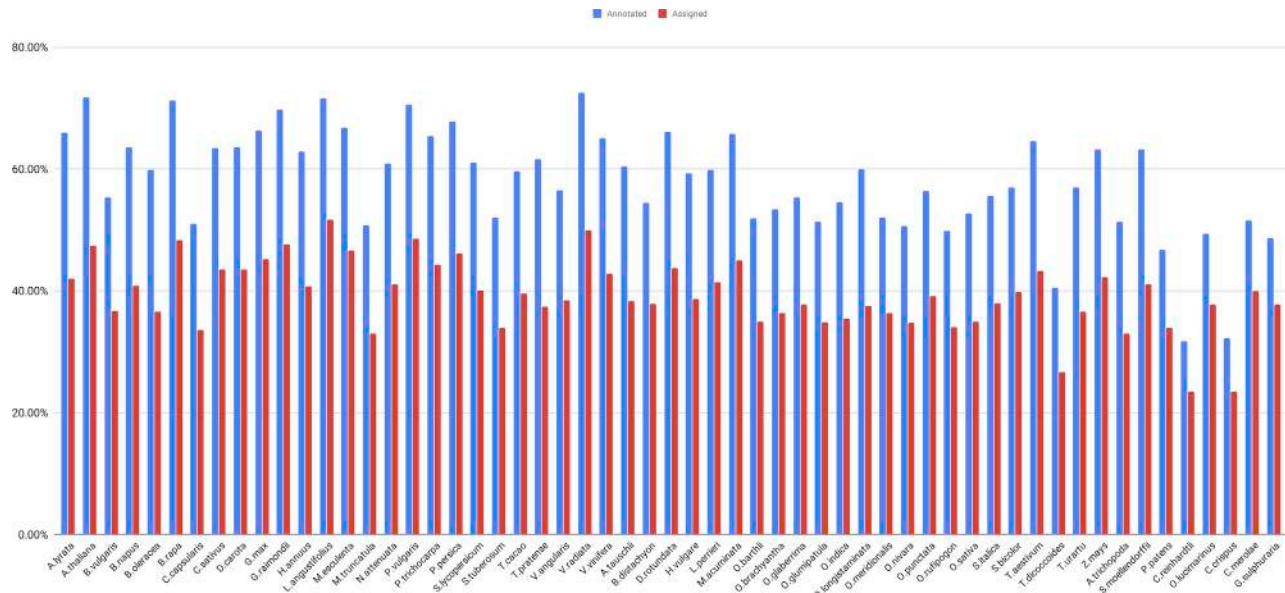
Figure 2. Mercator4 Annotation Rates for Plant Proteomes.
Fifty-seven plant proteomes from Ensembl Plants v41 categorized by the Mercator4 annotation. The diagram distinguishes between "classified" proteins (assignable to a MapMan4 Bin-1/27 and Bin-50) and "annotated" proteins (assignable to a MapMan4 Bin-1/27, Bin-50, or a Swiss-Prot protein entry). The rate is given as proportion of the total number of proteins.

jobs can be simultaneously submitted and monitored. Users have the option of providing an email address that will be notified when their jobs are complete.

Once a job has finished, a summary of the protein categorization is provided along with a bar chart. This bar chart shows how many of the leaf categories belonging to each top-level bin contain at least one protein. In cases where a complete proteome was submitted, this chart can immediately suggest whether some biological pathways are missing. However, this insight relies on the completeness of the submitted proteome because underpopulated categories could also be an indication of an incomplete proteome. Table 1 depicts these values for *A. thaliana*, tomato, wheat, and corn, and demonstrates that these values are usually above 90% for well-conserved processes.

For a more detailed view of the protein classification, the user can launch the Mercator4 Tree Viewer (Figure 4). This visualization shows the number of proteins assigned to each bin, displayed on the MapMan4 hierarchical tree. The tree structure can also be used to compare multiple proteomes. To support this, a selection of reference proteomes is provided for comparison purposes. A download option is provided that creates a tab-delimited text file for all jobs and for the selected reference proteomes. The data can easily be loaded into a variety of statistical programs to allow a more detailed analysis or to perform comparison between proteomes.

Protein classification and annotation is performed on an HPC cluster that was recently upgraded. Further speed enhancements were achieved by reducing the number of sequence comparison tools and reference databases used during the bin assignment and annotation. The post-processing code

was also optimized, resulting in a dramatic reduction in disk I/O. These hardware and software enhancements have resulted in speed improvements such that a typical diploid plant proteome (∼30 000 proteins) can be processed within a few minutes.

Legacy Versions

Access to the original Mercator v.3 will continue to be supported, and a separate tool (Legacy Mercator4) is available to support legacy versions of Mercator4 (https://www.plabipd.de/portal/legacy-mercator4). Providing access to legacy Mercator4 versions ensures that any analyses carried out will remain reproducible in the future. However, given that users could potentially run a job against a variety of versions, the Mercator4 tree viewer is not supported in the legacy version.

Visualization in the MapMan Desktop Application

Given the extensive redesign of the MapMan framework for this release, it was necessary to create new MapMan4 pathway diagrams to reflect these changes and provide compatibility with the MapMan desktop application. A series of new pathway diagrams have been released that enable the visualization of transcripts using the MapMan desktop application. These include, for example, new kinase families and nitrogen uptake. The new pathway diagrams are available via the MapMan website (https://mapman.gabipd.org/mapmanstore).

As data analyses are often performed online nowadays, we have developed a feature-reduced web version of the MapMan desktop application. This allows visualization of functional responses based on client web browser technology and can easily integrate into functional genomics analysis platforms. As all rendering and business logic is performed on the client side
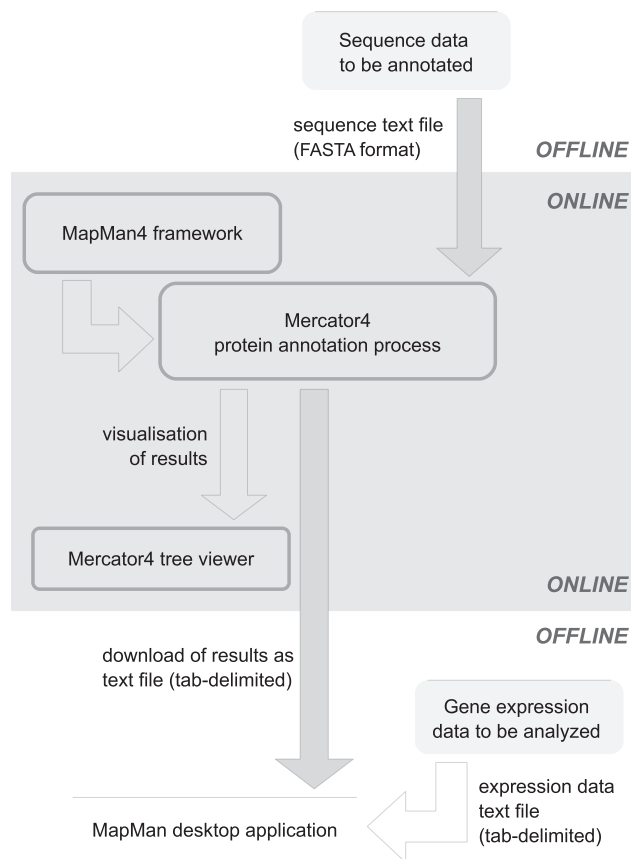
Figure 3. Overview of the Mercator4 Protein Annotation Process.

Mercator4 assigns MapMan4 categories to protein sequences provided by the user. The resulting categorization can be visualized online by the Mercator4 tree viewer (Figure 4) or downloaded as a tab-delimited text file. On a local computer, the file together with a corresponding expression data file can be loaded into the MapMan desktop application to visualize the expression for each MapMan4 category.

(i.e., on the web browser), and this component can be integrated into simple analysis platforms or static websites. Alternatively, the demo website can be downloaded and used as an embedded MapMan desktop application (https://usadellab.github.io/MapManJS/ultramicro.html).

Performance Measures and Comparison with Other Annotation Frameworks

Gene Network-Based Assessment of MapMan4 Protein Functional Annotation

A significant challenge with evaluating the effectiveness of protein functional annotation is the lack of large evaluation datasets that are independent of the protein sequence. One strategy to elucidate such sequence-independent information is by using "guilt-by-association" approaches based on expression information (Di Salle et al., 2017). It has previously been reported by Klie and Nikoloski (2012) that the MapMan v.3 ontology outperforms GO in this area, providing a higher annotation rate of unknown proteins based on, for example, a simple k-nearest neighbor approach. As the same design principles were adhered to in the redesigned and expanded MapMan4

ontology, we expected it to perform even better in the analysis of biological networks.

We compared the performance of MapMan4 and the original MapMan v.3 release. Using a similar approach to that described by Klie and Nikoloski (2012), we calculated simple gene networks based on Pearson and Spearman correlation for the model plant *A. thaliana*, based on data downloaded from GeneCAT (Mutwil et al., 2008). For each gene pair, we assessed where both corresponding protein members had strong context bin assignments and how many of these pairs shared at least one MapMan top-level category. As can be seen in Figure 5, the new MapMan4 framework consistently outperformed the old MapMan v.3 annotation in terms of precision, regardless of whether the gene networks were constructed using Pearson or Spearman correlation thresholds.

Assessment of MapMan4 Using the Gene Ontology Framework

To bridge and compare the MapMan4 annotation with the corresponding GO annotation, 572 412 reference protein sequences from the public Swiss-Prot database were annotated by Mercator4 (using 3989 distinct MapMan4 leaf categories). On average, 143 reference proteins were assigned to each MapMan4 bin. The GO terms for each individual protein of a MapMan4 category were extracted, proteins without GO annotations discarded, and the GO terms shared by all proteins assigned to the MapMan4 bin. On average, 41 GO terms were assigned to each bin. Less than 1% of the MapMan4 bins were assigned no GO term at all. This was either because there was a single reference protein without GO annotations assigned to the bin (0.1%) or because the reference proteins did not share any GO terms (0.75%).

Approximately 34% of the MapMan4 bins showed a non-unique pattern of GO terms (i.e., the GO term collection was shared with at least one other bin). In some cases, these shared GO term patterns are the result of the same protein function existing in different biological contexts. In other cases, they were caused by the fine granularity of the MapMan4 bins—for example, the individual ribosomal proteins are categorized into many specifiable MapMan4 bins that all share the same GO term pattern. Sometimes, the shared term patterns may have been a result of the stringent filtering requiring a GO term to be shared by all proteins before getting included. This therefore resulted in an enrichment of the higher-level, more generic GO terms that would be expected to be present in a number of bins.

As an assessment of specificity or detail of the protein function description formed by the overall GO annotation of a bin, we computed the depth of each GO term. The depth of a given GO term was measured as the minimum number of edges separating the root node of the GO graph from the respective GO term. The relationship between the depth of a GO term and its information content was previously analyzed by Klie and Nikoloski (2012) using GO annotations from the *A. thaliana* proteome. They found that a depth value of 6 corresponds to an information content of 90% of its maximum content, when terms from the GO categories "Biological Process" and "Cellular Component" are assessed. The depth value of the MapMan4 bins (compound GO annotation) was calculated to

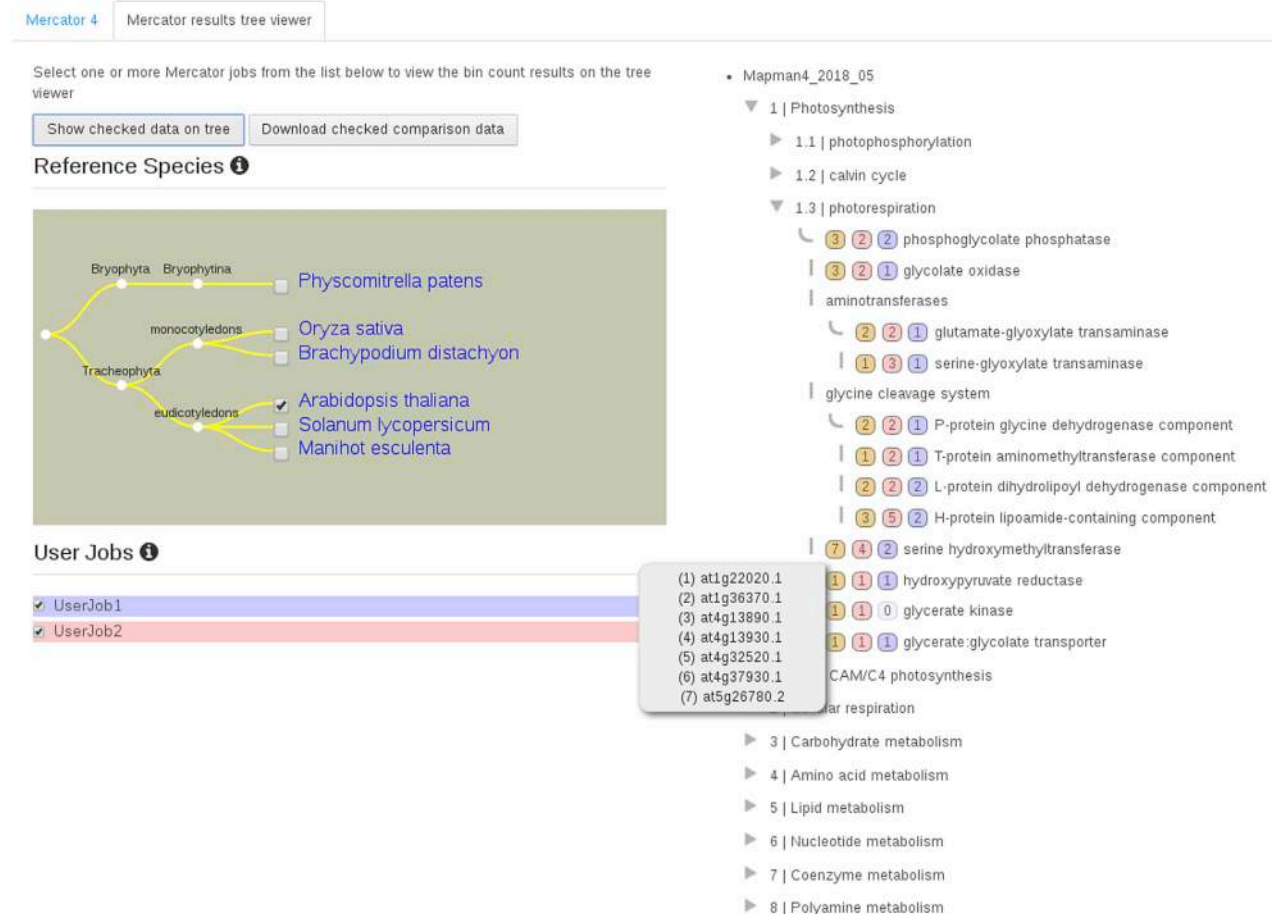# Mercator4 v1.0 Release Version

Figure 4. Mercator4 Tree Viewer.
Screenshot of the Mercator4 tree viewer that compares annotations from user jobs and—this example—from the reference species *Arabidopsis thaliana*. Each entry (colored circles within the tree) has the number of proteins assigned to a certain MapMan4 category, displayed with the option to see the protein names by hovering over the protein number.

be 6.6 on average, thus suggesting a very detailed description of the protein function.

To assess the quality of Mercator4-inferred GO terms, we used the *Oryza sativa* proteins, from UniProt/trEMBL, as a test dataset. To avoid bias, 3467 proteins that overlapped with Swiss-Prot were removed. As a gold standard for the annotation, we downloaded the latest *O. sativa* GO term annotation (http://geneontology.org/gene-associations/gene_association.gramene_oryza.gz) and retained only those annotations that had experimental evidence or were made by a human expert curator. For reasons of completeness, the remaining GO annotations were extended with their respective ancestral terms as obtainable from the directed GO graph. This resulted in a gold standard of 19 629 distinct GO annotations. Due to removing electronically inferred annotations, 68 572 proteins were left unannotated and were also removed. In total, 616 proteins from the original dataset of 72 655 remained.

Mercator4 was used to classify the test proteins, and GO terms were inferred from these assignments. An independent GO term assignment was also performed using InterProScan. Mercator4 and InterProScan inferences were compared with the gold standard annotation, and resulted in a Matthews Correlation Coefficient (MCC) of 0.16 for both Mercator4 and InterProScan (with Mercator4 yielding a slightly higher value), indicating that GO terms inferred are indeed meaningful and not biased.

Strikingly, when the MapMan4 categories were mapped onto GO terms using only a subset of 41 898 Swiss-Prot plant proteins, the GO annotation results were similar to the results obtained using all Swiss-Prot references (see above). This suggests that the MapMan4 framework is not only able to annotate protein functions present in plants, but also to assign these functions to conserved non-plant proteins. However, protein functions not found in plants are beyond the scope of the MapMan4 framework.

Usage Example: Detecting Gene Losses in *Cuscuta* Species
Data mining for genome changes should be promising approach when analyzing gene loss in plants with a reduced gene content requirement such as parasitic plants. In a test case, we used the
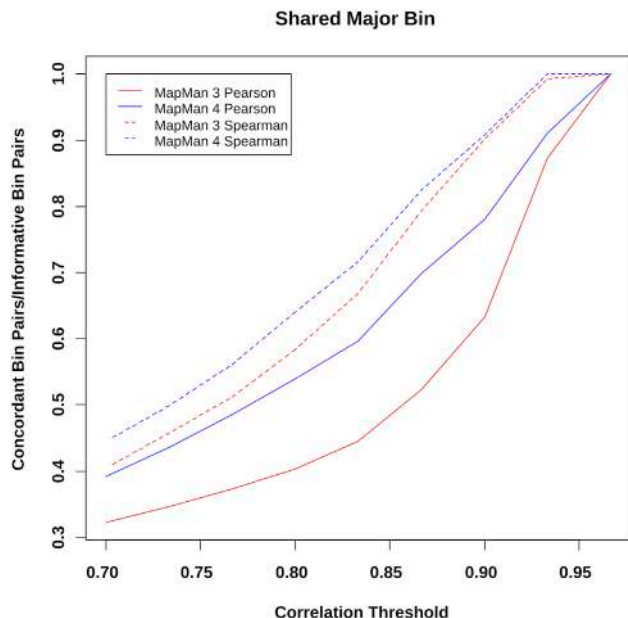
Shared Major Bin

Figure 5. Analysis of Concordant Bin Pairs in Correlation Networks.

Analysis was performed for different correlation thresholds in the *Arabidopsis* GeneCAT expression dataset. Gene pairs are displayed (as fraction of all informative pairs) for which both corresponding protein sequences were assigned to the same top-level category. The original MapMan v.3 annotation is shown in red and the MapMan4 annotation in blue. Solid lines represent Pearson correlation and dotted lines represent Spearman correlation.

Mercator4 to analyze the proteomes deduced from the genome of the holoparasitic species *Cuscuta campestris* (Vogel et al., 2018) and *Cuscuta australis* (Sun et al., 2018), the plastid-encoded proteomes of *Cuscuta gronovii* (Funk et al., 2007) and *Cuscuta obtusiflora* (McNeal et al., 2007), and the proteomes of the related autotrophic species *Ipomoea nil* (Hoshino et al., 2016), *Solanum lycopersicum* (Tomato Genome Consortium, 2012), and the plastidial proteome of *Ipomoea batatas* (Yan et al., 2015).

Transcription in plastids is mediated by two different RNA polymerases, a nucleus-encoded RNA polymerase, and a plastid-encoded RNA polymerase (PEP) complex (Yu et al., 2014; Pfannschmidt et al., 2015). The PEP is predominantly responsible for transcribing genes involved in photosynthesis (Yagi and Shiina, 2014) and consists of the conserved plastid-encoded core subunits rpoA, rpoB, rpoC1, and rpoC2. All the core subunits have been reported as missing in the plastid genomes of *C. gronovii* (Funk et al., 2007) and *C. obtusiflora* (McNeal et al., 2007). The Mercator4 analysis with visualization of Bin-15.9 (*RNA biosynthesis.organelle machineries.RNA polymerase activities*) confirms the lack of the PEP core subunits (Figure 6B) in these *Cuscuta* species. In comparison, the plastid genome of the related autotrophic *I. batatas* contains the genes for the PEP subunits (Figure 6B). In addition, Mercator4 reveals that the nuclear genomes of *C. campestris* and *C. australis* have experienced major losses of the PAP/pTAC PEP-associated co-factors (Figure 6A). Moreover, the putative regulatory co-factors PrdA and Prin2, as well as the six Sigma-

like factors required for the initiation of the plastidial transcription by PEP, are lacking. Interestingly, the transcriptionally active plastid chromosome proteins pTAC9 and pTAC17 were identified by Mercator4 in both *Cuscuta* proteomes (Figure 6A).

It has been shown that many plastidial RNA editing sites were abolished in *Cuscuta* species (Tillich and Krause, 2010). The visualization of Bin-16.10.2 (*RNA processing.organelle machineries.RNA editing*) confirms major losses of plastidial RNA editing factors in *Cuscuta* species. In accordance with this reduction in editing sites, 11 of the 15 editing factors found in tomato and *I. nil* are lacking in *C. campestris* and *C. australis* (Figure 6C).

Comparison Using Other Ontologies and Automated Annotation Tools

To determine whether the biological insights described above, regarding the PEP complex and plastidial RNA editing factors, could potentially be discovered with other annotation pipelines and frameworks, we reannotated the 27 PEP core subunits and associated factors and 32 plastidial RNA editing factors from *S. lycopersicum*, using both the previous Mercator v.3 release and the KAAS pipeline (Moriya et al., 2007).

Mercator4 had assigned these 59 proteins to 42 different bins, illustrating the fine granularity of the MapMan4 framework structure. Furthermore, these categories were coherently structured under two branch nodes, representing the PEP complex components and RNA editing factors, respectively, making the loss of these mechanisms readily apparent by comparing the gene counts across species.

In contrast, Mercator v.3 could annotate 33 of the 59 proteins. Of these, 17 proteins were classified in broadly appropriate bins (*RNA.transcription*, *RNA.regulation of transcription*, *RNA.RNA binding*) while 16 others were assigned to unrelated bins. Loss of the PEP core subunits, which are assigned to *RNA.transcription*, would however be difficult to notice as this bin is quite general, and thus many other proteins are assigned to the same category. Likewise, any signal from the loss of PEP-associated components and plastidial RNA editing factors would be difficult to discern, due to the low annotation rate and large number of other proteins in the relevant categories.

Annotation using the KAAS pipeline could assign only 11 of the 59 proteins. Of these, the four PEP core subunits were correctly assigned to the PEP RNA polymerase KO groups, while others were assigned to *spliceosome*, *signal transduction*, and *chaperones* KO groups. Loss of the PEP core subunits would be clear within the RNA polymerase KO groups; however, the related loss of PEP-associated components and loss of plastidial RNA editing factors would likely be missed.

DISCUSSION

The rapidly expanding number of available plant genome sequences (https://www.plabipd.de/plant_genomes_pa.ep) offers an opportunity for unraveling protein function through comparative gene regulatory network analysis (Ferrari et al., 2018). In addition, it widens the application of transcriptomics and proteomics tools (Sheth and Thaker, 2014) and plant genome scale metabolic prediction (De Oliveira Dal'molin and Nielsen,
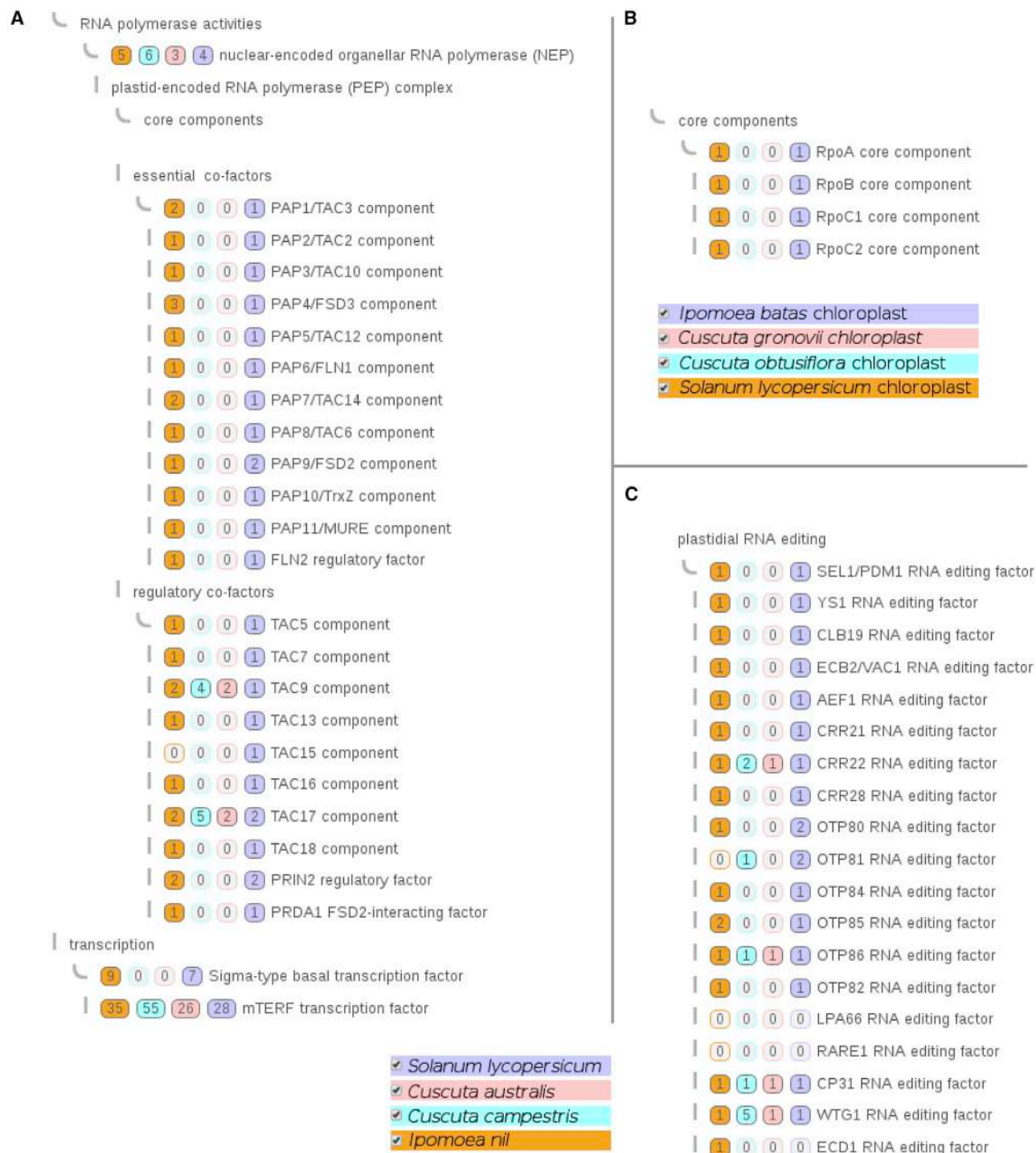
Figure 6. Lacking Components of the Plastidial RNA Polymerase Complex and Plastidial RNA Editing Factors in *Cuscuta* spp.

(A) Screenshot of the Mercator4 tree viewer for Bin-15.9 (RNA biosynthesis.organelle machineries). Most of the plastid transcriptionally active chromosome (TAC) components are not available in *C. campestris* and *C. australis* (inner columns), while the proteomes of *I. nil* (outer left column) and *S. lycopersicum* (outer right column) contain all nuclear encoded components of the RNA polymerase PEP complex.

(B) Screenshot as in **(A)** but for plastid-encoded proteomes. The plastid-encoded core components of the RNA polymerase (PEP) complex are available in the plastid proteomes of *S. lycopersicum* and *I. batatas* but not the parasitic *C. gronovii* and *C. obtusiflora* (inner columns).

(C) Screenshot of the Mercator4 tree viewer for Bin-16.10.2 (RNA processing.organelle machineries.RNA editing). Many plastidial RNA editing factors, while present in autotrophic Solanales species (outer columns), are not available in parasitic *Cuscuta* species (inner columns). The Mercator4 tree viewer also includes the RNA editing factors LPA66 and RARE1, which do not seem to occur in any of the Solanales plant species.

2013) to phylogenetically remote plants. Genomics and systems biology share common ground (Conesa and Mortazavi, 2014) and a draft genome is often the start for additional downstream analyses (Cuevas et al., 2016). However, protein sequences need to be put into a functional biological context to enable meaningful genome comparison within and between species. The demand for functional annotation is already visible in the use of the original Mercator tool, which has processed more than 12 000 datasets (i.e., genomes and transcriptomes) since its release (Lohse et al., 2014). While there is likely to be considerable redundancy between these datasets with regard to species, a recent study (Rai et al., 2017) came to the conclusion that almost 1000 different plant species have been studied by RNA-sequencing technology. This is likely to be an underestimate as it was based merely on publicly available data extracted from the NCBI Sequence Read Archive (https://www. ncbi.nlm.nih.gov/sra).

An increasing number of studies are coupling metabolite and transcriptome profiling data. This has been done in model plants for some years (Gibon et al., 2006; Satou et al., 2014). A rapidly expanding area is secondary metabolism in medicinal or ornamental plants, with the goal of discovering novel pathway members and gaining insights into pathway regulation (Polturak et al., 2018; Scossa et al., 2018).

Thus, the need for functional annotation and tools to analyze and interpret multi-omics data for model and in particular non-model plant species is increasing (Bolger et al., 2018). This is especially necessary to bridge between omics data and research questions about physiology (Do Amaral and Souza, 2017). Mercator was initially developed for MapMan against this background (Lohse et al., 2014), but its extension was restricted due to the underlying design of MapMan. Thus, we have redesigned the MapMan framework and have adapted Mercator4 to reflect these changes. While the current version of Mercator4 does not yet capture as many protein sequences as the original version (Supplemental Table 2), the new classifications are more specific.

The increased specificity together with the design goals offer several advantages. (1) Statistical representation analyses such as an over-representation analysis is straightforward and can be restricted to any level on the hierarchical bin tree. Although the previous version of MapMan v.3 already performed reasonably well for protein functional annotation (Klie and Nikoloski, 2012), performance has been substantially improved by a complete redesign of the framework. (2) The new MapMan4 framework is fully deterministic and can be applied to any land plant proteome to allow a consistent comparison between proteomes. Finally, (3) the annotation process performed by the online tool Mercator4 takes only minutes to annotate a complete plant proteome.

To demonstrate the usefulness of the MapMan4 framework, we analyzed the deduced proteomes of two *Cuscuta* species. For survival, these holoparasitic plants strictly rely on a nutrient supply from their host plants, which has relieved them from an autotrophic lifestyle. In line with this evolutionary step, the plastid and nuclear genomes have experienced extensive losses in genes related to photosynthesis and other processes not needed

for a parasitic lifestyle (Krause, 2008; Vogel et al., 2018). With the new MapMan4 framework, substantial protein losses in categories containing components of the PEP complex and protein factors involved in plastidial RNA editing are detectable in detail and can be visualized by the Mercator4 tree viewer (Figure 6). The Mercator4 annotation enables a quick and reliable survey of which proteins within a certain biological context are available or not in a given plant proteome.

METHODS

Design of the MapMan4 Framework

The MapMan4 hierarchical category structure was designed based on published experimental evidence and textbook knowledge (Figure 1). This also provided initial reference proteins that were used as seeds to find orthologs in high-quality annotated plant genomes across the plant kingdom. Determining whether a protein is a true ortholog involved manual examination of multiple sequence alignments from many plant species. The curated set of orthologous proteins identified for each MapMan4 bin was used to create one or more bin-specific Hidden Markov Models (HMMs), which can identify orthologous proteins from additional species.

The main functional bins in MapMan4 were designed to contain proteins that have a strong biological context, e.g., the well-defined function of a protein within a pathway. However, it was necessary to relax this criterion for transcription factors, which are simply classified using their canonical transcription factor family as context. Another case where limited functional context is available are the large enzyme families, which were currently gathered in Bin-50. This category includes proteins that are known to belong to enzyme families, but information pertaining to their specific function may not have been ascertained.

Automatic Annotation of Plant Proteomes

Each protein to be classified is tested against the bin-specific HMMs, using *hmmscan* from the HMMER3 software package (Eddy, 2011). If the provided sequences are nucleotides, a 6-frame translation of the sequence is generated for testing against the HMMs.

Many proteins that are assigned to the subcategories under the 27 main categories, based on their biological role, will also have assignment to categories under Bin-50 on the basis of their enzymatic activity. In these cases, assignments to the 27 main categories are considered to have a higher priority, and the redundant enzymatic activity bin assignment is filtered out.

The remaining unclassified proteins are subsequently compared with protein sequences contained in the Swiss-Prot database (The UniProt Consortium, 2018) using BLASTP. Positive matches (using a BLAST bit score >80) inherit the annotation from the matching Swiss-Prot hit and move to Bin-35.1 (not assigned.annotated). This provides an annotation for a number of proteins that were not assigned to a functional MapMan4 bin. Proteins that remain unannotated are placed in Bin-35.2 (not assigned.not annotated).

Mercator4 Implementation

The Mercator4 annotation web tool was implemented using the Java Portlet technology on a Liferay Portal (https://www.liferay.com). The front-end website is written in HTML with interactive components provided using the Javascript library D3 (https://d3js.org). The Javascript code for the collapsible tree visualization module is based on code created by Kate Morley (http://code.iamkate.com).

Submitted FASTA-formatted text files are split and distributed to a HPC cluster running Grid Engine. All back-end pre- and post-processing of

the jobs has been written in Java. The Distributed Resource Management Application API (DRMAA) is used to submit and monitor the jobs. The results are evaluated and collated before generation of the output file, which is presented to the user for download.

Proteome Annotation

All available plant proteomes were downloaded from Ensembl Plants (release 41, ftp://ftp.ensemblgenomes.org/pub/plants/release-41). Genomes that contained multiple splice variants were further processed to remove all but the longest form. Before being processed by Mercator4, each FASTA-formatted file was validated using the Mercator FASTA Validator, with records shorter than 5 amino acids or longer than 25000 amino acids removed (https://www.plabipd.de/portal/web/guest/mercator-fasta-validator). The validated files were submitted to the online Mercator4 annotation tool (https://www.plabipd.de/portal/mercator4).

The protein sequences used for the Mercator4 usage example were from the genomes of *C. campestris* (Vogel et al., 2018), *C. australis* (Sun et al., 2018), *S. lycopersicum* (Tomato Genome Consortium, 2012), and *I. nil* (Hoshino et al., 2016). Because plastid-encoded protein sequences from *C. campestris*, *C. australis*, and *I. nil* were not available, plastid-encoded protein sequences from *C. gronovii* (Funk et al., 2007), *C. obtusiflora* (McNeal et al., 2007), and *I. batatas* (Yan et al., 2015) were included instead.

Analysis of Concordant Bin Pairs in Correlation Networks

Within the expression dataset for *A. thaliana* (file ExpMatAra.exp, available at http://aranet.mpimp-golm.mpg.de/download.html), all ambiguous gene code assignments were deleted. In addition, for genes measured by multiple probes only one was retained. The resulting gene expression matrix was loaded into R and a correlation network was calculated using the R cor (correlation between matrices) function both for Spearman and Pearson correlation. The resulting matrix was then inspected for correlation values between genes with correlation thresholds of 0.7–0.966 in steps of 0.033 and only pairs where both proteins had an assignment in Bin-1 to 27 for MapMan4, and in Bin-1 to 34 (without Bin-26) for MapMan v.3. The pair was counted as concordant if at least one MapMan top-level bin assignment between the two proteins was shared and discordant otherwise. Code and data are available from https://github.com/usadellab/MapMan-Mercator-4.

Web-Available MapMan Application

To allow the integration of the MapMan application into web services, we have ported some basic MapMan application components into Javascript. In addition, we have implemented code for data visualization that relies on Javascript D3 (https://d3js.org) and allows quick rendering of simple data formats. An over-representation analysis is also included based on Javascript D3 and relies on a Fisher's exact test that uses the Lanczos approximation to compute the gamma function (Lanczos, 1964). Values and code were translated from the GNU Scientific Library. All components were written to be maximally portable and light, therefore pathway files, for example, need to comply with the formatting as provided by MapMan to allow simpler parsing. The code in addition to a working version is available from https://github.com/usadellab/usadellab.github.io.

MapMan4 Comparison with Gene Ontology

The complete Swiss-Prot protein sequence dataset was annotated by Mercator4 to find sets of reference proteins with each set representing a MapMan4 leaf category. The subsequent GO annotation for each reference protein also includes ancestral GO terms, i.e., terms found in all paths leading from the GO root term to the respective descendant term. Reference proteins without any GO annotation were removed from the sets. The collected GO terms for a reference protein form what we call the compound GO annotation of the protein. Finally, a MapMan4 category was annotated by the GO terms shared by all reference proteins of the

category. To evaluate how detailed the GO annotation of a MapMan4 category describes a protein function, we measured the depth of each GO term appearing in the bin. The depth is defined as the minimum number of edges leading from the root of the GO graph to the respective GO term.

We assessed the quality of protein function predictions obtained using Mercator4 and InterProScan (Quevillon et al., 2005). Performance was assessed on a gold standard of rice proteins that were not present in Swiss-Prot and that had manually curated GO annotations on the GO website. As performance measures we used the MCC (Matthews, 1975; Powers, 2011). Before calculating these measures, all GO term predictions were extended to include related ancestral terms. All material, code, documentation, and results are available as R-package (MapMan2GO, http://github.com/usadellab/MapMan2GO).

REFERENCES

Alseekh, S., and Fernie, A.R. (2018). Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J. **94**:933–942.

Altmann, M., Altmann, S., Falter, C., and Falter-Braun, P. (2018). High-quality yeast-2-hybrid interaction network mapping. Curr. Protoc. Plant Biol. **3**:e20067.

Ariel, F.D., Manavella, P.A., Dezar, C.A., and Chan, R.L. (2007). The true story of the HD-Zip family. Trends Plant Sci. **12**:419–426.

Bolger, M.E., Arsova, B., and Usadel, B. (2018). Plant genome and transcriptome annotations: from misconceptions to simple solutions. Brief. Bioinform. **19**:437–449.

Bolser, D.M., Staines, D.M., Perry, E., and Kersey, P.J. (2017). Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomic data. Methods Mol. Biol. **1533**:1–31.

Conesa, A., and Mortazavi, A. (2014). The common ground of genomics and systems biology. BMC Syst. Biol. **8**:S1.

Cuevas, D.A., Edirisinghe, J., Henry, C.S., Overbeek, R., O'Connell, T.G., and Edwards, R.A. (2016). From DNA to FBA: how to build your own genome-scale metabolic model. Front. Microbiol. **7**:907.

De Oliveira Dal'Molin, C.G., and Nielsen, L.K. (2013). Plant genome-scale metabolic reconstruction and modelling. Curr. Opin. Biotechnol. **24**:271–277.

Di Salle, P., Incerti, G., Colantuono, C., and Chiusano, M.L. (2017). Gene co-expression analyses: an overview from microarray collections in *Arabidopsis thaliana*. Brief. Bioinform. **18**:215–225.

Do Amaral, M.N., and Souza, G.M. (2017). The challenge to translate OMICS data to whole plant physiology: the context matters. Front. Plant Sci. **8**:2146.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. **7**:e1002195.

Ezer, D., Jung, J.H., Lan, H., Biswas, S., Gregoire, L., Box, M.S., Charoensawan, V., Cortijo, S., Lai, X., Stöckle, D., et al. (2017). The evening complex coordinates environmental and endogenous signals in *Arabidopsis*. Nat. Plants **3**:17087.

Fernie, A.R., and Stitt, M. (2012). On the discordance of metabolomics with proteomics and transcriptomics: coping with increasing complexity in logic, chemistry, and network interactions scientific correspondence. Plant Physiol. **158**:1139–1145.

Ferrari, C., Proost, S., Ruprecht, C., and Mutwil, M. (2018). PhytoNet: comparative co-expression network analyses across phytoplankton and land plants. Nucleic Acids Res. **46**:W76–W83.

Fernie, A.R., and Tohge, T. (2017). The genetics of plant metabolism. Annu. Rev. Genet. **51**:287–310.

Funk, H.T., Berg, S., Krupinska, K., Maier, U.G., and Krause, K. (2007). Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. BMC Plant Biol. **7**:45.

Gene Ontology Consortium. (2014). Gene ontology consortium: going forward. Nucleic Acids Res. **43** (Database issue):D1049–D1056.

Gibon, Y., Usadel, B., Blaesing, O.E., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. Genome Biol. **7**:R76.

Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., Shin, T., Minakuchi, Y., Koda, Y., Nagano, A.J., et al. (2016). Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. Nat. Commun. **7**:13295.

Jaiswal, P., and Usadel, B. (2016). Plant pathway databases. Methods Mol. Biol. **1374**:71–87.

Jantzen, S.G., Sutherland, B.J., Minkley, D.R., and Koop, B.F. (2011). GO Trimming: systematically reducing redundancy in large Gene Ontology datasets. BMC Res. Notes **4**:267.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. **45**:D1040–D1045.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45**:D353–D361.

Klie, S., and Nikoloski, Z. (2012). The choice between MapMan and gene ontology for automated gene function prediction in plant science. Front. Genet. **3**:115.

Krause, K. (2008). From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. Curr. Genet. **54**:111–121.

Lanczos, C. (1964). A precision approximation of the gamma function. SIAM J. Numer. Anal. **1**:86–96.

Ling, M.H., Rabara, R.C., Tripathi, P., Rushton, P.J., and Ge, X. (2013). Extending MapMan ontology to tobacco for visualization of gene expression. Dataset Pap. Biol. **2013**:706465.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A.R., Stitt, M., and Usadel, B. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ. **37**:1250–1258.

Lopez-Obando, M., Ligerot, Y., Bonhomme, S., Boyer, F.D., and Rameau, C. (2015). Strigolactone biosynthesis and signaling in plant development. Development **142**:3615–3619.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., et al. (2013). CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. **41**:D348–D352.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta **405**:442–451.

McNeal, J.R., Kuehl, J.V., Boore, J.L., and de Pamphilis, C.W. (2007). Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Plant Biol. **7**:57.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. **35**:W182–W185.

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol. Biol. **396**:59–70.

Mutwil, M., Obro, J., Willats, W.G., and Persson, S. (2008). GeneCAT - novel webtools that combine BLAST and co-expression analyses. Nucleic Acids Res. **36**:W320–W326.

Naithani, S., Preece, J., D'Eustachio, P., Gupta, P., Amarasinghe, V., Dharmawardhana, P.D., Wu, G., Fabregat, A., Elser, J.L., Weiser, J., et al. (2017). Plant Reactome: a resource for plant pathways and comparative analysis. Nucleic Acids Res. **45**:D1029–D1039.

Pérez-Rodríguez, P., Riano-Pachon, D.M., Corrêa, L.G.G., Rensing, S.A., Kersten, B., and Mueller-Roeber, B. (2010). PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. **38**:D822–D827.

Pfannschmidt, T., Blanvillain, R., Merendino, L., Courtois, F., Chevalier, F., Liebers, M., Grübler, B., Hommel, E., and Lerbs-Mache, S. (2015). Plastid RNA polymerases: orchestration of enzymes with different evolutionary origins controls chloroplast biogenesis during the plant life cycle. J. Exp. Bot. **66**:6957–6973.

Polturak, G., Heinig, U., Grossman, N., Battat, M., Leshkowitz, D., Malitsky, S., Rogachev, I., and Aharoni, A. (2018). Transcriptome and metabolic profiling provides insights into betalain biosynthesis and evolution in *Mirabilis jalapa*. Mol. Plant **11**:189–204.

Powers, D.M.W. (2011). Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. J. Mach. Learn Tech. **2**:37–63.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res. **33**:W116–W120.

Rai, A., Saito, K., and Yamazaki, M. (2017). Integrated omics analysis of specialized metabolism in medicinal plants. Plant J. **90**:764–787.

Rhee, S.Y., and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. Trends Plant Sci. **19**:212–221.

Satou, M., Enoki, H., Oikawa, A., Ohta, D., Saito, K., Hachiya, T., Sakakibara, H., Kusano, M., Fukushima, A., Saito, K., et al. (2014). Integrated analysis of transcriptome and metabolome of *Arabidopsis* albino or pale green mutants with disrupted nuclear-encoded chloroplast proteins. Plant Mol. Biol. **85**:411–428.

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol. **173**:2041–2059.

Scossa, F., Benina, M., Alseekh, S., Zhang, Y., and Fernie, A.R. (2018). The Integration of metabolomics and next-generation sequencing data to elucidate the pathways of natural product metabolism in medicinal plants. Planta Med. **84**:855–873.

Sheth, B.P., and Thaker, V.S. (2014). Plant systems biology: insights, advances and challenges. Planta **240**:33–54.

Sun, G., Xu, V.Y., Liu, H., Sun, T., Zhang, J., Hettenhausen, C., Shen, G., Qi, J., Qin, Y., Li, J., et al. (2018). Large-scale gene losses underlie the genome evolution of parasitic plant *Cuscuta australis*. Nat. Commun. **9**:2683.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. **37**:914–939.

Tillich, M., and Krause, K. (2010). The ins and outs of editing and splicing of plastid RNAs: lessons from parasitic plants. New Biotechnol. **27**:256–266.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**:635–641.

Tzfadia, O., Bocobza, S., Defoort, J., Almekias-Siegl, E., Panda, S., Levy, M., Storme, V., Rombauts, S., Jaitin, D.A., Keren-Shaul, H., et al. (2018). The TranSeq 3′-end sequencing method for high-throughput transcriptomics and gene space refinement in plant genomes. Plant J. **96**:223–232.

The UniProt Consortium. (2018). UniProt: the universal protein knowledgebase. Nucleic Acids Res. **46**:2699.

Vanderschuren, H., Lentz, E., Zainuddin, I., and Gruissem, W. (2013). Proteomics of model and crop plant species: status, current limitations and strategic advances for crop improvement. J. Proteomics **93**:5–19.

Vogel, A., Schwacke, R., Denton, A.K., Usadel, B., Hollmann, J., Fischer, K., Bolger, A., Schmidt, M.H.W., Bolger, M.E., Gundlach, H., et al. (2018). Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. Nat. Commun. **9**:2515.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell **29**:944–959.

Yagi, Y., and Shiina, T. (2014). Recent advances in the study of chloroplast gene expression and its evolution. Front. Plant Sci. **5**:61.

Yan, L., Lai, X., Li, X., Wei, C., Tan, X., and Zhang, Y. (2015). Analyses of the complete genome and gene expression of chloroplast of sweet potato [*Ipomoea batatas*]. PLoS One **10**:e0124083.

Yu, Q.B., Huang, C., and Yang, Z.N. (2014). Nuclear-encoded factors associated with the chloroplast transcription machinery of higher plants. Front. Plant Sci. **5**:316.

Zhang, H., Lang, Z., and Zhu, J.K. (2018). Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. **19**:489–506.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P., Banf, M., Dai, X., Martin, G.B., Giovannoni, J.J., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol. Plant **9**:1667–1670.