



Development of a statistical framework for mass spectrometry data analysis in untargeted Metabolomics studies

Dissertation

for the award of the degree

„Doctor rerum naturalium“

of the Georg-August-Universität Göttingen

within the doctoral program

„Biomolecules: Structure – Function – Dynamics“

of the Georg-August-University School of Science (GAUSS)

submitted by

Alexander Kaever

from

Hannover, Germany

Göttingen 2014

Members of the Thesis Committee

Prof. Dr. Burkhard Morgenstern (1st Reviewer)

Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-Universität
Göttingen

Prof. Dr. Ivo Feussner (2nd Reviewer)

Department of Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-
August-Universität Göttingen

Prof. Dr. Helmut Grubmüller

Department Theoretical and Computational Biophysics, Max Planck Institute for Biophysical
Chemistry

Date of the oral examination: June 6th, 2014

Affidavit

I hereby confirm that this thesis has been written independently and with no other sources and aids than quoted.

Göttingen, November 2014

Alexander Kaever

Contents

1	Abstract	1
2	Publication List	3
3	Introduction	5
3.1	Mass spectrometry-based Metabolomics	5
3.2	Statistical and exploratory data analysis	11
3.3	Functional annotation and integration of other omics platforms	15
3.4	The wound response of Arabidopsis thaliana: A case study experiment	17
3.5	Objectives and overview	19
4	MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data	21
5	Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets	31
6	MarVis-Pathway: Integrative and Exploratory Pathway Analysis of Non-Targeted Metabolomics Data	45
7	Applications of the MarVis-Suite	61
8	Discussion	63
8.1	Adduct correction and annotation of LC/MS data	64
8.2	Statistical ranking and filtering of intensity profiles	67
8.3	Combination of multi-omics data sets, clustering, and visualization	70
8.4	Pathway annotation, visualization, and enrichment analysis	71
8.5	Meta-analysis of pathway enrichment	74
8.6	Workflow and platforms	77

9	References	79
10	Acknowledgments	97
11	Supplementary Material	99
11.1	MarVis-Suite handbook	99
12	Curriculum Vitae	173

Abstract

A central objective in the analysis of mass spectrometry-based untargeted Metabolomics data is the detection of intensity patterns that differ between experimental conditions and the identification of underlying metabolites and biochemical processes. In this context, the identification of metabolites is a major bottleneck and needs to be guided by expert knowledge and tools for explorative data analysis. The integration of data sets from other omics platforms, e.g. DNA microarray-based Transcriptomics, can thereby provide valuable hints and support the reconstruction of related metabolic pathways, which then form the biochemical context for metabolite identification. In this work, a statistical framework and user interfaces for exploratory evaluation of mass spectrometry-based non-targeted Metabolomics data in combination with data sets from other omics platforms are introduced. The developed methods and tools were combined in the highly interactive MarVis-Suite software. The MarVis-Filter interface includes functions for the adduct and isotope correction of mass spectrometry data, molecular formula prediction, statistical ranking, filtering, and combination of cross-omics data sets. Within MarVis-Cluster, intensity profiles associated with ion species or microarray spots (features) in filtered and combined data sets can be clustered, visualized, interactively inspected and labeled. By means of MarVis-Pathway, data set features may be annotated in the context of organism-specific metabolic pathways. For statistical analysis, which forms a counterweight to the highly interactive and selective MarVis workflow, an extensive framework for meta-analysis of multi-omics data sets based on pathway enrichment analysis was developed. The methods and tools were successfully applied to several liquid chromatography/mass spectrometry data sets in combination with DNA microarray data in the context of plant wounding. The integration of Transcriptomics data thereby significantly supported the analysis and interpretation of non-targeted Metabolomics data sets.

Publication List

- **A. Kaefer**, M. Landesfeind, K. Feussner, A. Mosblech, I. Heilmann, B. Morgenstern, I. Feussner, and P. Meinicke. MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics*, 2014. doi: 10.1007/s11306-014-0734-y
- **A. Kaefer**, M. Landesfeind, K. Feussner, B. Morgenstern, I. Feussner, and P. Meinicke. Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets. *PLoS ONE*, 9(2):e89297, 2014
- **A. Kaefer**, M. Landesfeind, M. Possienke, K. Feussner, I. Feussner, and P. Meinicke. MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data. *Journal of Biomedicine and Biotechnology*, 2012:263910, 2012
- **A. Kaefer**, M. Landesfeind, K. Feussner, I. Feussner, and P. Meinicke. Metabolite clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. In *The Handbook of Plant Metabolomics*, pages 273-287. Wiley-Blackwell, 2013
- **A. Kaefer**, T. Lingner, K. Feussner, C. Göbel, I. Feussner, and P. Meinicke. MarVis: a Tool for Clustering and Visualization of Metabolic Biomarkers. *BMC Bioinformatics*, 10:92, 2009
- M. Landesfeind, **A. Kaefer**, K. Feussner, C. Thurow, C. Gatz, I. Feussner, and P. Meinicke. Integrative study of Arabidopsis thaliana metabolomic and transcriptomic data with the interactive MarVis-Graph software. *PeerJ*, 2(e239), 2014
- S. König, K. Feussner, **A. Kaefer**, M. Landesfeind, C. Thurow, P. Karlovsky, C. Gatz, A. Polle, and I. Feussner. Soluble phenylpropanoids are involved in the defense response of Arabidopsis against Verticillium longisporum. *New Phytologist*, 202(3):823-837, 2014

- J. Gamir, V. Pastor, **A. Kaefer**, M. Cerezo, and V. Flors. Targeting novel chemical and constitutive primed metabolites against *Plectosphaerella cucumerina*. *The Plant Journal*, 78(2):227-240, 2014
- V.-T. Tran, S. A. Braus-Stromeyer, H. Kusch, M. Reusche, **A. Kaefer**, A. Kühn, O. Valerius, M. Landesfeind, K. Abhauer, M. Tech, K. Hoff, T. Pena-Centeno, M. Stanke, V. Lipka, and G. H. Braus. Verticillium transcription activator of adhesion vta2 suppresses microsclerotia formation and is required for systemic infection of plant roots. *New Phytologist*, 202(2):565-581, 2014
- S. König, K. Feussner, M. Schwarz, **A. Kaefer**, T. Iven, M. Landesfeind, P. Ternes, P. Karlovsky, V. Lipka, and I. Feussner. Arabidopsis mutants of sphingolipid fatty acid α -hydroxylases accumulate ceramides and salicylates. *New Phytologist*, 196(4):1086-1097, 2012
- K. Nahlik, M. Dumkow, Ö. Bayram, K. Helmstaedt, S. Busch, O. Valerius, J. Gerke, M. Hoppert, E. Schwier, L. Opitz, M. Westermann, S. Grond, K. Feussner, C. Göbel, **A. Kaefer**, P. Meinicke, I. Feussner, and G. H. Braus. The COP9 signalosome mediates transcriptional and metabolic response for hormones, oxidative stress protection and cell wall rearrangement during fungal development. *Molecular Microbiology*, 78:964-979, 2010
- P. Meinicke, T. Lingner, **A. Kaefer**, K. Feussner, C. Göbel, I. Feussner, P. Karlovsky, and B. Morgenstern. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology*, 3:9, 2008

Introduction

3.1 Mass spectrometry-based Metabolomics

In recent years, Metabolomics, the study of all small molecules (metabolites) representing educts, intermediates, and products of metabolism, has become a key methodology to investigate an organism's reaction under different experimental conditions, such as environmental or genetic perturbations [1, 2, 3]. In contrast to the more established fields Transcriptomics and Proteomics, which focus on the analysis of the whole set of transcripts (transcriptome) and proteins (proteome) in organisms, Metabolomics allows the characterization of the biochemical properties of a cell in terms of metabolite abundance measurements, e.g. as result of modifications on the genetic level. Therefore, it was named the link between genotypes and phenotypes [4]. Metabolomics comprises two different approaches: targeted and non-targeted studies. Targeted Metabolomics [5] focuses on the quantification of a set of known metabolites. Non-targeted or untargeted studies aim to identify and characterize unknown or so far not described metabolites for particular experimental conditions [6]. In this context, the terms metabolic fingerprinting and profiling are often used to indicate the type of analysis [4, 7]. A metabolic fingerprint thereby represents all measurements obtained for a particular sample and all detected but not yet identified metabolites. These fingerprints can then be used to discriminate different groups of samples and extract measurements that are responsible for this distinction. In a final step, the underlying metabolites may be identified. In metabolic profiling studies, classes of known compounds, e.g. belonging to the same metabolic pathway, are analyzed and their abundance is compared between different samples.

Before analyzing metabolic samples, the metabolites have to be extracted from the biological material [8, 9, 10] (see figure 3.1, first part). This step includes the quenching of metabolic reactions, e.g. by means of rapid freezing in liquid nitrogen, and the removal of macromolecules, such as proteins, which would interfere with the detection of the small-weight metabolites. For targeted analysis, the extraction method is usually highly optimized and tailored to the target

compounds [11]. In order to extract as many metabolites as possible for non-targeted analysis, different classes of metabolites are often extracted separately. For example, by means of a two-phase extraction with methanol, chloroform and water [12], non-polar metabolites are extracted in the chloroform and polar metabolites in the methanol phase. In order to compare different genotypes, environmental perturbations, or developmental stages, multiple independent samples are prepared for different experimental conditions (see figure 3.1 and section 3.2). In many applications, an experiment comprises more than two conditions, e.g. when performing studies on time series [13].

For the analysis of metabolic samples, nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) techniques are widely used [14, 15]. It is important to state that currently no single platform can measure all classes of metabolites occurring in typical metabolic samples. NMR spectroscopy allows the non-destructive analysis of metabolites and also provides structural information [16]. However, NMR requires large amounts of analytes and is therefore not applicable for the detection of low-concentration metabolites, such as plant hormones. In contrast, MS analysis features a higher sensitivity and allows to detect and quantify hundreds of metabolites in small sample volumes. In this type of analysis, the molecules in a sample are ionized, separated by means of the mass-to-charge (m/z) ratios of corresponding ions, and finally detected and quantified. MS analysis can be performed in positive and negative ionization mode, which result in positively or negatively charged ions, respectively. During ionization, different analytes can interact and particular ions may be suppressed [17], which hinders the detection or may result in an incorrect quantification. Therefore, the analytes are usually separated before MS analysis using chromatographic techniques.

The most common combinations here are gas chromatography/mass spectrometry (GC/MS) [18] and liquid chromatography/mass spectrometry (LC/MS) [19, 20]. GC/MS allows the highly reproducible separation of volatile metabolites or less volatile compounds in combination with derivatization (chemical adding of a volatile functional group). The gaseous analytes together with a carrier gas (mobile phase) are transported through a column and chronologically separated by means of the interaction with the column material (stationary phase). GC/MS was used in the earliest studies in Metabolomics [1].

In LC analysis, the analytes are dissolved in solvents (mobile phase) and then led through a column. Depending on the interaction with the stationary phase and the solvent system, different compounds take more or less time to travel the column and a chronological separation is achieved. LC/MS is suitable for the analysis of most metabolites that can be dissolved and does not require derivatization of non-volatile compounds. However, the chromatographic separation is often less reproducible compared to GC and shifts in retention time (rt), the time a particular compound takes to traverse the column, are observed between different analyses.

The t_r depends for example on the size of particles in the column material, the column length, the solvent system, and the polarity of analytes. In order to allow high-throughput LC/MS analysis, ultra (high) performance liquid chromatography (UPLC or UHPLC) systems, which use columns packed with very small particles in combination with high pressure, have been developed [20].

In most applications, (UP)LC is combined with electrospray ionization (ESI) and time-of-flight (TOF) analysis in the mass spectrometer [20]. Details on other ionization and detection technologies applied in LC/MS analysis can be found in [19, 20]. ESI is a soft ionization technique that results in rather small amounts of fragmentations of the analytes. The compounds in the liquid sample are ionized by means of high voltage and form drops, which dissolve into smaller droplets until single ions are isolated. Once these ions reach the TOF mass analyzer, they are accelerated in an electric field and the time until they reach a detector plate is measured. This time is proportional to the square root of the m/z ratio of each ion. Finally, the ions that reach the detector are recorded. The abundance of ions (intensity) is either measured as ion counts or electric current. This information can be combined in the so-called total ion chromatogram (TIC), which summarizes the ion intensities over all m/z values along the t_r axis (see figure 3.2 and 3.1). For each t_r , an MS spectrum is recorded (see lower plot in figure 3.2).

After soft ionization by ESI-MS, a particular metabolite is often detected as protonated (positive ionization mode) or deprotonated (negative mode) ion. However, a single metabolite species is often represented by multiple ions with different m/z ratios. These related ions may represent isotopologues, molecules having the same structure but containing different numbers of isotopes. Also the aggregation of multiple target molecules or multiply charged ions, which are not so common for ESI-MS of small metabolites, are possible. Additionally, a molecule may be fragmented during ionization resulting in the loss of mass. In case the fragment is not subsequently ionized and can therefore not be detected by MS, this phenomenon is called neutral loss. Especially for ESI-based (UP)LC/MS analysis, the formation of adducts is often observed [20, 21]. These adducts result from the addition of another molecule to the target during ionization, e.g. the addition of formate in negative or ammonium and sodium in positive mode (see lower plot in figure 3.2). Since isotopologues interact in the same way with the stationary phase of the LC column and the other described alterations occur in the MS ion source, all ions representing the same metabolite species are detected at about the same t_r .

In order to compare different samples or conditions, the results from UPLC TOF-MS analysis of each sample (see figure 3.2) have to be combined [7] (see figure 3.1). This data preprocessing includes the detection and integration of intensity peaks and the chromatographic alignment of sample-specific peak lists [22]. Because of limited chromatographic precision,

ions belonging to a particular metabolite species are detected in an *rt* range around a maximum intensity (see figure 3.2). The average width of such a peak in UPLC TOF-MS analysis is about 10 seconds (depending on the actual platform). In comparison, a typical UPLC run takes only a few minutes. For each sample, peaks are detected in the *rt*-*m/z* spectrum and the corresponding intensities are either integrated/summed or the maximum value (peak height) is taken. The sample-specific peak lists, which contain the *rt*s, *m/z* ratios, and integrated intensities for all detected peaks, have then to be aligned and corrected for *rt* shifts, which are very common for the less reproducible (UP)LC analysis compared to GC. Finally, the aligned peaks can be stored in a feature matrix of intensity profiles (see figure 3.1). Each ion peak feature (column in the feature matrix) is associated with an *rt* and *m/z* value, averaged over all aligned samples, and a profile containing the intensities for all samples [12]. These intensities can be used to compare the relative abundance of the corresponding metabolite species between different samples and conditions. Ion features representing the same metabolite species, e.g. adducts or isotopologues, usually show a very similar intensity profile. The ion features are also referred to as markers [7] or marker candidates [12]. A typical UPLC TOF-MS data set contains a few thousand features, depending on the preprocessing and additional filters, e.g. when considering only peaks above a predefined intensity threshold.

For the described data preprocessing, machine and vendor-specific software platforms, such as the MarkerLynx Application Manager for the MassLynx software (Waters Corporation) and the Mass Hunter Workstation in combination with the Mass Profiler Professional software (Agilent Technologies Corporation), are available. Additionally, open software packages, such as the popular XCMS platform [23], have been developed for this purpose.

The workflow of data acquisition (see figure 3.1) has to be repeated for different extraction phases of metabolites. Additionally, the samples of each extraction phase are analyzed in positive and negative ionization mode of the mass spectrometer. This repeated analysis results in multiple data matrices, e.g. four data sets for the analysis of the polar and non-polar extraction phase in positive and negative mode, respectively. Since some metabolite species can be detected in positive and negative ionization mode, represented by different ion types, and can even occur in the polar as well as the non-polar extraction phase, the resulting data sets are not independent. Because the corresponding metabolites are detected as different ion species, e.g. representing unknown adducts [24], and *rt* shifts occur, the related features cannot be easily merged across data sets.

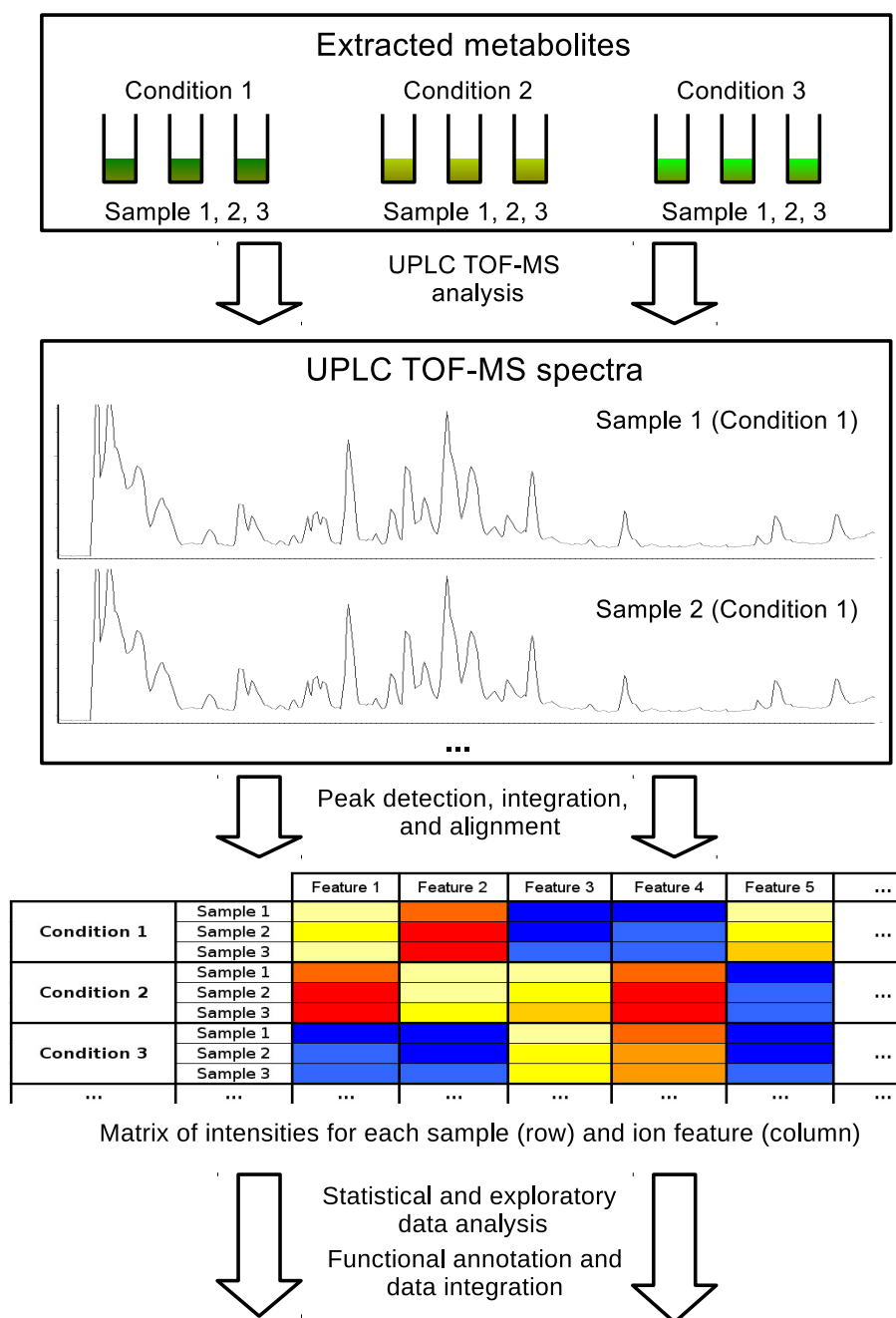


Figure 3.1: Workflow for UPLC TOF-MS analysis of multiple samples, associated with different experimental conditions, and data processing (chromatogram provided by Dr. Kirstin Feussner, feature matrix adapted from [12]). Each cell in the feature matrix represents the integrated intensity for the corresponding ion feature (column) and sample (row). The intensities are color-coded, e.g. red color represents high and blue color low intensities.

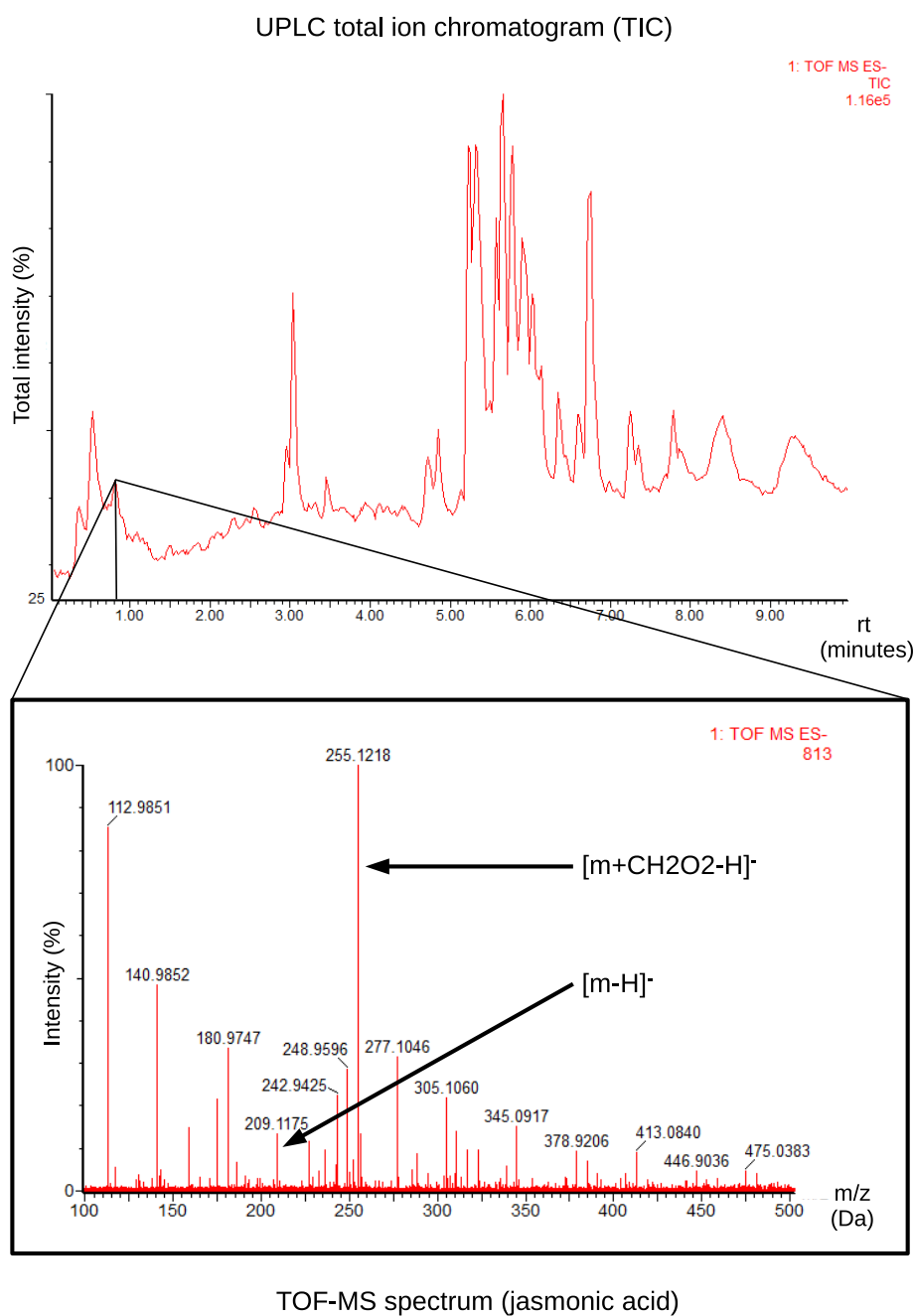


Figure 3.2: Example TIC and MS spectrum of jasmonic acid for an UPLC TOF-MS analysis of a single sample (non-polar extraction phase in negative ionization mode, chromatogram and spectrum provided by Dr. Kirstin Feussner). The total intensity for a particular rt in the TIC (upper plot) summarizes the underlying MS spectrum. The lower plot shows the corresponding spectrum for jasmonic acid (rt 0.79 minutes, monoisotopic mass 210.1256 Dalton). Two ionization products, the deprotonated molecule $[m - H]^-$ and the formate adduct $[m + CH_2O_2 - H]^-$, are marked by arrows. Notably, the formate adduct shows a much higher intensity compared to the deprotonated jasmonate.

3.2 Statistical and exploratory data analysis

In most applications, data set features associated with differential intensity profiles are of high interest. These profiles show large differences between the experimental conditions in comparison to small variations within each condition. In order to detect such candidates, statistical ranking and filtering methods are applied [25, 26]. The corresponding significance tests can be divided into parametric methods, which are based on assumptions on the intensity distribution, e.g. the normal or log-normal distribution, and non-parametric methods, which utilize intensity ranks. The parametric t-test and the analysis of variance (ANOVA) in case of more than two conditions or the rank-based Mann-Whitney and Kruskal-Wallis test are popular methods in this context. In the context of DNA microarray analysis, also non-parametric tests based on the random permutation of sample labels (assignments of samples to conditions), such as the significance analysis of microarrays (SAM) [27], are often applied.

A p-value for a particular feature thereby represents the probability of obtaining an equally or more differential profile for the corresponding feature by chance. The actual p-values strongly depend on the assumptions associated with the null hypothesis of the test (that there is no difference between the conditions), e.g. normal or log-normal distributed intensities in case of ANOVA. The p-values can be used to rank the data set features and filter them by means of an error threshold after adjustment for multiple testing [28]. The p-values should be adjusted because low p-values are expected to be observed when testing a large number of features in parallel. For example, if 10,000 independent tests with true null hypotheses (no differences between the conditions) are performed, one feature with a p-value of 0.0001 is expected to be found by chance. Different methods have been developed for the adjustment in a multiple testing scenario. The conservative Holm-Bonferroni method [29] controls the familywise error rate (FWER), which represents the probability of falsely rejecting a true null hypothesis (declaring one or more features significant although they do not represent real differences between the conditions). The less conservative Benjamini-Hochberg procedure [30] controls the false discovery rate (FDR), which represents the expected rate of false discoveries (features with rejected but true null hypothesis). The FWER or FDR can be estimated for each feature and used to filter a data set, e.g. by means of an error threshold of 0.01 or 0.05. For random permutation tests, the error rates can be directly estimated based on the observed feature-specific test statistics and the corresponding values obtained in the permutations [27].

Despite filtering, a typical data set still contains hundreds of features, which are associated with complex multivariate intensity profiles. In order to identify interesting intensity patterns,

unsupervised data mining approaches, such as principal component analysis (PCA), clustering algorithms, and self-organizing maps are often employed [1, 7, 12, 26].

PCA and related methods are applied for dimensionality reduction and visualization of the samples in a data set. The samples are initially represented as high-dimensional vectors containing the intensities for all features. By means of an eigenvalue decomposition of the estimated covariance matrix, the first k (typical two) orthogonal eigenvectors which represent most of the variance are extracted. The sample vectors (rows in the matrix in figure 3.1) can then be projected onto this low-dimensional coordinate plane and visualized, e.g. as 2D scatter plot (see figure 3.3). The coordinates after projection are called principal components (PCs) and the corresponding visualization the PCA score plot. This plot is often used for quality control. Samples of the same condition should cluster together, which indicates an overall higher variation between than within the conditions (see figure 3.3). However, PCA is based on a linear projection and complex non-linear relationships cannot be represented. Additionally, the first k (e.g. two) principal components may represent only a small fraction of the total variance in the data.

Similar to the score plot, the extracted eigenvectors, which contain the projection weights (loadings) for all data set features, can be visualized in a so-called loading plot (see figure 3.4). Here, the features and not the samples are represented as scatter points and features with high absolute loadings, which represent large parts of the captured variance, can be identified. In addition, clusters of features, which are associated with similar loadings, may be inspected.

A central disadvantage of the score and loading plot visualization is that the intensity profiles responsible for the separation of conditions and for high absolute loadings cannot be directly inspected [12]. In the loading plot in figure 3.4, for example, it is not clear what kind of intensity patterns are associated with particular features.

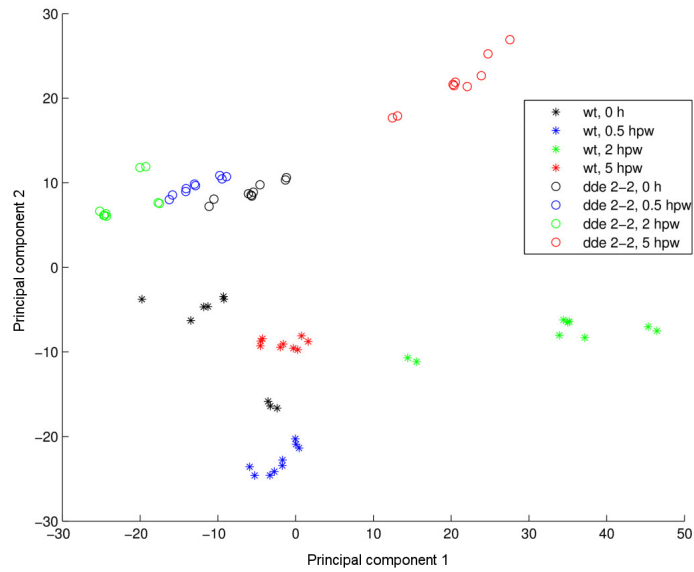


Figure 3.3: Example score plot of the first two principal components of 72 samples associated with eight conditions (adapted from [12]). The samples are represented as condition-specific colored markers and the plot shows a clear separation of the samples for most of the conditions (see [12] for details).

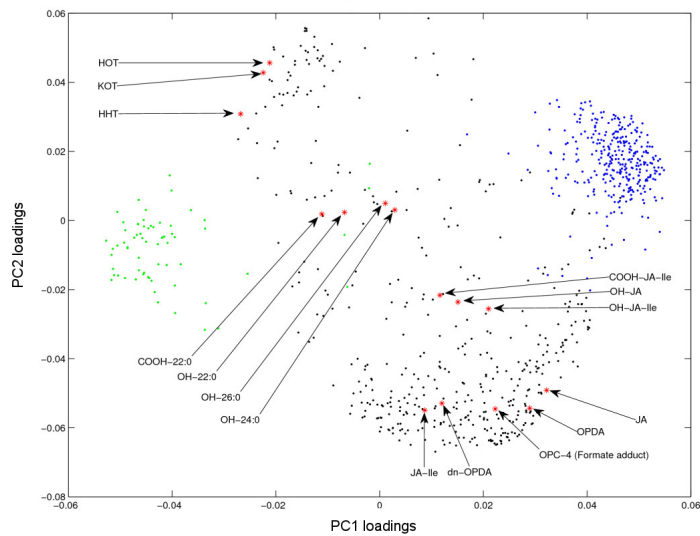


Figure 3.4: Example loading plot for the first two principal components of 837 ion features (adapted from [12]). Every feature is represented as scatter point. Features associated with identified metabolites are marked with arrows. Related features identified in cluster analysis are marked in green and blue color (see [12] for details).

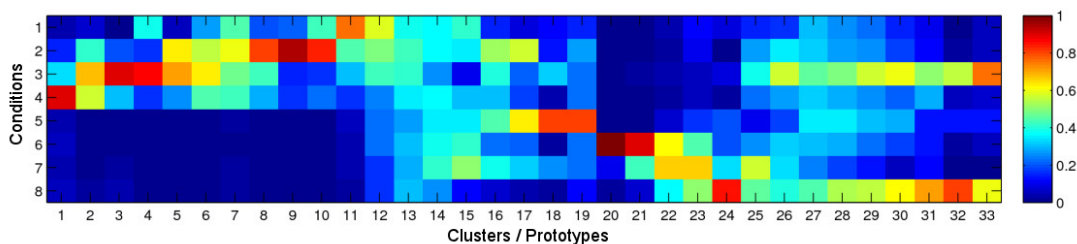


Figure 3.5: Example heatmap visualization of ordered prototype profiles after 1D-SOM clustering (adapted from [12]). Each row represents an experimental condition, each column a prototype/cluster. The cells show color-coded normalized intensities for each condition and prototype (see color mapping on the right-hand side). High relative intensities are represented by red and low intensities by blue color.

For the direct visualization and inspection of intensity patterns, the clustering and visualization by means of one-dimensional self-organizing maps (1D-SOMs) was introduced [12] and has proven to be highly valuable for exploratory data analysis in MS-based non-targeted Metabolomics studies. For each feature, the intensities per condition are averaged and the resulting vector is normalized to unit Euclidean length (normalized profile). Then, a one-dimensional array of k connected prototype vectors (the 1D-SOM) is fitted into the space of normalized profiles and each feature is associated with a representative prototype. After this training procedure, each prototype represents a cluster of features with similar profiles and the average profiles per cluster (prototype profiles) can be visualized as heatmap (see figure 3.5). Due to the linear order of the 1D-SOM, the prototype profiles in the heatmap visualization are sorted according to similarity. This allows a convenient overview on prominent intensity patterns in the data set and the identification of blocks of neighboring clusters that share a similar profile. Because of the aggregation of sample intensities per condition (averaging), the patterns can be directly interpreted in the experimental context.

Compared to the classical approach of hierarchical clustering [31] combined with the popular k-means algorithm [32, 33], the 1D-SOM training results in a more robust clustering and visualization of feature profiles [12]. The 1D approach allows the direct heatmap visualization of all experimental conditions (y-axis in figure 3.5) along the linear array of prototypes (x-axis in figure 3.5). In case of applications of the classical 2D-SOMs [34, 35], the multivariate prototype profiles cannot be directly integrated in a single heatmap since the x-axis and y-axis are used for the visualization of the 2D grid of prototypes.

Besides the described unsupervised methods, also supervised methods for machine learning, which for example use the sample labels (assignments of samples to conditions) in order to train a classifier, are applied in this context [26]. The partial least squares discriminant analysis (PLS-DA) [36], support vector machines (SVMs) [37], and the random forests classifier [38,

39] are popular examples. After training of the corresponding model, feature weights or scores can be visualized and candidates which discriminate between the experimental conditions may be extracted [40]. However, this work focuses on unsupervised methods for clustering and visualization and especially the 1D-SOM approach.

The 1D-SOM method was implemented in the MarVis-Cluster (**Marker Visualization**) tool [41]. After 1D-SOM training, MarVis-Cluster facilitates the interactive selection and export of clusters of data set features which are associated with an interesting prototype profile. The 1D-SOM order thereby significantly supports this selection. Similar to the cluster analysis of gene expression patterns [31], the functional analysis of features associated with a particular intensity pattern, e.g. represented by single or blocks of clusters, is of high interest. Especially in case of non-targeted LC/MS data, the interactive interface of MarVis-Cluster facilitates the integration of the user's expert knowledge, e.g. for the identification of adducts and other related features found in the same cluster [12, 41].

3.3 Functional annotation and integration of other omics platforms

In order to identify metabolites associated with differential intensity profiles or interesting patterns in a particular experimental context, the ion features have to be annotated. For this annotation, the accurate m/z ratios are mapped to exact masses of known metabolites [21, 24]. The highly machine-specific and often poorly reproducible rt values from LC analysis are usually not used for this purpose. For mass-based mapping, the feature-specific m/z ratios have to be transformed into putative monoisotopic masses. This transformation includes the correction for the ion charge, e.g. when detected with double charge, included isotopes, adducts, and neutral mass loss (see section 3.1). For this purpose, common adduct and ionization rules, which describe the formation of frequently observed ion species in the form $[xm + y]^{z[+/-]}$, are applied in combination with isotope correction [42, 43, 44, 45]. These rules, e.g. $[m + H]^+$ for protonation, formally describe the building blocks of a particular ionization product. They include the number of combined target molecules (x) or charges (z) and the addition or subtraction of other molecules (y) and allow the calculation of the potential mass (m) of the corresponding metabolite (see lower plot in figure 3.2). For the prediction of ionization rules, isotope patterns, and accurate feature masses in LC/MS data, software tools, such as AStream [46] and CAMERA [47], have recently been developed.

A more sophisticated method, compared to the mass matching, is to predict molecular formulas based on the feature masses and match these with known metabolites [24, 42]. In order to reduce the number of possible formulas for a given tolerance, heuristic filtering based on chemical rules can be applied [48]. For the mass matching and formula prediction, a high mass precision, the machine-dependent accurateness of the measured m/z values, is of high importance. Even with high precision, e.g. 1 mDalton, the mapping is often ambiguous because of isomer metabolites, which share the same molecular formula but have a different chemical structure.

For this reason, the context in which a particular metabolite occurs is of high importance [42]. This includes basic information whether a compound has any biological relevance, has been described previously for the organism under study, or is part of a metabolic reaction or pathway, which is associated with specific genes. Especially organism-specific metabolic pathways, which are stored in public databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [49, 50] and BioCyc [51], provide a valuable context for metabolite annotation. Therefore, the mapping of corrected ion features to metabolic pathways was implemented as a tool [52]. The mapping of ion features to different metabolites in the same pathway thereby increases the confidence of the putative identification, especially when the corresponding profiles show the same intensity pattern.

Besides KEGG and BioCyc, other published databases provide context-specific metabolite data. AraCyc [53], which is part of the BioCyc collection, is an *Arabidopsis*-specific pathway database. The Human Metabolome Database (HMDB) [54] contains endogenous metabolites for human derived from literature and experiments. In the LIPID MAPS structure database [55], biologically relevant lipids are organized in different classes. The KNApSAcK database [56] provides metabolite entries associated with different species. A more general biochemical database is PubChem [57], which covers more than 700,000 compounds and associated biological test/screening results. In the context of GC/MS and MS/MS analysis, more specialized databases, which also include compound-specific MS fragmentation spectra, are available [58, 59, 60, 61]. In recent years, several web-based platforms [62, 63, 64, 65] have been developed, which allow the online preprocessing, statistical and exploratory analysis, and functional annotation of MS-derived Metabolomics data in the context of the described databases.

The introduced methods for analysis of non-targeted LC/MS data imply a couple of challenges. The mass-based mapping of ion features to metabolites and associated pathways allows the fast annotation of large data sets and the generation of working hypotheses. However, the mappings based on exact masses are only putative and error-prone. First, the predicted feature masses may not be correct because of errors in the adduct and isotope detection. Second, the

mapping to a particular metabolite can be erroneous because of isomers or similar compound masses within the machine-specific mass precision. Additionally, many metabolites are not yet represented in public databases [3]. Another challenge is that Metabolomics data sets which are used for first hypothesis generation often comprise only a small number of independent biological replicates.

The integration of data sets from other omics platforms which complement the Metabolomics approach, such as DNA microarray and RNA-seq based Transcriptomics [66, 67] and MS-based Proteomics [68], is a promising direction to cope with these challenges [69, 70]. A comprehensive data analysis should thereby take advantage of all available omics data sets for a particular experimental context [71, 72]. In comparison to the relatively young Metabolomics discipline, DNA microarray based Transcriptomics and corresponding methods for data analysis are much more established. This includes statistical tests for differentially expressed genes [73, 74, 75] and the cluster analysis of corresponding expression patterns [76, 77, 78].

For the knowledge-based functional analysis of sets of related genes, the overrepresentation [79, 80] and gene set enrichment analysis [81] are popular methods [82, 83]. The objective of this type of data analysis is to identify gene sets, e.g. genes associated with a particular pathway or gene ontology [84] term, which show an overrepresentation or enrichment of genes differentially expressed under particular experimental conditions. The methodology was also transferred to Metabolomics [45, 85, 86] by analyzing sets of metabolites, e.g. metabolic pathways. For integrative analysis, pathways, which summarize biochemical reactions and associated enzymes, genes, and metabolites, represent a convenient link between data sets from different omics platforms [62, 87].

In order to integrate results from multiple independent studies in the same experimental context, methods in the field of statistical meta-analysis [88, 89] have been developed. Meta-analysis has been applied to independent DNA microarray studies in order to extract genes which are differentially expressed considering multiple data sets [90, 91, 92, 93]. For the combination of independent pathway-specific p-values from enrichment analysis of microarray data, a framework was introduced in [94]. However, in the context of MS-based Metabolomics, the dependence of data sets plays an important role (see section 3.1).

3.4 The wound response of *Arabidopsis thaliana*: A case study experiment

The introduced methods and tools were evaluated and successfully applied to cross-omics data in the context of wounding of *Arabidopsis thaliana* (see the following chapters). The wound

response, which is part of the plant's defense against insects, has been studied by means of Transcriptomics as well as Proteomics experiments [95, 96, 97]. It is well known that the response is mainly regulated by the isoleucin conjugate of jasmonic acid (JA-Ile) [98, 99, 100]. Metabolic pathways which describe the biosynthesis of jasmonic acid are available in KEGG (alpha-linolenic acid metabolism) and AraCyc (jasmonic acid biosynthesis). These pathways also contain important precursor metabolites, such as 12-oxo-10,15-phytodienoic acid (12-OPDA) and 3-Oxo-2-(pent-2'-enyl)-cyclopentane-1-octanoic acid (OPC-8:0), and show a good coverage of genes coding for enzymes in the JA-Ile biosynthesis. For these reasons, the context of plant wounding is used as a model system for the evaluation of methods for LC/MS-based Metabolomics [12].

The data sets, which were derived from UPLC TOF-MS Metabolomics and DNA microarray Transcriptomics analysis, comprise conditions for wild type (wt) and the jasmonate-deficient *dde2-2* mutant plants [101] harvested at different time points after wounding. Table 3.1 gives an overview on the data sets ordered according to two scenarios of data analysis. In scenario A (see chapter 4 and 5), the available samples were analyzed only by UPLC TOF-MS and independent DNA microarray data sets [102] obtained from the ArrayExpress [103] repository were integrated (chapter 5). In scenario B (see chapter 6), the Metabolomics (UPLC TOF-MS analysis) and Transcriptomics data (DNA microarray analysis) were derived from the same biological samples and a more integrated data analysis, e.g. by means of the clustering of cross-omics feature profiles, was possible.

The Metabolomics experiments and preprocessing of raw UPLC TOF-MS data (peak detection and alignment) were performed in the Department of Plant Biochemistry¹ by Dr. Kirstin Feussner and coworkers. The DNA microarray data sets used in chapter 5 were downloaded from the ArrayExpress website and preprocessed (preparation of the data matrix) by Dr. Corinna Thurow² and Manuel Landesfeind³. The Transcriptomics experiment described in chapter 6 was designed by Prof. Dr. Ivo Feussner, Prof. Dr. Ingo Heilmann, Dr. Kirstin Feussner, and Dr. Alina Mosblech (Department of Plant Biochemistry) and conducted by Dr. Alina Mosblech (including RNA preparation). Microarray analysis and data preprocessing (including quantile-normalization) were performed by Dr. Lennart Opitz and Dr. Gabriela Salinas-Riester⁴.

¹Department of Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-August-University Göttingen

²Department for Plant Molecular Biology and Physiology, Schwann-Schleiden-Research-Center for Molecular Cell Biology, Georg-August-University Göttingen

³Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University Göttingen

⁴DNA Microarray and Deep-Sequencing Facility, Department of Developmental Biochemistry, Georg-August-University Göttingen

Label	Times	Platform	Extraction	Ionization	Reference
A_M1	0.5h 2h 5h	UPLC TOF-MS	non-polar	negative	chapter 4 and 5
A_M2	0.5h 2h 5h	UPLC TOF-MS	non-polar	positive	chapter 4 and 5
A_T1	1h	DNA microarray	-	-	[102], E-ATMX-9
A_T2	3h	DNA microarray	-	-	E-MEXP-1475
B_M1	0.5h 2h	UPLC TOF-MS	non-polar	negative	chapter 6
B_M2	0.5h 2h	UPLC TOF-MS	non-polar	positive	chapter 6
B_M3	0.5h 2h	UPLC TOF-MS	polar	negative	chapter 6
B_M4	0.5h 2h	UPLC TOF-MS	polar	positive	chapter 6
B_T1	0.5h 2h	DNA microarray	-	-	chapter 6

Table 3.1: Overview on data sets in the cross-omics case study used for evaluation and application of the developed methods and tools. The second column (Times) shows the time points when the wounded plants (wt and mutant) were harvested. All data sets also contain a condition of control samples for unwounded wt and *dde2-2* plants. The columns Extraction and Ionization indicate the extraction phase and ionization mode for the Metabolomics data sets. The A_T1 and A_T2 data sets can be obtained from the ArrayExpress website (see IDs in the last column).

3.5 Objectives and overview

The objective of this work is the development of a statistical framework for the described workflow in non-targeted Metabolomics studies and the integration of other omics platforms. The main focus lies on the statistical analysis of intensity profiles and the pathway enrichment and meta-analysis of multiple independent or dependent data sets from UPLC TOF-MS analyses. The already successfully applied methods for exploratory data analysis and the interactive MarVis-Cluster interface should be integrated in the new framework. Ideally, the framework should extend the user-driven exploratory data analysis by functional annotations and be a counterweight and statistical control of the highly interactive and selective MarVis-Cluster-based analysis.

In the following publications, the extension of the MarVis-Cluster tool to the powerful MarVis-Suite toolbox, the statistical framework, and selected applications are introduced. The first paper (chapter 4) describes the MarVis-Filter tool, which features the raw data import, ranking, filtering, adduct and isotope correction, and combination of processed data sets before analysis in MarVis-Cluster. The second publication (chapter 5) introduces the statistical framework based on the meta-analysis of pathway enrichment utilizing independent and dependent multi-omics data sets. Chapter 6 contains a publication describing the MarVis-Pathway tool, which is used for the reconstruction and statistical analysis of pathways for ranked, filtered, or

selected data set features, and further extensions of the MarVis-Suite and the statistical framework. Figure 1 in chapter 6 shows an overview on the workflow of data analysis within the new MarVis-Suite. Chapter 7 summarizes selected coauthor publications on the application of the MarVis-Suite. In chapter 8, the overall results of the publications are discussed. As supplementary material, chapter 11 contains the complete MarVis-Suite 2.0 handbook, which describes all methods, tools, and graphical user interfaces in detail.

MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data

The following paper was published 2012 in the *Journal of Biomedicine and Biotechnology* [104]. Supplementary material are available on <http://marvis.gobics.de>.

The sections on the biological interpretation of results from data analysis were written and the tables were created by Alexander Kaever, Dr. Kirstin Feussner, and Prof. Dr. Ivo Feussner in close collaboration. The algorithm for adduct and isotope correction includes concepts of an earlier version [45] and the MarVis-Filter tool includes corresponding redesigned prototype functions. The interfaces for prediction of molecular formulas from exact masses were implemented by Lars Söder (Department of Bioinformatics) under supervision of Alexander Kaever and Dr. Peter Meinicke (Department of Bioinformatics). The article was critically revised by all coauthors.

Methodology Report

MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data

Alexander Kaever,¹ Manuel Landesfeind,¹ Mareike Possienke,² Kirstin Feussner,^{2,3}
Ivo Feussner,² and Peter Meinicke¹

¹ Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University Göttingen,
37077 Göttingen, Germany

² Department for Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-August-University Göttingen,
37077 Göttingen, Germany

³ Department of Molecular Microbiology and Genetics, Institute of Microbiology and Genetics,
Georg-August-University Göttingen, 37077 Göttingen, Germany

Correspondence should be addressed to Alexander Kaever, alex@gobics.de

Received 28 July 2011; Revised 18 January 2012; Accepted 18 January 2012

Academic Editor: Brad Upham

Copyright © 2012 Alexander Kaever et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Statistical ranking, filtering, adduct detection, isotope correction, and molecular formula calculation are essential tasks in processing mass spectrometry data in metabolomics studies. In order to obtain high-quality data sets, a framework which incorporates all these methods is required. We present the MarVis-Filter software, which provides well-established and specialized methods for processing mass spectrometry data. For the task of ranking and filtering multivariate intensity profiles, MarVis-Filter provides the ANOVA and Kruskal-Wallis tests with adjustment for multiple hypothesis testing. Adduct and isotope correction are based on a novel algorithm which takes the similarity of intensity profiles into account and allows user-defined ionization rules. The molecular formula calculation utilizes the results of the adduct and isotope correction. For a comprehensive analysis, MarVis-Filter provides an interactive interface to combine data sets deriving from positive and negative ionization mode. The software is exemplarily applied in a metabolic case study, where octadecanoids could be identified as markers for wounding in plants.

1. Introduction

A central aim of untargeted Metabolomics and Metabonomics studies is the identification of marker metabolites which play a crucial role in the experimental context [1, 2]. Mass spectrometry combined with either gas chromatography (GC/MS) or liquid chromatography (LC/MS) has become a key technology for metabolome analysis under different experimental conditions [3, 4]. A typical data set after peak detection and sample alignment [5–7] consists of several thousand marker candidates which are characterized by a retention time (RT), a mass-to-charge value (m/z), and a multivariate intensity profile of abundance levels per condition, respectively [8]. The

experimental conditions are represented by replicate samples and may correspond to environmental disease or genetic perturbations [9–11]. In order to obtain a high-quality data set of experiment-related marker candidates, the raw data set is usually ranked and filtered using supervised machine learning techniques such as Random Forest classification [12, 13] or statistical analysis based on ANOVA or Kruskal-Wallis tests [14–16]. The filtered marker candidates are then annotated according to known metabolites from public biological and biomedical compound databases [17–21]. A central task of annotation is the calculation of actual molecular masses corresponding to each marker candidate by correcting the m/z ratios according to the ionization mode, potential adduct formation, and included natural

isotopes [22]. This problem can be addressed by applying the ionization rules $[xm + y]^{z[+/-]}$ [23], where x denotes the number of combined target molecules, y the mass of attached molecules (adduct formation), and z the degree of ionization (e.g., single or double). Additionally, the number of included isotopes has to be estimated in order to query databases which contain monoisotopic compound masses. Based on a potential ionization rule with parameters x , y , z and the number of included isotopes, the corresponding compound mass can be calculated.

For the corrected masses which cannot be assigned to particular compounds, the identification can be supported by calculating possible molecular formulas. The number of considered formulas can be significantly reduced by incorporating information from preprocessing as well as rules for heuristic filtering of molecular formulas [24], respectively. A major step in this process is the estimation of the number of included carbon atoms based on the intensity profiles of previously detected isotopologues.

There are a great number of software packages available, which provide tools for statistical analysis of multivariate experimental data [25, 26]. A number of tools for peak detection and sample alignment of mass spectrometry data, such as MetAlign or OpenMS, also support the deconvolution of isotopologues and statistical analysis [27, 28]. For the XCMS platform [7], a package for the annotation of LC/ESI-MS mass signals based on adduct rules has been implemented [23]. The calculation of possible ionization products and the rule-based heuristic filtering of molecular formulas is provided by several software packages [22, 24]. However, to the best of our knowledge, there is no software available which incorporates all of these methods in a single user-friendly tool as offered by MarVis-Filter.

2. Materials and Methods

In the following sections, the algorithm for adduct/isotope correction and the implementation of MarVis-Filter are described in detail.

2.1. Algorithm for Adduct and Isotope Correction. The algorithm is based on the input of the retention times, m/z ratios, and raw intensity profiles of all marker candidates in a data set and calculates as output the potential monoisotopic mass, ionization rule, and number of included ^{13}C -isotopes for every candidate. The approach is based on a greedy strategy which minimizes the number of potential molecular masses and simultaneously maximizes the similarity of intensity profiles between candidates with a similar retention time and actual mass. This concept follows the paradigm that in mass spectrometry analysis a metabolite is usually represented by several marker candidates with a similar retention time and intensity profile, but different m/z ratios according to the various possibilities of ionization and number of included isotopes. As parameters, the algorithm expects a list of ionization/adduct rules sorted according to their relevance, the assumed maximal number of ^{13}C -isotopes per marker candidate, a mass tolerance, an RT tolerance, and a minimal

cosine similarity of intensity profiles. The isotopologues correction is restricted to the detection of ^{13}C .

For storage of pairwise cosine similarities between candidate profiles, the algorithm utilizes a five-dimensional matrix M . Each entry $M_{(m,a_1,i_1,a_2,i_2)}$ corresponds to the maximal cosine similarity between the intensity profile of candidate m , assuming ionization rule a_1 and i_1 ^{13}C -isotopes, and another candidate, which has a similar retention time (within tolerance) and corrected mass (within tolerance) assuming ionization rule a_2 and i_2 ^{13}C -isotopes. For each candidate m , the algorithm then chooses the ionization rule and number of ^{13}C -isotopes which is supported by the highest sum of cosine similarities. In the following, the algorithm is described in detail.

- (1) Initialize M with zeros.
- (2) Calculate all possible masses by applying all ionization rules and number of ^{13}C -isotopes to all candidate m/z ratios.
- (3) Consider all pairs of potential masses under the following constraints and fill M with pairwise cosine similarities of corresponding candidate profiles.
 - (i) Consider only pairs of different marker candidates.
 - (ii) Consider only pairs within the mass and RT tolerance.
 - (iii) Consider only pairs with at least the requested cosine similarity.
 - (iv) Consider only pairs with different combinations of adduct rules and number of isotopes.
 - (v) For each entry in M hold only the maximum cosine similarity.
- (4) Calculate the reduced three-dimensional matrix M^{red} with summed entries:

$$M_{(m,a_1,i_1)}^{\text{red}} = \sum_{a_2,i_2} M_{(m,a_1,i_1,a_2,i_2)}. \quad (1)$$
- (5) Choose for each candidate m : the adduct rule and isotope number with the maximal sum of similarities $c_{\text{max}} = \max_{a_1,i_1} (M_{(m,a_1,i_1)}^{\text{red}})$. If $c_{\text{max}} = 0$, use the first ionization rule and zero ^{13}C -isotopes as default.
- (6) Calculate the masses according to chosen rules and isotope numbers.

In order to avoid apparently false associations between marker candidates, negative cosine similarities are disregarded. If for a given candidate different selections of the ionization rule and the number of isotopes maximize the sum of cosine similarities, the ionization rules with the highest relevance and the minimal number of ^{13}C -isotopes are selected.

Following the annotation of the ionization rules and ^{13}C -isotopes, the number of carbon atoms per candidate is estimated by comparing the raw intensities of marker candidates with zero predicted ^{13}C -isotopes (I_M) and the

respective marker candidates including one ^{13}C -isotope (I_{M+1}) according to the following formula:

$$n_C = \frac{98.9 I_{M+1}}{1.1 I_M}, \quad (2)$$

corresponding to the natural abundances of carbon isotopes. Given a pair of candidates, annotated as isotopologues (M and $M + 1$) and with the same ionization rule, a robust estimation of the number of carbon atoms is obtained by calculating the median n_C over all samples included in both intensity profiles.

2.2. Implementation. MarVis-Filter is implemented in the Matlab and C programming language and has been compiled together with the MarVis-Cluster tool [29] for Microsoft Windows XP/Vista/7. Execution of the software requires installation of the Matlab Compiler Runtime, which is provided with the software. The installation packages, the documentation, and example data sets can be downloaded from the project home page <http://marvis.gobics.de/>.

For data import and export MarVis-Filter uses the CSV (Comma Separated Values) file format, which can easily be processed by statistical analysis software and spreadsheet applications. MarVis-Filter also supports the direct import of aligned mass spectrometry samples from MarkerLynx Application Manager of MassLynx (Waters Corporation, Milford). For interactive analysis, ranking and filtering of multivariate intensity profiles MarVis-Filter provides the well-known one-way ANOVA and Kruskal-Wallis tests [14] combined with methods for P value adjustment for multiple-hypothesis testing [30, 31]. Based on customizable lists of ionization rules, the adduct/isotope correction can be performed on raw or filtered data sets. The ionization rules are imported as text files and can easily be adapted or extended.

Figure 1 shows the main window of MarVis-Filter after import and ranking. The “Ranking plot” (1) displays the adjusted P values (y -axis) of all candidate intensity profiles in the current data set sorted in ascending order. The data set can interactively be filtered according to a user-defined significance level by selecting a marker, sliding the red separator line or jumping to a predefined level. The “Profile plot” (2) shows the raw intensity profile of the currently selected marker candidate. Intensity values of replicated samples belonging to the same experimental conditions are marked in the same color. The “Marker information box” (3) displays information about all marker candidates of the data set arranged according to the P values and characterized by the m/z ratio, RT and additional user-defined scores, which can be imported along with the data set. After adduct and isotope correction, the additional annotations are displayed in this listbox as well. The “Data set clipboard listbox” (4) shows data sets which are currently held in the MarVis clipboard. The current (filtered or unfiltered) data set can simply be added or removed to/from this list. The data set clipboard supports an adduct and isotope correction of selected data sets in a batch mode. Data sets which were corrected based on different sets of ionization rules (e.g.,

positive and negative ionization) may be combined into one single data set.

For selected candidate profiles, bar plots, standard error plots, and boxplots can easily be inspected and exported in various image formats. For detailed analysis, the user can zoom into all plots. Additionally, MarVis-Filter provides a convenient interface for quick candidate search based on the ID, RT, m/z , or mass value.

MarVis-Filter also provides a molecular formula calculator, which is based on the Seven Golden Rules [24] and utilizes the estimated number of carbon atoms per marker candidate obtained after adduct and isotope correction.

MarVis-Filter and MarVis-Cluster [29] are combined in the MarVis-Suite which features the direct data exchange between preprocessing in MarVis-Filter and convenient visualization of multivariate intensity profiles and high-level cluster analysis in MarVis-Cluster.

3. Results and Discussion

The functionality of MarVis-Filter is demonstrated using two data sets of a metabolomic case study for plant wounding experiments [8]. The data sets are available on the project homepage <http://marvis.gobics.de/> together with a detailed description of the extraction and UPLC-TOF method. Additionally, the data sets are available for import in MarVis-Filter after installation of the MarVis-Suite (wound_neg_raw.csv and wound_pos_raw.csv in the examples directory).

3.1. Case Study and Data Sets. The case study reflects a wounding time course of *Arabidopsis thaliana* wild-type (WT) plants as well as of mutant plants (dde 2-2), which are deficient in the biosynthesis of the plant wound hormone jasmonic acid and its derivatives [32]. The wounding time course represents eight experimental conditions. The first four conditions reflect the metabolic situation within a wounding time course of wild-type (WT) plants, starting with the unwounded control plants (abbreviation wt_0) followed by the plants harvested 0.5 (wt_30), 2 (wt_2), and 5 hours past wounding (wt_5). The conditions 5 to 8 represent the analogous time course for the jasmonate deficient mutant plant dde 2-2 (aos_0, aos_30, aos_2, aos_5). Each condition contains nine replicate samples.

3.2. Data Import and Analysis in MarVis-Filter. The two data sets are imported sequentially in MarVis-Filter using the “Import raw CSV data” entry in the “File” menu with the following options: Delimiter: “;”; Start row: 5, Start column: 3; ID label: “id”; Generate IDs: activated; x column: 2; x label: “rt”; y column: 3; y label: “ m/z ”; Condition identifiers: “wt_0, wt_30, wt_2, wt_5, aos_0, aos_30, aos_2, aos_5”.

After data import, the marker candidates are sorted and ranked according to the P values of a Kruskal-Wallis test and the Bonferroni-Holm adjustment for multiple hypothesis testing [30] by selecting the corresponding checkboxes in the “Filter dialog” and the “Adjustment for multiple testing” dialog.

Adduct and isotope correction are performed on the full data sets separately using predefined sets of adduct

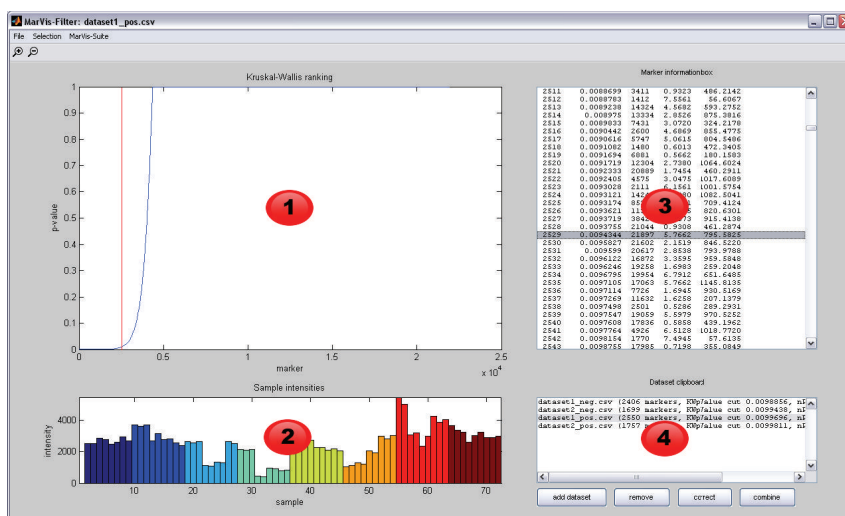


FIGURE 1: The main window of MarVis-Filter after data import and ranking. The “Ranking plot” (1) displays the adjusted P values (y -axis) of all candidate intensity profiles. The “Profile plot” (2) shows the raw intensity profile of the currently selected marker candidate. The “Marker listbox” (3) displays information about the current marker candidate. The “Data set clipboard listbox” (4) shows data sets which are currently held in the MarVis clipboard.

TABLE 1: Overview on data sets from the metabolomic case study for plant wounding experiments. The columns “Candidates” and “Filtered candidates” contain the number of marker candidates in the raw data set and the number of significant candidates in the filtered data set, respectively.

Data set	Ionization	Candidates	Filtered candidates	Samples per condition
1	Negative	24796	1719	9
2	Positive	23325	1785	9

rules for the negative (Table 2) and positive ionization mode (Table 3), an RT tolerance of 0.04 minutes, a mass tolerance of 0.005 Da, a minimal cosine similarity of 0.75, and a maximum number of two ^{13}C -isotopes per candidate. The adduct rules had been determined in previous targeted UPLC-TOF-MS experiments. After correction, the data sets are filtered according to a significance level for adjusted P values of 0.01 (“Goto level” entry in “Selection” menu) and added to the MarVis data set clipboard. Table 1 shows the initial number of imported marker candidates and the number of high-quality marker candidates after filtering. Finally, the two data sets in the MarVis clipboard are concatenated using the “combine” button. The combined data set can be sorted according to a user-defined method once again and is then presented in a new MarVis-Filter window. After selecting the whole data set, the combined subset of 3504 high-quality marker candidates can be exported as a CSV file, and clustered as well as visualized using MarVis-Cluster (“Goto MarVis-Cluster” entry in the “MarVis-Suite” menu). Figure 2 shows the results from clustering of the filtered and combined data in MarVis-Cluster.

TABLE 2: List of adduct rules for correction of data measured in negative ionization mode.

Rule	Description	Rule
1	Deprotonation	$[m - H]^-$
2	Formate adduct	$[m + \text{CH}_2\text{O}_2 - H]^-$
3	Formate adduct with sodium	$[m + \text{CH}_2\text{O}_2 - 2H + \text{Na}]^-$

TABLE 3: List of adduct rules for correction of data measured in positive ionization mode.

Rule	Description	Rule
1	Protonation	$[m + H]^+$
2	Ammonium adduct	$[m + \text{NH}_4]^+$
3	Sodium adduct	$[m + \text{Na}]^+$

3.3. Identification of Metabolites. The corrected, filtered, and combined data sets were used to identify metabolites which show a significant change of abundance in the wound time course in WT and/or jasmonate deficient mutant plants. First, the corrected masses of marker candidates were matched to molecular masses of all compounds recorded in the KEGG [17] and AraCyc [18] database or literature [33] based on a tolerance of 0.005 Da. The identity of marker candidates was confirmed based on the isotopic pattern and coelution with identical standards or MS/MS fragmentation [34]. Thus, a number of oxylipins could be identified as wound-induced metabolite markers (see Table 4). Oxylipins are metabolites deriving from lipid peroxidation and are involved in regulating developmental processes as well as environmental responses, like the inflammatory or wound response, in nearly every organism. Among these bioactive

TABLE 4: Identified metabolites in the combined and filtered data set. The retention time is measured in minutes and the exact compound mass is stated in Dalton. The columns “Negative” and “Positive” contain the number of associated marker candidates/ions obtained in the negative or positive ionization mode. The column “Ions” contains the sum of associated marker candidates/ions per compound. The column “*P* value” contains the minimal adjusted *P* value of the Kruskal-Wallis test over all associated marker candidates, respectively. The column “Mass error” contains the absolute difference between the corrected mass of the marker candidate with the minimal adjusted *P* value and the exact compound mass in Dalton.

RT	Exact mass	Mass error	Name	Formula	Ions	Negative	Positive	<i>P</i> value
0.73	210.1256	0.0015	Jasmonic acid	C12H18O3	5	5	0	6.67e-8
2.08	292.2038	0.0021	OPDA	C18H28O3	8	4	4	3.41e-8
1.85	310.2144	0.0016	13-HPOT	C18H30O4	1	1	0	2.04e-4
2.49	292.2038	0.0027	13-KOT	C18H28O3	4	4	0	1.70e-7
1.33	264.1725	0.0037	dn-OPDA	C16H24O3	5	3	2	7.65e-8
0.5	226.1205	0.0009	11/12-Hydroxy jasmonic acid	C12H18O4	4	4	0	2.87e-8
0.51	339.2046	0.0008	12-Hydroxy jasmonoyl isoleucine	C18H29NO5	1	1	0	3.44e-5
0.51	353.1838	0.001	12-Carboxy jasmonoyl isoleucine	C18H27NO6	1	1	0	2.20e-5
4.02	760.4762	0.005	18:3/dn-OPDA-MGDG	C43H68O11	4	0	4	1.96e-6
2.85	774.4554	0.0022	OPDA/dn-OPDA-MGDG	C43H66O12	8	0	8	1.93e-7
3.26	802.4867	0.0034	OPDA/OPDA-MGDG	C45H70O12	7	0	7	1.79e-7
4.59	1048.6487	0.0033	OPDA/dn-OPDA-MGDG-OPDA	C61H92O14	9	0	9	1.48e-6
4.89	1076.68	0.002	OPDA/OPDA-MGDG-OPDA	C63H96O14	8	0	8	2.13e-6
2.36	936.5083	0.0023	OPDA/dn-OPDA-DGDG	C49H76O17	4	0	4	1.86e-6
2.76	964.5396	0.0021	OPDA/OPDA-DGDG	C51H80O17	6	0	6	1.98e-7

The identified oxylipins are found in literature under the following synonyms: Jasmonic acid (3-Oxo-2*R*-(2*Z*)-2-penten-1*R*-yl-cyclopentaneacetic acid), OPDA (12-Oxo-10,15(*Z*)-phytyldienoic acid or 4-Oxo-5 α -(2(*Z*)-pentenyl)-2-cyclopentene-1 α -octanoic acid), 13-HPOT (13-Hydroperoxy-octadeca-9(*Z*),11(*Z*),15(*Z*)-trienoic acid), 13-KOT (13-Keto-octadeca-9(*Z*),11(*Z*),15(*Z*)-trienoic acid), dn-OPDA (4-Oxo-5*S*-(2*Z*)-2-penten-1-yl-2-cyclopentene-1*S*-hexanoic acid), 18:3/dn-OPDA-MGDG (Arabidopsis F, Monogalactosyldiacylglycerol), OPDA/dn-OPDA-MGDG (Arabidopsis A, Monogalactosyldiacylglycerol), OPDA/OPDA-MGDG (Arabidopsis B, Monogalactosyldiacylglycerol), OPDA/dn-OPDA-MGDG-OPDA (Arabidopsis E, Acylated Monogalactosyldiacylglycerol), OPDA/OPDA-MGDG-OPDA (Arabidopsis G, Acylated Monogalactosyldiacylglycerol), OPDA/dn-OPDA-DGDG (Arabidopsis C, Digalactosyldiacylglycerol), and OPDA/OPDA-DGDG (Arabidopsis D, Digalactosyldiacylglycerol).

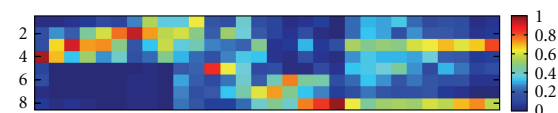


FIGURE 2: Prototype plot of the filtered and combined data in MarVis-Cluster using 30 prototypes for clustering. Every column represents the average intensity profile (prototype) of associated marker candidates. The prototypes are ordered according to similarity based on a one-dimensional self-organizing map. The first prototypes represent marker candidates in a WT-specific wound time course (high intensities in the first four conditions and almost no intensities in the last four conditions).

lipids, the mammalian and plant oxylipins are the best characterized ones. Mammals use predominantly C20 fatty acids (eicosanoids), while in plants C18 fatty acids are most abundantly used for the biosynthesis of oxylipins or so-called octadecanoids [35]. The identified oxylipins (see Table 4) are part of the α -linolenic acid metabolism or members of the compound class of mono- and digalactosyldiacylglycerols. They are described in the context of plant wounding [33, 34, 36]. Thirteen of the fifteen identified oxylipins could only be detected in either the negative or the positive ionization mode. On average, five ions/marker candidates could be

assigned per compound. The findings are supported by very low adjusted *P* values from the Kruskal-Wallis test of the intensity profiles (see previous section and Table 4).

4. Conclusions

MarVis-Filter combines essential preprocessing tools for mass spectrometry data analysis within a single user-friendly tool. Large data sets from the negative and positive ionization mode can easily be imported, corrected, filtered, and combined. Lists of ionization rules for adduct correction can be customized, extended, and commented in a convenient way using a standard text editor. Within the MarVis-Suite filtered and combined data sets can directly be clustered, visualized, and analyzed in detail using the MarVis-Cluster tool. In a case study 75 high-quality marker candidates could be clearly assigned to fifteen compounds of the oxylipin class based on the adduct and isotope correction in MarVis-Filter. The combination of data sets deriving from the negative and positive ionization mode is an important step for further data analysis. In the case study, most of the identified metabolites could only be detected in either the negative or the positive mode. The significance of the selected wound markers is supported by a high number of annotated and assigned ions/marker candidates and by very low adjusted *P* values from the Kruskal-Wallis test. The statistical filtering of

marker candidates reduced the complexity of the data sets from about 48000 to 3500 significant candidates (about 7 percent).

Acknowledgments

This work was partially supported by the DFG FOR-546-FE 446/2-3 to I. Feussner and by the Federal Ministry of Education and Research (BMBF 0315595A) to P. Meinicke and I. Feussner. A. Kaever and M. Landesfeind were supported by the Biomolecules program of the Göttingen Graduate School for Neurosciences, Biophysics, and Molecular Biosciences (GGNB). The authors are grateful to Lars Söder for support in software development and testing, to Farina Schill for critical reading the manuscript and to Pia Meyer for expert technical assistance.

References

- [1] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer, "Metabolite profiling for plant functional genomics," *Nature Biotechnology*, vol. 18, no. 11, pp. 1157–1161, 2000.
- [2] J. K. Nicholson, J. C. Lindon, and E. Holmes, "Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica*, vol. 29, no. 11, pp. 1181–1189, 1999.
- [3] J. Lisec, N. Schauer, J. Kopka, L. Willmitzer, and A. R. Fernie, "Gas chromatography mass spectrometry-based metabolite profiling in plants," *Nature Protocols*, vol. 1, no. 1, pp. 387–396, 2006.
- [4] R. C. H. De Vos, S. Moco, A. Lommen, J. J. B. Keurentjes, R. J. Bino, and R. D. Hall, "Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry," *Nature Protocols*, vol. 2, no. 4, pp. 778–791, 2007.
- [5] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, 2007.
- [6] M. Katajamaa and M. Orešić, "Data processing for mass spectrometry-based metabolomics," *Journal of Chromatography A*, vol. 1158, no. 1-2, pp. 318–328, 2007.
- [7] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787, 2006.
- [8] P. Meinicke, T. Lingner, A. Kaever et al., "Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps," *Algorithms for Molecular Biology*, vol. 3, no. 1, article 9, 2008.
- [9] V. Shulaev, D. Cortes, G. Miller, and R. Mittler, "Metabolomics for plant stress response," *Physiologia Plantarum*, vol. 132, no. 2, pp. 199–208, 2008.
- [10] L. Tarpley, A. L. Duran, T. H. Kebrom, and L. W. Sumner, "Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period," *BMC Plant Biology*, vol. 5, article 8, 2005.
- [11] K. Nahlik, M. Dumkow, Ö Bayram et al., "The COP9 signalosome mediates transcriptional and metabolic response to hormones, oxidative stress protection and cell wall rearrangement during fungal development," *Molecular Microbiology*, vol. 78, no. 4, pp. 964–979, 2010.
- [12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] M. Beckmann, D. P. Enot, D. P. Overy, and J. Draper, "Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 9, pp. 3444–3451, 2007.
- [14] J. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, CRC Press, 2003.
- [15] A. Koulman, B. A. Tapper, K. Fraser, M. Cao, G. A. Lane, and S. Rasmussen, "High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics," *Rapid Communications in Mass Spectrometry*, vol. 21, no. 3, pp. 421–428, 2007.
- [16] D. A. MacKenzie, M. Defernez, W. B. Dunn et al., "Relatedness of medically important strains of *Saccharomyces cerevisiae* as revealed by phylogenetics and metabolomics," *Yeast*, vol. 25, no. 7, pp. 501–512, 2008.
- [17] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.
- [18] P. Zhang, H. Foerster, C. P. Tissier et al., "MetaCyc and AraCyc. Metabolic pathway databases for plant research," *Plant Physiology*, vol. 138, no. 1, pp. 27–37, 2005.
- [19] M. Sud, E. Fahy, D. Cotter et al., "LMSD: LIPID MAPS structure database," *Nucleic Acids Research*, vol. 35, no. 1, pp. D527–D532, 2007.
- [20] D. S. Wishart, C. Knox, A. C. Guo et al., "HMDB: a knowledgebase for the human metabolome," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D603–D610, 2009.
- [21] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Research*, vol. 37, no. 2, pp. W623–W633, 2009.
- [22] J. Draper, D. P. Enot, D. Parker et al., "Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'," *BMC Bioinformatics*, vol. 10, article 227, 2009.
- [23] R. Tautenhahn, C. Böttcher, and S. Neumann, "Annotation of LC/ESI-MS mass signals," in *1st International Conference on Bioinformatics Research and Development (BIRD '07)*, vol. 4414 of *Lecture Notes in Computer Science*, pp. 371–380, Berlin, Germany, March 2007.
- [24] T. Kind and O. Fiehn, "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry," *BMC Bioinformatics*, vol. 8, article 105, 2007.
- [25] R. Gentleman, R. Ihaka et al., <http://www.r-project.org/>.
- [26] B. Jones, *MATLAB: Statistics Toolbox User's Guide*, MathWorks, 1993.
- [27] A. Lommen, "Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing," *Analytical Chemistry*, vol. 81, no. 8, pp. 3079–3086, 2009.
- [28] M. Sturm, A. Bertsch, C. Gröpl et al., "OpenMS—an open-source software framework for mass spectrometry," *BMC Bioinformatics*, vol. 9, article 163, 2008.
- [29] A. Kaever, T. Lingner, K. Feussner, C. Göbel, I. Feussner, and P. Meinicke, "MarVis: a tool for clustering and visualization of metabolic biomarkers," *BMC Bioinformatics*, vol. 10, article 92, 2009.

- [30] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.
- [31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, pp. 289–300, 1995.
- [32] B. Von Malek, E. Van Der Graaff, K. Schneitz, and B. Keller, "The Arabidopsis male-sterile mutant dde2-2 is defective in the ALLENE OXIDE SYNTHASE gene encoding one of the key enzymes of the jasmonic acid biosynthesis pathway," *Planta*, vol. 216, no. 1, pp. 187–192, 2002.
- [33] C. Göbel and I. Feussner, "Methods for the analysis of oxylipins in plants," *Phytochemistry*, vol. 70, no. 13-14, pp. 1485–1503, 2009.
- [34] A. Ibrahim, A. Schütz, J. Galano et al., "The alphabet of galactolipids in Arabidopsis thaliana," *Frontiers in Plant Physiology*, vol. 2, article 95, 2011.
- [35] A. Andreou, F. Brodhun, and I. Feussner, "Biosynthesis of oxylipins in non-mammals," *Progress in Lipid Research*, vol. 48, no. 3-4, pp. 148–170, 2009.
- [36] G. A. Howe and G. Jander, "Plant immunity to insect herbivores," *Annual Review of Plant Biology*, vol. 59, pp. 41–66, 2008.

Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets

The following paper was published 2014 in *PLoS ONE* [105]. The Supporting Information are available on <http://dx.doi.org/10.1371/journal.pone.0089297>.

The sections on the biological interpretation of results from data analysis were written and all tables created in close collaboration of Alexander Kaever, Dr. Kirstin Feussner, and Prof. Dr. Ivo Feussner. The parsing of KEGG and BioCyc flatfiles, the calculation of monoisotopic masses for pathway database construction and the method for database query were partially implemented by Manuel Landesfeind. The article was critically revised by all coauthors.

Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets

Alexander Kaever^{1*}, Manuel Landesfeind¹, Kirstin Feussner², Burkhard Morgenstern¹, Ivo Feussner², Peter Meinicke¹

1 Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University, Göttingen, Germany, **2** Department of Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-August-University, Göttingen, Germany

Abstract

A major challenge in current systems biology is the combination and integrative analysis of large data sets obtained from different high-throughput omics platforms, such as mass spectrometry based Metabolomics and Proteomics or DNA microarray or RNA-seq-based Transcriptomics. Especially in the case of non-targeted Metabolomics experiments, where it is often impossible to unambiguously map ion features from mass spectrometry analysis to metabolites, the integration of more reliable omics technologies is highly desirable. A popular method for the knowledge-based interpretation of single data sets is the (Gene) Set Enrichment Analysis. In order to combine the results from different analyses, we introduce a methodical framework for the meta-analysis of p-values obtained from Pathway Enrichment Analysis (Set Enrichment Analysis based on pathways) of multiple dependent or independent data sets from different omics platforms. For dependent data sets, e.g. obtained from the same biological samples, the framework utilizes a covariance estimation procedure based on the nonsignificant pathways in single data set enrichment analysis. The framework is evaluated and applied in the joint analysis of Metabolomics mass spectrometry and Transcriptomics DNA microarray data in the context of plant wounding. In extensive studies of simulated data set dependence, the introduced correlation could be fully reconstructed by means of the covariance estimation based on pathway enrichment. By restricting the range of p-values of pathways considered in the estimation, the overestimation of correlation, which is introduced by the significant pathways, could be reduced. When applying the proposed methods to the real data sets, the meta-analysis was shown not only to be a powerful tool to investigate the correlation between different data sets and summarize the results of multiple analyses but also to distinguish experiment-specific key pathways.

Citation: Kaever A, Landesfeind M, Feussner K, Morgenstern B, Feussner I, et al. (2014) Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets. PLoS ONE 9(2): e89297. doi:10.1371/journal.pone.0089297

Editor: Andrew C. Gill, University of Edinburgh, United Kingdom

Received: September 6, 2013; **Accepted:** January 20, 2014; **Published:** February 28, 2014

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

Funding: Alexander Kaever and Manuel Landesfeind were funded by the German Federal Ministry of Education and Research (BMBF BioFung project 0315595A), and were supported by the Biomolecules program of the Göttingen Graduate School for Neurosciences, Biophysics, and Molecular Biosciences (GGNB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: alex@gobics.de

Introduction

High-throughput omics platforms, such as mass spectrometry (MS) based Metabolomics and Proteomics or DNA microarray or RNA-seq-based Transcriptomics, allow the comprehensive analysis of an organism's reaction under different experimental conditions [1–5]. A current major challenge in systems biology is the combination and integrative analysis of the large data sets obtained from these platforms [6–8]. A single data set usually contains the intensity/expression profiles (intensities for all measured samples) of thousands of features, such as different ion species in MS or spots in DNA microarray analysis. After individual preprocessing of each data set, which includes the statistical analysis, ranking, or filtering of features according to the relevance of their profiles [9–11], the features have to be assigned to known biological entities [12], such as metabolites, genes, or proteins. Especially in MS-based Metabolomics, a major bottleneck is the identification of metabolites in non-targeted experiments [13]. In many applications, the putative monoisotopic masses of measured ion species cannot unambiguously be mapped to metabolite entries in public databases. The integration of data

from other omics platforms which provide a more reliable mapping, such as DNA microarrays, can significantly support the metabolite identification in this case. After annotation, the results are usually interpreted in the context of current knowledge, e.g. known biochemical pathways or processes [14–16]. A popular method for this knowledge-based interpretation of single data sets is the Gene Set Enrichment Analysis [17] or Overrepresentation Analysis [18,19]. Many similar approaches have been developed and the methodology was transferred to other omics platforms [20–23]. In general, the enrichment analysis is based on sets of entities, e.g. pathways with associated metabolites, and results in a list of relevant sets which are enriched in high-ranking features (in comparison to all features in the data set). In most methods, the enrichment level of a single set is expressed as p-value. Modelling metabolic pathways as well-defined sets of biological entities, e.g. metabolites, enzymes, and corresponding genes, has shown to be a powerful approach to interpreting complex omics data sets. Furthermore, the concept of pathways associated with different types of biological entities facilitates the joint analysis of different data sets [24].

Table 1. Overview on data sets.

Label	Number of features	Times	Platform	Ionization mode	Reference
M1	24796	0.5 h, 2 h, 5 h	Mass spectrometry	negative	[11]
M2	23325	0.5 h, 2 h, 5 h	Mass spectrometry	positive	[11]
T1	25392	1 h	DNA microarray	-	[32], E-ATMX-9
T2	25392	3 h	DNA microarray	-	E-MEXP-1475

The table gives an overview on the four data sets used for evaluation and application. The third column (Times) summarizes the different points in time when the wounded plants were harvested in the respective experiment. The T1 and T2 data sets can be obtained from the ArrayExpress [44] website. doi:10.1371/journal.pone.0089297.t001

The combination of results from different studies sharing the same experimental design in terms of null and alternative hypothesis (meta-analysis) is a central task in various statistical applications [25–27]. In case of the combination of independent p-values, Fisher’s method [28] or Stouffer’s method [29], also known as normal, Z-method, or Z-transform test, are often applied. For dependent p-values and known covariances, in [30] an extended version of Fisher’s method was proposed (Brown’s method). In order to increase statistical power, meta-analysis has been applied to Pathway Enrichment Analysis (Set Enrichment Analysis utilizing pathways as sets) in the context of cancer studies

[31]. The proposed methods were focused on the combination of independent p-values based on DNA microarray data. In contrast, we introduce a general methodical framework for the meta-analysis of multiple dependent or independent data sets resulting from different omics platforms applied to Pathway Enrichment Analysis. In order to cope with dependent data sets, such as obtained from the same biological samples analyzed by MS in negative and positive ionization mode, the framework utilizes a covariance estimation procedure based on the nonsignificant pathways in single data set enrichment analysis. The framework is applied and evaluated on two Metabolomics MS data sets [32]

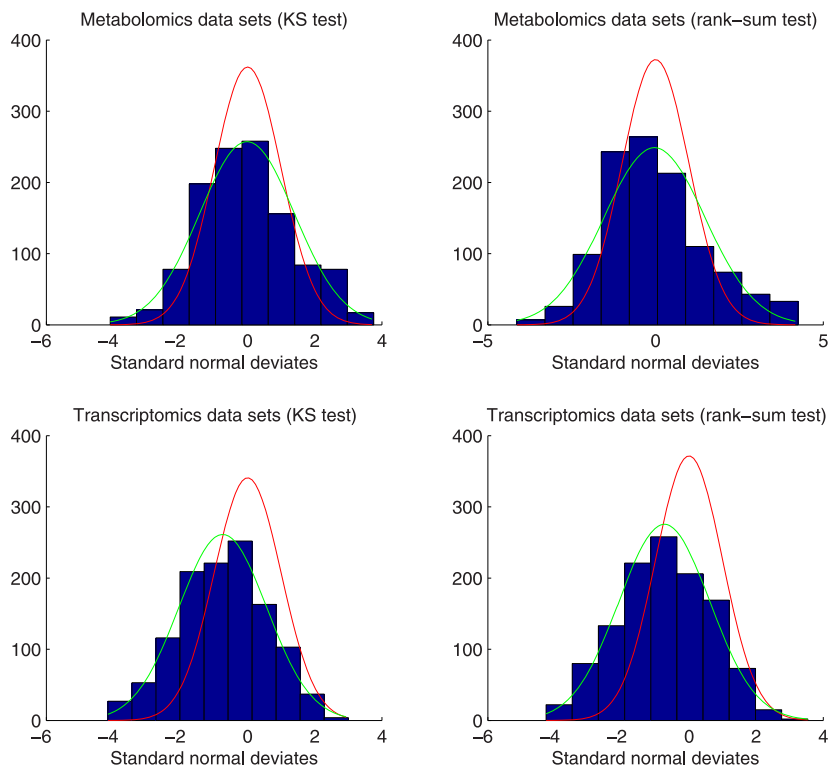


Figure 1. Histograms of standard normal deviates for the Metabolomics and Transcriptomics data sets. For the p-value calculation, the Kolmogorov-Smirnov (KS) and rank-sum tests were utilized. The p-values were restricted to the range $(10^{-5}, 1 - 10^{-5})$. The red graph represents the expected density assuming the standard normal distribution. The green graph shows the expected density assuming a normal distribution with the sample mean and standard deviation as parameters. The histograms for both tests are similar and confirm the normal-like distribution of deviates. In both cases however, the sample standard deviation is higher than the unit standard deviation used for the transformation. Additionally, the sample mean for the combined Transcriptomics data sets is smaller than zero. doi:10.1371/journal.pone.0089297.g001

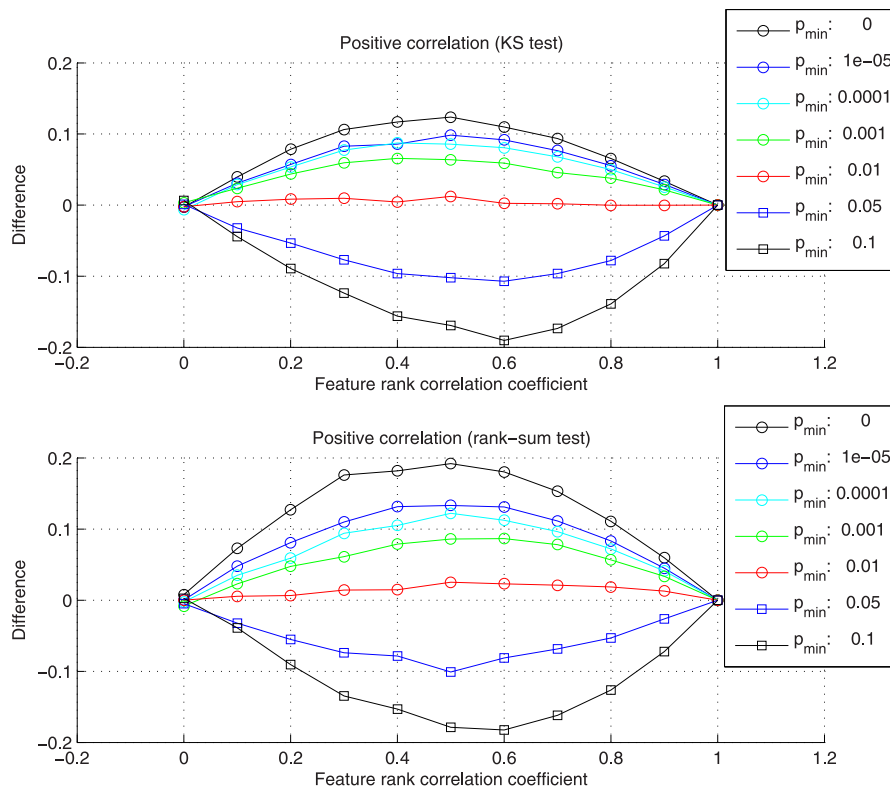


Figure 2. Differences between the reconstructed correlation coefficients from pathway enrichment and the introduced positive feature correlation. The differences were calculated for different p_{min} values and the Kolmogorov-Smirnov (KS) and rank-sum test. The best reconstruction, corresponding to differences near zero, can be observed for $p_{min}=0.01$.
doi:10.1371/journal.pone.0089297.g002

and two Transcriptomics DNA microarray studies [11] in the context of wounding of *Arabidopsis thaliana*. The main focus of this exemplary meta-analysis lies on the enhancement of MS based Metabolomics results by means of the microarray studies.

Materials and Methods

Data sets and preprocessing

For application and evaluation of the meta-analysis, two Metabolomics MS data sets (M1 and M2) [11] and two Transcriptomics DNA microarray data sets (T1 and T2) [32] were used (see Table 1 and Dataset S1 for details). All studies investigate the wounding of *Arabidopsis thaliana* wild type and the jasmonate-deficient *dde 2-2* mutant plants [33], the experimental designs comprise conditions for control plants as well as plants harvested at different times after wounding (see Table 1). The two Metabolomics data sets derive from an Ultra Performance Liquid Chromatography (UPLC) analysis coupled to a Time-Of-Flight (TOF) MS detection. With this method, the non-polar extraction phase of one set of samples was analyzed in positive and negative ionization mode. Since some metabolites may have been measured in both ionization modes following different (partially unknown) ionization rules [34], the level of dependence between both data sets is not clear. In case of the MS data sets, a single feature corresponds to a particular ion species, which is characterized by an exact mass-to-charge ratio and a retention time. A single

metabolite may be represented by multiple features, e.g. corresponding to different adduct formations and isotopologues. The features in the microarray data sets correspond to different spots on the array containing DNA probes that match a particular sequence. Also in this case, a single transcript may be represented by multiple features corresponding to particular sequences of the respective gene. The feature profiles of all data sets were ranked separately utilizing a signal-to-noise ratio (similar to the method described in [9], see TechnicalDescription S1).

Pathway enrichment analysis

The ranked features were mapped to the pathway entries in AraCyc [35] and the *Arabidopsis*-specific pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [14] (see TechnicalDescription S1). In case of the Metabolomics MS data sets, all potential monoisotopic masses were calculated per feature based on the ionization rules and number of isotopes used in [11] and mapped to the metabolite masses in the databases. In case of the Transcriptomics DNA microarray data, the features were mapped to the *A. thaliana* genes utilizing their CATMA IDs [36]. Based on the mappings, a set of feature ranks was extracted for each pathway and data set. In order to test for an over-representation of high-ranked features, a p-value was calculated for each set of ranks (pathway) utilizing a one-sided Kolmogorov-Smirnov (KS) or Wilcoxon rank-sum test (also known as Mann-Whitney U test) [21]. In case of the KS test, the empirical

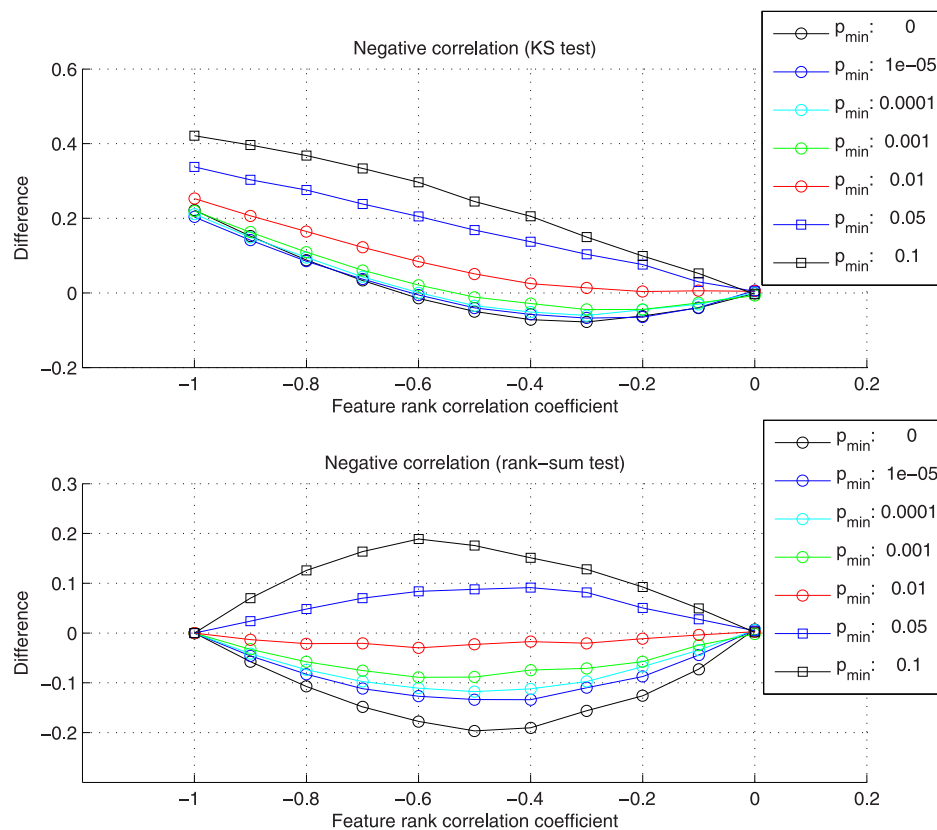


Figure 3. Differences between the reconstructed correlation coefficients from pathway enrichment and the introduced negative feature correlation. The differences were calculated for different p_{min} values and the Kolmogorov-Smirnov (KS) and rank-sum test. The best reconstruction, corresponding to differences near zero, can be observed for $p_{min}=0.01$. The KS test is not able to fully reconstruct strong negative feature correlations.

doi:10.1371/journal.pone.0089297.g003

distribution of ranks in a given set is compared to the distribution of ranks in the respective data set. In case of the rank-sum test, the sum of feature ranks within a given set is evaluated. Especially for Gene Set Enrichment Analysis of DNA microarrays, many methods have been published [20]. Most of these methods are based on KS-like or average gene-specific statistics. For a general meta-analysis and in order to combine the Metabolomics and Transcriptomics data sets in a robust way, we decided to utilize the rank-based KS and rank-sum test. However, more specialized methods for the pathway-specific p-value calculation may be employed as well. The resulting p-values for the dependent Metabolomics data sets were used for the covariance estimation (see corresponding section). The covariances between both Transcriptomics data sets and between the Metabolomics and Transcriptomics data sets, which were obtained from independent biological samples, were set to zero.

Meta-analysis of p-values

In statistical meta-analysis, the most common methods for combining independent p-values from related tests are Fisher's [28] and Stouffer's method [29]. In Fisher's method, the meta-p-value is calculated based on a chi-squared distribution (see TechnicalDescription S1). In Stouffer's method, the test statistic is

the sum of p-values transformed into normally distributed random variables (standard normal deviates). For dependent p-values, a powerful approach is Brown's method [30], which is an extension of Fisher's method based on a scaled chi-squared distribution and modified degrees of freedom utilizing a known covariance matrix for standard normal deviates. The given p-values can be transformed into standard normal deviates by means of the inverse cumulative distribution function of the standard normal distribution. The covariance matrix of the standard normal deviates can also be utilized in order to extend Stouffer's method to dependent p-values.

Estimation of covariances

In most applications with dependent data sets, the covariance matrix is not known and has to be estimated. In our proposed procedure, the pairwise covariance between two data sets is estimated based on the standard normal deviates of the pathway-specific p-values, which were obtained for each single data set in Pathway Enrichment Analysis. This estimation is expected to be biased by the alternative hypothesis since the similar or same experimental setup of the data sets imposes a certain dependence and significant pathways associated with very low p-values will strongly influence the results. In order to minimize this bias in the

Table 2. Results from meta-analysis of pathway enrichment (Brown's method).

Rank	DB	Pathway	Hits	KS	Rank-sum
1	KEGG	alpha-Linolenic acid metabolism	214	0.0001321	2.383e-05
2	AraCyc	jasmonic acid biosynthesis	176	0.003676	0.0003101
3	AraCyc	glycolipid desaturation	325	0.0007046	0.0102
4	KEGG	Linoleic acid metabolism	147	0.5252	0.3968
5	AraCyc	superpathway of phenylalanine, tyrosine and tryptophan biosynthesis	86	0.5364	0.5253
6	AraCyc	traumatin and (Z)-3-hexen-1-yl acetate biosynthesis	131	0.4538	0.5557
7	KEGG	2-Oxocarboxylic acid metabolism	578	0.5252	0.767
8	AraCyc	glucosinolate biosynthesis from dihomomethionine	153	0.5364	0.7857
9	KEGG	Starch and sucrose metabolism	335	0.5252	0.7857
10	KEGG	Proteasome	114	0.5252	0.7857

The table contains the high-ranking pathways from meta-analysis of pathway enrichment (Brown's method) based on the Kolmogorov-Smirnov (KS) and rank-sum test utilizing all data sets. The p-values per data set were restandardized. The pathways are sorted according to the meta-p-values derived from the rank-sum test. The second column (DB) contains the name of the source database, the fourth column (Hits) the number of feature assignments. The last two columns comprise the false discovery rates calculated from the meta-p-values.

doi:10.1371/journal.pone.0089297.t002

estimation of covariances for dependent data sets, and the meta-analysis based on the previous results were applied and evaluated on the four Metabolomics/Transcriptomics data sets (see previous section). First, in order to check the distribution of transformed p-values, the histograms of the standard normal deviates were inspected. Because of significant pathways which are highly relevant in this context, the p-values are expected to be not fully uniformly distributed, which may result in a distribution of transformed p-values that deviates from the standard normal distribution. In this case, the p-values/normal deviates should be corrected for significance analysis. Second, the performance of the introduced method in reconstructing simulated data set correlations was evaluated for different p_{min} values. This performance was not clear, since the proposed correlation estimation includes several complex steps, such as the mapping of a proportion of feature ranks to pathways of different size, the calculation and restriction of p-values, and the transformation into normal deviates. Additionally, the p_{min} parameter might have a strong influence on the results. Therefore, another objective of the simulation studies was the identification of an appropriate

parameter value for the real data sets. Third, the correlation estimation and meta-analysis were applied to all four real data sets. All data sets, containing the annotation information from the pathway mapping, and the results from Pathway Enrichment Analysis are available as comma-separated-values files (see Dataset S1 and Table S1). The source code of functions for the meta-analysis of p-values can be found in File S1.

Distribution of standard normal deviates

Figure 1 shows the histograms of the transformed p-values (standard normal deviates) from Pathway Enrichment Analysis for the two Metabolomics and two Transcriptomics data sets within the p-value range ($10^{-5}, 1 - 10^{-5}$). The histograms for the KS and the rank-sum test are similar and confirm the normal-like distribution of deviates. In both cases however, the sample standard deviation is higher than the unit standard deviation used for the transformation. Additionally, the sample mean for the combined Transcriptomics data sets is smaller than zero. This difference may be caused by pathways which are directly or indirectly influenced by the experimental setup. Although the

Table 3. Results from meta-analysis of pathway enrichment (Stouffer's extended method).

Rank	DB	Pathway	Hits	KS	Rank-sum
1	KEGG	alpha-Linolenic acid metabolism	214	6.708e-05	1.127e-05
2	AraCyc	jasmonic acid biosynthesis	176	0.001774	7.328e-05
3	AraCyc	glycolipid desaturation	325	0.02043	0.2122
4	AraCyc	traumatin and (Z)-3-hexen-1-yl acetate biosynthesis	131	0.4545	0.4326
5	AraCyc	superpathway of phenylalanine, tyrosine and tryptophan biosynthesis	86	0.4545	0.4326
6	KEGG	Linoleic acid metabolism	147	0.7866	0.499
7	AraCyc	glucosinolate biosynthesis from dihomomethionine	153	0.4545	0.6282
8	AraCyc	glucosinolate biosynthesis from tryptophan	132	0.4545	0.6583
9	AraCyc	glucosinolate biosynthesis from phenylalanine	115	0.6522	0.6583
10	AraCyc	glucosinolate biosynthesis from tetrahomomethionine	114	0.4545	0.6583

The table contains the high-ranking pathways from meta-analysis of pathway enrichment (Stouffer's extended method) based on the Kolmogorov-Smirnov (KS) and rank-sum test utilizing all data sets. The last two columns comprise the false discovery rates calculated from the meta-p-values.

doi:10.1371/journal.pone.0089297.t003

Table 4. Results from pathway enrichment analysis of data set M1.

Rank	DB	Pathway	Hits	KS	Rank-sum
1	KEGG	alpha-Linolenic acid metabolism	65	0.02084	0.03717
2	AraCyc	jasmonic acid biosynthesis	68	0.1531	0.1503
3	KEGG	Linoleic acid metabolism	43	0.4598	0.8524
4	AraCyc	indole-3-acetyl-amino acid biosynthesis	29	0.4598	0.8524
5	AraCyc	traumatol and (Z)-3-hexen-1-yl acetate biosynthesis	38	0.4598	0.8524
6	AraCyc	galactosylcyclitol biosynthesis	14	0.4598	0.8524
7	AraCyc	glycolipid desaturation	144	0.4598	0.8524
8	KEGG	Porphyryr and chlorophyll metabolism	222	0.8841	0.8524
9	AraCyc	poly-hydroxy fatty acids biosynthesis	59	0.9248	0.8524
10	KEGG	Lysine degradation	46	0.4598	0.8524

The table contains the high-ranking pathways from enrichment analysis of data set M1 based on the Kolmogorov-Smirnov (KS) and rank-sum test. The pathways are sorted according to the restandardized p-values derived from the rank-sum test. The last two columns comprise the false discovery rates calculated from the restandardized p-values.

doi:10.1371/journal.pone.0089297.t004

highly significant pathways with p-values below the threshold 10^{-5} were left out, many other pathways are expected to be indirectly affected by the wounding process. Another explanation would be the dependence of feature ranks used for p-value calculation, e.g. introduced by the dependence of different microarray spots representing the same gene or by gene-gene

correlations [17]. In order to eliminate the observed bias, the p-values were restandardized [37] for significance analysis by means of the sample mean and sample standard deviation of observed normal deviates per data set and retransforming of the standardized deviates into corrected p-values. This is a conservative correction because the observed bias also includes the pathways which are directly influenced by the wounding process.

Table 5. Selected feature mappings from data set M1.

Rank	rt	m/z	Mappings
1	0.73	255.1218	Jasmonic acid
3	0.73	209.1168	Jasmonic acid
7	0.73	256.1264	Jasmonic acid
8	2.08	337.1999	OPDA, EOTrE
11	2.08	338.2044	OPDA, EOTrE
321	5.66	986.6145	18:3/18:1-DGD, 18:2/18:2-DGD
324	5.78	822.5428	18:2/16:0-MGD, 18:1/16:1-MGD
410	5.53	820.5295	18:3/16:0-MGD, 18:2/16:1-MGD, 18:1/16:2-MGD
447	2.33	339.2155	OPC-8:0
540	5.64	960.5985	18:3/16:0-DGD
542	6.02	823.5541	18:1/16:0-MGD, 18:0/16:1-MGD
554	5.67	858.5064	18:3/18:3-MGD
563	5.74	795.5232	18:3/18:1-MGD, 18:2/18:2-MGD
650	6.18	939.5986	18:2/18:3-DGD
846	5.89	962.613	18:2/16:0-DGD
879	6.23	859.5155	18:2/18:3-MGD
899	1.86	309.2055	HpOTrE
1445	6.17	964.6258	18:1/16:0-DGD
1727	7.53	278.2245	Linolenic acid
2142	0.52	239.0895	9-Oxononanoic acid

The table shows selected mappings of features from data set M1 (24796 features) to entries in the first three pathways in tables 2 and 3. The first column contains the feature rank. The second and third column show the corresponding retention times and mass-to-charge ratios. Multiple mappings correspond to different ionization rules or isotopologues.

doi:10.1371/journal.pone.0089297.t005

Estimation of data set correlation

In simulated studies (see TechnicalDescription S1 for details), the correlation estimation was evaluated by calculating the pairwise Pearson correlation coefficients between all four data sets and a copy of the respective data set with different percentages of feature ranks randomly permuted. For each original and permuted data set, the p-values were calculated for all pathways using the KS or rank-sum test. The correlation coefficient between each original and permuted data set was computed based on the respective standard normal deviates (not restandardized) and the restriction of p-values utilizing different parameter values p_{min} . As measurement of the introduced artificial correlation, the correlation coefficient between the feature ranks of each data set and the permuted ranks (feature rank correlation) was calculated and averaged, respectively. The whole procedure was repeated for negative correlation by randomly permuting a percentage of the inverted original feature ranks per data set.

Table S2 shows the average results over all data sets in detail. Figure 2 and 3 summarize the differences between the reconstructed correlation coefficients from pathway enrichment and the introduced positive or negative feature rank correlation. In comparison to the average feature rank correlation coefficients (x-axis), the absolute correlation is overestimated for low p_{min} values and underestimated for high values. A p_{min} value of 0.01 results in the best reconstruction of data set correlation, the absolute difference between the correlation coefficients from pathway enrichment and the feature rank correlation is close to zero for both tests. In case of the observed overestimation for low p_{min} values, the relevant pathways, which are associated with many top-ranking features, are assigned a low p-value, even when randomly permuting some of the features, and have a high influence on the correlation estimation. In case of the underestimation for high p_{min} values, the introduced correlation over all

Table 6. Results from pathway enrichment analysis of data set M2.

Rank	DB	Pathway	Hits	KS	Rank-sum
1	AraCyc	glycolipid desaturation	167	0.0009173	0.002862
2	AraCyc	antheraxanthin and violaxanthin biosynthesis	63	0.1033	0.2477
3	KEGG	Carotenoid biosynthesis	389	0.3608	0.8365
4	AraCyc	zeaxanthin biosynthesis	29	0.5251	0.8365
5	AraCyc	lutein biosynthesis	34	0.5251	0.8365
6	AraCyc	capsanthin and capsorubin biosynthesis	38	0.5251	0.8365
7	AraCyc	brassinosteroids inactivation	20	0.5251	0.8365
8	KEGG	Porphyrin and chlorophyll metabolism	236	0.8693	0.8365
...
12	KEGG	alpha-Linolenic acid metabolism	89	0.8693	0.8365
13	AraCyc	jasmonic acid biosynthesis	54	0.8693	0.8365

The table contains the high-ranking pathways from enrichment analysis of data set M2 based on the Kolmogorov-Smirnov (KS) and rank-sum test. The last two columns comprise the false discovery rates calculated from the restandardized p-values.
doi:10.1371/journal.pone.0089297.t006

features and pathways cannot be fully recovered when restricting the range of p-values and number of utilized pathways too much.

For the KS test and small negative feature rank correlations, the estimated coefficients from enrichment are considerably larger, e.g. showing a difference between 0.2 and 0.4 in case of a feature rank correlation of -1 (see Figure 3). This can be explained by the non-symmetric properties of the one-sided KS test. A set enriched in both high-ranking and low-ranking features would receive a low p-value when performing the one-sided KS test on the original as well as the inverted ranks. The rank-sum test, on the contrary,

would result in an average p-value in both cases because the sum of ranks in the set is near the expected value. For a p_{min} value of 0.01 and negative correlation, the KS test is still able to reconstruct feature rank correlation coefficients between 0 and -0.3 with a difference near zero.

For the correlation estimation between the two dependent Metabolomics data sets, a p_{min} value of 0.01, which showed the best reconstruction in the simulations, was utilized. The estimation resulted in relatively small coefficients, 0.12 (KS test) and 0.08 (rank-sum test).

Table 7. Selected feature mappings from data set M2.

Rank	rt	m/z	Mappings
2	2.08	310.2377	OPDA, EOTrE
8	2.08	293.2117	OPDA, EOTrE
11	2.08	311.2422	OPDA, EOTrE
48	2.08	315.1932	OPDA, EOTrE
180	6.17	942.6175	18:1/16:0-DGD
211	6.17	941.6124	18:2/18:3-DGD
231	6.22	772.5912	18:2/16:0-MGD, 18:1/16:1-MGD
248	5.51	776.5365	18:3/16:0-MGD, 18:2/16:1-MGD, 18:1/16:2-MGD
295	6.00	960.6576	18:3/18:1-DGD, 18:2/18:2-DGD
297	4.69	772.5034	18:3/16:2-MGD
310	5.07	937.5843	18:3/18:3-DGD
330	5.87	935.6452	18:2/16:0-DGD
413	6.15	915.5996	16:0/18:1-DGD
459	6.45	774.6054	18:1/16:0-MGD, 18:0/16:1-MGD
507	5.72	768.56	18:3/16:1-MGD, 18:2/16:2-MGD, 18:1/16:3-MGD
615	5.18	748.5052	18:3/16:3-MGD
699	1.41	441.3184	Volicitin

The table contains selected feature mappings from data set M2 (23325 features) to the first three pathways in tables 2 and 3. Multiple mappings correspond to different ionization rules or isotopologues.
doi:10.1371/journal.pone.0089297.t007

Meta-analysis of pathway enrichment

Tables 2 and 3 show the results from meta-analysis of pathway enrichment utilizing Brown's and Stouffer's extended method integrating the correlation estimation for the Metabolomics data sets. The pathways are sorted according to the False Discovery Rate (FDR) [38] calculated based on the meta-p-values. Pathways with more than 500 associated entries were left out in this analysis for better interpretability. For both methods, the top-ranked pathways are the "alpha-Linolenic acid metabolism" (KEGG, 214 feature hits), the "jasmonic acid biosynthesis" (AraCyc, 176 feature hits), and the "lycolipid desaturation" (AraCyc, 325 feature hits). These pathways specifically describe parts of the biosynthesis of the well-known wound hormone jasmonate [39]. The first two pathways cover all biosynthetic steps from the fatty acid alpha-linolenic acid to jasmonic acid. The first committed step is catalyzed by the allene oxide synthase (AOS), whose gene is mutated in the *dde 2-2* mutant plants [33]. The glycolipid desaturation pathway describes the formation of the alpha-linolenic acid via sequential steps of glycolipid-linked desaturation. The FDRs for these key pathways are much lower compared to the following pathways. Tables 4, 5, 6, 7, 8, 9, 10, and 11 show the results from enrichment analysis of the four single data sets and selected mappings of top-ranked features which were assigned to entries in the three key pathways, respectively. The enrichment analysis of the M1 data set (negative ionization mode, see Table 4) provides a major contribution to the results from meta-analysis. The first two pathways are also top-ranked but associated with much higher FDRs. The high-ranked features associated with jasmonic acid and its precursor metabolites, such as OPDA and OPC-8:0, are mainly responsible for this ranking (see Table 5). However, the mapping of putative monoisotopic feature masses to

Table 8. Results from pathway enrichment analysis of data set T1.

Rank	DB	Pathway	Hits	KS	Rank-sum
1	KEGG	Glycolysis/Gluconeogenesis	108	0.3527	0.2952
2	KEGG	Proteasome	57	0.2885	0.2952
3	KEGG	Protein processing in endoplasmic reticulum	176	0.2885	0.2952
4	KEGG	Ribosome	220	0.02489	0.2952
5	KEGG	Oxidative phosphorylation	118	0.2885	0.2952
6	KEGG	Phenylalanine, tyrosine and tryptophan biosynthesis	54	0.492	0.2952
7	AraCyc	superpathway of phenylalanine, tyrosine and tryptophan biosynthesis	43	0.4966	0.3302
...
14	AraCyc	jasmonic acid biosynthesis	27	0.8201	0.35
...
23	KEGG	alpha-Linolenic acid metabolism	30	0.6202	0.4782
...
420	AraCyc	glycolipid desaturation	7	0.976	0.9758

The table contains the high-ranking pathways from enrichment analysis of data set T1 based on the Kolmogorov-Smirnov (KS) and rank-sum test. The last two columns comprise the false discovery rates calculated from the restandardized p-values.
doi:10.1371/journal.pone.0089297.t008

metabolites is error-prone and ambiguous. For example, OPDA, EOTrE, and a couple of other metabolites provided by KEGG and AraCyc share the same sum formula and single ion features cannot be unambiguously assigned without further information. In contrast to the alpha-linolenic acid metabolism pathway (KEGG), the very similar jasmonic acid biosynthesis pathway (AraCyc) is

Table 9. Selected feature mappings from data set T1.

Rank	ID	Mappings
6	AT2G06050	12-oxophytodienoate reductase 3
12	AT3G11170	fatty acid desaturase 7
16	AT5G42650	allene oxide synthase
18	AT1G17420	lipoxygenase 3
82	AT2G06050	12-oxophytodienoate reductase 3
120	AT4G15440	hydroperoxide lyase 1
226	AT5G48880	peroxisomal 3-keto-acyl-CoA thiolase 5
241	AT2G44810	phospholipase A1
316	AT1G20510	OPC-8:0 CoA ligase 1
436	AT1G76680	12-oxophytodienoate reductase 1
638	AT1G72520	lipoxygenase 4
737	AT4G16760	peroxisomal acyl-coenzyme A oxidase 1
744	AT1G17420	lipoxygenase 3
1037	AT3G45140	lipoxygenase 2
1487	AT1G13280	allene oxide cyclase 4
2116	AT2G06925	phospholipase A2-ALPHA
2788	AT2G31360	delta 9 acyl-lipid desaturase 2
3146	AT3G15290	3-hydroxyacyl-CoA dehydrogenase
4263	AT5G04040	triacylglycerol lipase SDP1
4276	AT1G76150	enoyl-CoA hydratase 2

The table contains selected feature mappings from data set T1 (25392 features) to the first three pathways in tables 2 and 3. Multiple mappings correspond to different spots on the microarray.
doi:10.1371/journal.pone.0089297.t009

associated with a much higher FDR. This can be explained by a number of additional entries found only in the AraCyc version of the pathway and representing general substrates, such as acetyl-CoA, intermediate products which could not be measured with a high signal-to-noise ratio, such as OPC6-3-hydroxyacyl-CoA, or other side products. The glycolipid desaturation pathway, which can be found at position seven, is associated with a very high FDR. Most of the glycolipid species show higher intensities and signal-to-noise ratios in positive compared to negative ionization mode, which results in a very low FDR in pathway enrichment analysis of the M2 data set (see Tables 6 and 7). In contrast, jasmonate and many direct precursor metabolites cannot be measured in positive ionization mode with sufficient intensity, which explains the less prominent ranking of the alpha-linolenic acid metabolism (rank 12) and jasmonic acid biosynthesis (rank 13). Nonetheless, metabolites such as OPDA can be measured in both ionization modes with high signal-to-noise ratio and these findings confirm the corresponding pathways in meta-analysis. Integrating the Transcriptomics data sets T1 and T2 results in a much more comprehensive data interpretation (see Tables 9 and 11). Figure 4 exemplarily shows the pathway map of the alpha-linolenic acid metabolism with marked entries matched by high-ranking features from all data sets. In this combination, the ambiguous mapping of the MS data is supported by unambiguously matching transcripts. Almost all of the transcripts corresponding to enzymes in the alpha-linolenic acid metabolism can be found in the T1 and T2 data sets with relatively high signal-to-noise ratios. This results in much lower FDRs for the jasmonate-specific pathways in meta-analysis compared to the results from single Metabolomics data set analysis. Also in the analysis of the single Transcriptomics data sets (see Tables 8 and 10), these two pathways are associated with relatively high FDRs. In case of the T1 data set, both pathways can be found at less prominent positions (rank 14 and 23, see Table 8). For both Transcriptomics data sets, the glycolipid desaturation is ranked in the middle of all pathways (rank 420 and 161). Only a small number of transcripts associated with fatty acid desaturase show a high signal-to-noise ratio (see Tables 9 and 11).

In case of both methods for meta-analysis, the pathways “Linoleic acid metabolism” and “traumatol and (Z)-3-hexen-1-yl acetate biosynthesis” can be found in the list of top-ten. These

Table 10. Results from pathway enrichment analysis of data set T2.

Rank	DB	Pathway	Hits	KS	Rank-sum
1	KEGG	alpha-Linolenic acid metabolism	30	0.5885	0.0794
2	KEGG	Starch and sucrose metabolism	142	0.6748	0.3277
3	AraCyc	jasmonic acid biosynthesis	27	0.7192	0.3277
4	KEGG	Linoleic acid metabolism	11	0.7192	0.7457
5	AraCyc	glucosinolate biosynthesis from phenylalanine	16	0.7192	0.7457
6	AraCyc	glucosinolate biosynthesis from dihomomethionine	19	0.7192	0.7457
7	KEGG	Valine, leucine and isoleucine biosynthesis	19	0.7192	0.7457
8	AraCyc	glucosinolate biosynthesis from tryptophan	21	0.7192	0.7457
9	AraCyc	starch degradation I	37	0.7192	0.7457
...
161	AraCyc	glycolipid desaturation	7	0.8851	0.9666

The table contains the high-ranking pathways from enrichment analysis of data set T2 based on the Kolmogorov-Smirnov (KS) and rank-sum test. The last two columns comprise the false discovery rates calculated from the restandardized p-values.
doi:10.1371/journal.pone.0089297.t010

pathways are directly connected with the alpha-linolenic acid metabolism and affected by the AOS mutation as well [40]. However, it should be noted that the second pathway is only of limited relevance in this context because the used genotype Columbia is a natural mutant in its second enzymatic step, the fatty acid hydroperoxide lyase reaction [41]. The 2-Oxocarboxylic acid metabolism (Brown's method) and several pathways in the ranking based on Stouffer's extended method describe

glucosinolate biosynthesis, the major chemical defense reaction of Arabidopsis plants upon wounding that is regulated by jasmonates [42]. Though, these pathways are associated with comparably high FDRs.

Comparing the results based on the KS and the rank-sum test, no clear trend towards lower FDRs can be observed. In case of Brown's method, the glycolipid desaturation pathway is associated with a much lower FDR for both tests. In case of Stouffer's extended method, both jasmonate-specific pathways are scored with lower FDRs.

Table 11. Selected feature mappings from data set T2.

Rank	ID	Mappings
25	AT5G42650	allene oxide synthase
104	AT2G06050	12-oxophytodienoate reductase 3
355	AT1G76680	12-oxophytodienoate reductase 1
376	AT5G48880	peroxisomal 3-keto-acyl-CoA thiolase 5
426	AT1G17420	lipoygenase 3
484	AT3G15870	oxidoreductase
631	AT1G19640	jasmonic acid carboxyl methyltransferase
1019	AT3G11170	fatty acid desaturase 7
1263	AT5G04040	triacylglycerol lipase SDP1
1354	AT4G16760	peroxisomal acyl-coenzyme A oxidase 1
1371	AT1G17420	lipoygenase 3
1544	AT3G45140	lipoygenase 2
1812	AT3G15850	fatty acid desaturase 5
1940	AT2G06925	phospholipase A2-ALPHA
2139	AT4G30950	fatty acid desaturase 6
2413	AT2G06050	12-oxophytodienoate reductase 3
2653	AT3G15290	3-hydroxyacyl-CoA dehydrogenase
3022	AT1G67560	lipoygenase 3
3297	AT3G06860	3-hydroxyacyl-CoA dehydrogenase
3383	AT2G33150	peroxisomal 3-keto-acyl-CoA thiolase 2

The table contains selected feature mappings from data set T2 (25392 features) to the first three pathways in tables 2 and 3. Multiple mappings correspond to different spots on the microarray.
doi:10.1371/journal.pone.0089297.t011

Discussion

The meta-analysis of pathway enrichment was evaluated and applied on two Metabolomics and two Transcriptomics data sets in the context of plant wounding. The meta-analysis based on Brown's and Stouffer's extended method is able to incorporate information from different independent and dependent omics data sets and distinguish key pathways in the experimental context. The FDRs calculated based on the meta-p-values are much lower compared to the single data set analysis. Especially for the pathway analysis of non-targeted Metabolomics studies, where the identification of metabolites is a bottleneck, the integration of data from other omics platforms, such as DNA microarrays, increases the value and reliability of results. In this application, Brown's and Stouffer's extended method showed overall similar results. However, Brown's method seems to be more powerful in case of pathways which are associated with extreme p-values for only a proportion of the data sets. The glycolipid desaturation pathway for example is associated with very small p-values (KS and rank-sum test) for the M2, relatively small p-values for the M1, and much larger p-values for the T1 and T2 data sets (see Table S1). In case of Brown's method, this pathway is associated with smaller FDRs (0.0007 and 0.01) in comparison to Stouffer's method (0.02 and 0.21). In contrast, Stouffer's method seems to be more powerful in case a pathway is associated with comparably small p-values for all data sets (see alpha-linolenic acid metabolism and jasmonic acid biosynthesis pathways). The choice of method depends on the objective of the meta-analysis, e.g. focus on pathways which show a consensus for all data sets or also including pathways with significant p-values for only a single or small number of data sets [26,43]. In the context of heterogeneous omics

data sets, which contain entities that cannot be measured in all experiments, e.g. metabolites that can be ionized either in positive or negative ionization mode, and pathways that may be associated with only a small number of entries for a particular omics platform, Brown's (or Fisher's method in case of independent p-values) seems to be the better choice. In both meta-analyses, a couple of pathways related to the wounding process were detected with relatively large FDRs. In order to combine the Metabolomics and Transcriptomics data sets in a robust way, we utilized general rank-based tests and a conservative restandardization of p-values per data set. The introduced framework may also be combined with more powerful tests specialized on microarray data analysis [37]. The enrichment analysis of the single T1 and T2 data sets resulted in considerably different rankings. This is likely to be related to the different time points when the wounded plants have been harvested (one and three hours).

In the performed simulation studies, the introduced feature rank correlation could be fully reconstructed utilizing the correlation estimation from pathway enrichment. By restricting the range of p-values via the parameter $p_{min}=0.01$, leaving out significant pathways, the estimation bias could be reduced. The comparison of the two dependent Metabolomics data sets, which were obtained from the same biological samples analyzed in positive and negative ionization mode, resulted in relatively small positive correlation coefficients. This indicates that only a small proportion of metabolites could be detected in both ionization modes with comparable quality of intensity profiles and that data from both modes should be considered in a comprehensive analysis. In general, the statistical power of the meta-analysis increases with decreasing dependence of data sets. Therefore, nearly independent data sets are desirable.

Comparing the one-sided KS and rank-sum test, both tests resulted in a similar distribution of normal deviates. In the simulation studies, the one-sided KS test was not able to fully reconstruct strong negative feature correlations. In most applications however, this type of data set correlation is not expected.

References

1. Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey R, et al. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18: 1157–1161.
2. Meinicke P, Lingner T, Kaever A, Feussner K, Göbel C, et al. (2008) Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology* 3: 9.
3. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422: 198–207.
4. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21: 33–37.
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5: 621–628.
6. Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology* 7: 198–210.
7. Weckwerth W, Wenzel K, Fiehn O (2004) Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4: 78–83.
8. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, et al. (2010) Visualization of omics data for systems biology. *Nature Methods* 7: S56–S68.
9. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98: 5116–5121.
10. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3: 3.
11. Kaever A, Landesfönd M, Possienke M, Feussner K, Feussner I, et al. (2012) MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data. *Journal of Biomedicine and Biotechnology* 2012.
12. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4: P3.
13. Dunn WB, Erban A, Weber RJ, Creek DJ, Brown M, et al. (2013) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 9: 44–66.
14. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* 40: D109–D114.
15. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 40: D742–D753.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25–29.
17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102: 15545–15550.
18. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA (2003) Global functional profiling of gene expression. *Genomics* 81: 98–104.
19. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA, et al. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biology* 4: R70.
20. Ackermann M, Strimmer K (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10: 47.
21. Barry WT, Nobel AB, Wright FA (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21: 1943–1949.
22. Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99.

Supporting Information

File S1 Matlab source code for functions used in meta-analysis.

(GZ)

Dataset S1 Data sets with database entry and pathway annotations.

The archive file contains the data sets in comma separated values format. The first column contains the feature IDs, respectively. The rt and Former m/z columns (M1 and M2 data set) contain the retention times and mass-to-charge ratios from MS analysis. The raw intensities for each sample can be found in the following columns. The s/n column shows the feature-specific signal-to-noise ratios and the last columns contain the KEGG and AraCyc entries and pathways mapped to the corresponding features and separated by slash characters.

(ZIP)

Table S1 Pathways with p-values and FDRs from Pathway Enrichment Analysis.

The comma separated values file contains the p-values, restandardized p-values, meta-p-values, and corresponding FDRs for single data set and meta-analysis.

(CSV)

Table S2 Supplementary tables for simulation studies.

(PDF)

Technical Description S1 Technical description of methods.

(PDF)

Acknowledgments

We would like to thank Helmut Grubmüller for fruitful discussions.

Author Contributions

Conceived and designed the experiments: KF IF. Performed the experiments: KF IF. Analyzed the data: AK ML KF BM IF PM. Wrote the paper: AK KF IF. Conception and design of the work: AK. Revising the article critically: ML KF BM IF PM.

23. Xia J, Wishart DS (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research* 38: W71–W77.
24. Wägele B, Witting M, Schmitt-Kopplin P, Suhre K (2012) MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. *PLoS ONE* 7: e39860.
25. Hedges LV, Olkin I (1985) *Statistical Methods for Meta-Analysis*. San Diego: Academic Press.
26. Whitlock M (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* 18: 1368–1373.
27. Loughin TM (2004) A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis* 47: 467–485.
28. Fisher RA (1925) *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
29. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM Jr (1949) *The American soldier: adjustment during army life*. Princeton: Princeton University Press.
30. Brown MB (1975) A method for combining non-independent, one-sided tests of significance. *Biometrics* 31: 987–992.
31. Shen K, Tseng GC (2010) Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* 26: 1316–1323.
32. Yan Y, Stolz S, Chételat A, Reymond P, Pagni M, et al. (2007) A downstream mediator in the growth repression limb of the jasmonate pathway. *The Plant Cell* 19: 2470–2483.
33. von Malek B, van der Graaff E, Schneitz K, Keller B (2002) The Arabidopsis male-sterile mutant dde2-2 is defective in the ALLENE OXIDE SYNTHASE gene encoding one of the key enzymes of the jasmonic acid biosynthesis pathway. *Planta* 216: 187–192.
34. Draper J, Enot D, Parker D, Beckmann M, Snowdon S, et al. (2009) Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics* 10: 227.
35. Mueller LA, Zhang P, Rhee SY (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology* 132: 453–460.
36. Sclep G, Allemeersch J, Liechti R, De Meyer B, Beynon J, et al. (2007) CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. *BMC Bioinformatics* 8: 400.
37. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. *The Annals of Applied Statistics*: 107–129.
38. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57: 289–300.
39. Wasternack C, Hause B (2013) Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in *Annals of Botany*. *Annals of Botany* 111: 1021–1058.
40. Stumpe M, Feussner I (2006) Formation of oxylipins by cyp74 enzymes. *Phytochemistry Reviews* 5: 347–357.
41. Duan H, Huang MY, Palacio K, Schuler MA (2005) Variations in cyp74b2 (hydroperoxide lyase) gene expression differentially affect hexenal signaling in the columbia and landsberg erecta ecotypes of arabidopsis. *Plant Physiology* 139: 1529–1544.
42. Sonderby IE, Geu-Flores F, Halkier BA (2010) Biosynthesis of glucosinolates—gene discovery and beyond. *Trends in Plant Science* 15: 283–290.
43. Rice WR (1990) A consensus combined p-value test and the family-wide significance of component tests. *Biometrics*: 303–308.
44. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* 31: 68–71.

MarVis-Pathway: Integrative and Exploratory Pathway Analysis of Non-Targeted Metabolomics Data

The following paper was published 2014 in *Metabolomics* [106]. Supplementary material are available on <http://dx.doi.org/10.1007/s11306-014-0734-y>.

The sections on the biological interpretation of results from data analysis were written by Alexander Kaever, Dr. Kirstin Feussner, and Prof. Dr. Ivo Feussner in close collaboration. The MarVis-Pathway tool includes redesigned prototype functions [44, 45]. Figure 1 and 4 in the paper are partially based on drafts from Dr. Kirstin Feussner. The custom database (supplementary material 2), the table of additional metabolite hits from this database (supplementary material 3), and the MS/MS spectra of selected metabolites (supplementary material 4) as well as the corresponding method description were prepared by Dr. Kirstin Feussner and coworkers. The article was critically revised by all coauthors.

MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data

Alexander Kaefer · Manuel Landesfeind · Kirstin Feussner ·
Alina Mosblech · Ingo Heilmann · Burkhard Morgenstern ·
Ivo Feussner · Peter Meinicke

Received: 22 April 2014 / Accepted: 23 September 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract A central aim in the evaluation of non-targeted metabolomics data is the detection of intensity patterns that differ between experimental conditions as well as the identification of the underlying metabolites and their association with metabolic pathways. In this context, the identification of metabolites based on non-targeted mass spectrometry data is a major bottleneck. In many applications, this identification needs to be guided by expert knowledge and interactive tools for exploratory data analysis can significantly support this process. Additionally, the integration of data from other omics platforms, such as DNA microarray-based transcriptomics, can provide valuable hints and thereby facilitate the identification of metabolites via the reconstruction of related metabolic pathways. We here introduce the MarVis-Pathway tool, which allows the user to identify metabolites by annotation of pathways from cross-omics data. The analysis is supported by an extensive framework for pathway enrichment and meta-analysis. The tool allows the mapping of data set features by ID, name, and accurate mass, and can incorporate information from adduct and isotope correction of

mass spectrometry data. MarVis-Pathway was integrated in the MarVis-Suite (<http://marvis.gobics.de>), which features the seamless highly interactive filtering, combination, clustering, and visualization of omics data sets. The functionality of the new software tool is illustrated using combined mass spectrometry and DNA microarray data. This application confirms jasmonate biosynthesis as important metabolic pathway that is upregulated during the wound response of Arabidopsis plants.

Keywords Metabolomics · Metabolic fingerprinting · Mass spectrometry · Metabolic pathways · Set enrichment analysis · Transcriptomics

1 Introduction

Metabolomics studies (Dunn et al. 2013; Fiehn 2002) aim to identify and characterize all metabolites under specific experimental conditions, such as environmental or genetic perturbations or developmental stages (Tarpley et al. 2005; Nahlik et al. 2010; Watanabe et al. 2013; Bellaire et al. 2013; König et al. 2014). In this field, mass spectrometry (MS) coupled to gas or liquid chromatography (GC/MS and LC/MS) has become a key technology for detection, identification, and quantification of metabolites (Dunn et al. 2005). A typical non-targeted metabolomics experiment can be represented by a high-dimensional data matrix (Dettmer et al. 2007; Meinicke et al. 2008) comprising information on the identity of measured ion species (data set features) and intensities for each feature and sample. These intensities can be used as relative abundance measurements for the comparison of different samples or groups of samples. The features are characterized by means of the mass-to-charge (m/z) ratio, retention time (rt), and

Electronic supplementary material The online version of this article (doi:10.1007/s11306-014-0734-y) contains supplementary material, which is available to authorized users.

A. Kaefer (✉) · M. Landesfeind · B. Morgenstern ·
P. Meinicke
Department of Bioinformatics, Institute of Microbiology and
Genetics, Georg-August-University Göttingen, Goldschmidtstr.
1, 37077 Göttingen, Germany
e-mail: alex@gobics.de

K. Feussner · A. Mosblech · I. Heilmann · I. Feussner
Department of Plant Biochemistry, Albrecht-von-Haller-Institute
for Plant Sciences, Georg-August-University Göttingen,
Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

the respective intensity profiles. Data sets from other omics technologies, such as DNA microarray or RNA-seq-based transcriptomics (Brown and Botstein 1999; Mortazavi et al. 2008) and MS-based proteomics (Aebersold and Mann 2003), may be represented in a similar way. After pre-processing, the corresponding data set features, e.g. DNA microarray spots, can be identified with associated gene, protein, or transcript IDs. Similar to the non-targeted MS data from a metabolomics experiment, where a particular metabolite may be represented by multiple features standing for different isotopologues and adducts (Brown et al. 2009; Draper et al. 2009), a transcript may be associated with multiple spots containing specific DNA probes. The typical workflow in the analysis of omics data involves several steps for the identification and characterization of data set features that are relevant in a particular context. For this purpose, replicate samples for each experimental condition are statistically evaluated in order to identify features which show significant differences (Dudoit et al. 2002; Sugimoto et al. 2012; Kaever et al. 2012). In many applications, e.g. when analyzing time series, the experiments comprise more than two conditions and pre-processing results in large data sets of complex multivariate intensity profiles.

After detection of features, which significantly differ between conditions, the filtered data set can be analyzed by means of exploratory multivariate methods, such as clustering algorithms, principal, or independent component analysis (Eisen et al. 1998; Dettmer et al. 2007; Gürdeniz et al. 2013; Meinicke et al. 2008; Wijetunge et al. 2013) in order to identify prominent intensity patterns. Finally, annotations, e.g. in terms of metabolic pathways, may be used to explain or characterize particular groups of features in a functional context (Dahlquist et al. 2002; Suhre and Schmitt-Kopplin 2008). Pathway maps from public databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2012) and BioCyc (Caspi et al. 2012), contain information about metabolic reactions as well as the associated enzymes, genes, and metabolites, and can therefore interconnect almost all omics fields (Arakawa et al. 2005; Wägele et al. 2012). While the mapping of gene and protein IDs is in most cases straightforward, *m/z* ratios from non-targeted metabolomics experiments cannot be directly mapped to entries in the corresponding databases and the identification of metabolites is a major bottleneck in such experiments (Dunn et al. 2013; Scalbert et al. 2009). A common approach is to calculate putative monoisotopic masses and molecular formulas for all MS data set features and match these with known metabolites (Brown et al. 2011; Kuhl et al. 2012; Kaever et al. 2012; Lee et al. 2013). In order to identify relevant pathways, a popular approach is the Gene/Metabolite Set Enrichment Analysis (G/M SEA) and Over-

Representation Analysis (ORA) (Subramanian et al. 2005; Xia and Wishart 2010; Persicke et al. 2012; Khatri et al. 2012), where pathways are represented as sets of entries, e.g. metabolites in MSEA. The enrichment analysis aims to detect pathways which are enriched in significant or high-ranked features mapped to corresponding entries.

For the analysis of MS-derived metabolomics data, several web-based platforms have been published that cover all steps from preprocessing, data set management, statistical analysis, mapping of features to metabolic pathways, and enrichment analysis (Kessler et al. 2013; Xia et al. 2012; Kastenmüller et al. 2011; Wägele et al. 2012). Only recently, the stand-alone software MetaboNexus (Huang et al. 2014), which combines a workflow similar to the web-based platforms with the manual selection and database query of MS features, was introduced. MetaboNexus provides a browser-based user interface, but the analysis is performed on the local machine and without requiring the upload of data sets to a web server. In the context of DNA microarray analysis, software tools and libraries which allow the exploratory data analysis by means of cluster algorithms are available (Eisen et al. 1998; Saldanha 2004; Sturn et al. 2002; Hoon et al. 2004) and the methodology of GSEA and ORA was implemented in multiple packages (Huang et al. 2009; Ackermann and Strimmer 2009; Khatri et al. 2012). Powerful software suites, such as the TM4 platform (Saeed et al. 2003, 2006), allow the interactive and exploratory analysis of microarray data, e.g. the clustering and labeling of transcript profiles, in combination with ORA. In order to combine and integrate results from different omics platforms, many tools which focus on visualization, e.g. based on metabolic pathways, have been proposed (Gehlenborg et al. 2010; Thimm et al. 2004; Junker et al. 2006; Neuweger et al. 2009). Different platforms for the network-based visualization and analysis of metabolomics and transcriptomics data have been introduced (Gao et al. 2010; Landesfeind et al. 2014; Posma et al. 2014). The Cytoscape (Shannon et al. 2003) plug-in Metscape (Gao et al. 2010), for example, allows the extraction of pathway-specific subnetworks, the coloring of nodes according to intensities, and the animation of different condition-specific snapshots.

The MarVis-Suite tools (Kaever et al. 2009, 2012) were introduced for the extraction, clustering, and visualization of metabolic markers from data originating from non-targeted experiments. The MarVis-Suite thereby combines functionalities of previously described tools and platforms with the focus on three main themes: It provides highly interactive desktop user interfaces, e.g. for interactive inspection of data clusters, thus integrating the user's expert knowledge instead of generating static heatmap figures. For the analysis of data from non-targeted MS experiments, specialized functions are provided. These

tools are combined with more general functions that allow the straightforward integration of data sets from other omics platforms. In particular, the MarVis-Cluster interface provides a robust clustering based on one-dimensional self-organizing maps (1D-SOMs) (Meinicke et al. 2008), that is interactively used to investigate intensity patterns for a large number of multivariate feature profiles. Additionally, the MarVis-Filter interface features the adduct and isotope correction, filtering, and combination of multiple data sets, e.g. derived from positive and negative ionization mode. Several tools of the MarVis-Suite have been successfully applied for the identification of metabolite markers relevant in plant-pathogen-interaction (Djamei et al. 2011; Floerl et al. 2012; König et al. 2014) as well as for the characterization of mutants in lipid metabolism of *Arabidopsis* (König et al. 2012) and the COP9 signalosome of *Aspergillus* (Nahlik et al. 2010; Gerke et al. 2012).

In order to identify data set features in a functional context, we introduce the MarVis-Pathway tool, which allows the annotation and analysis of organism-specific pathways from the KEGG and BioCyc database collections in combination with an SEA meta-analysis framework for multi-omics data sets (Kaefer et al. 2014). The mapping of features to database entries is based on the matching of IDs, names, or accurate masses. MarVis-Pathway thereby completes the MarVis-Suite pipeline by providing a knowledge-based interpretation of results from explorative data analysis (see Fig. 1 for an overview on the interactive workflow). In addition, we introduce a signal-to-noise ratio-based ranking and filtering method for the MarVis-Filter tool, which features the statistical analysis of heterogeneous omics data based on minimal assumptions and which can be easily used for exploratory data analysis by modifying the signal definition. The proposed methods and tools are applied to data sets combining LC/MS with DNA microarray data in the context of a cross-omics study on the wound response of *Arabidopsis* plants, which represents a well-established model system. We show that the strength of MarVis-Pathway lies in the enhancement of analysis and interpretation of non-targeted LC/MS data sets in combination with transcriptomics data.

2 Materials and methods

2.1 Availability

Installation packages for the MarVis-Suite including MarVis-Pathway and a detailed handbook are available on the project homepage <http://marvis.gobics.de>. Data sets are available as comma separated values (CSV) files. Additionally, a detailed protocol of the corresponding data analysis within the MarVis-Suite and project files which

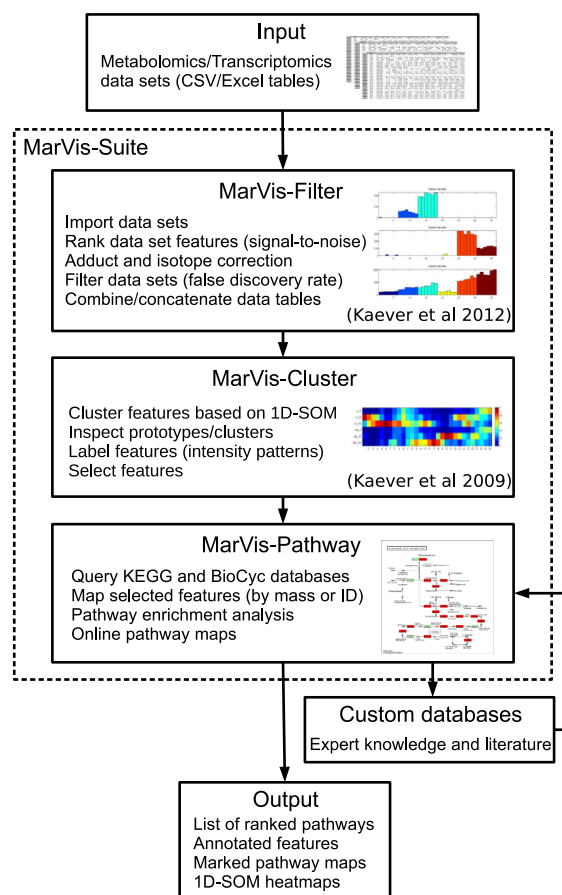


Fig. 1 Interactive workflow of data analysis within the MarVis-Suite

can be loaded directly into the MarVis-Suite interfaces (Load project function in the File menu) are provided.

2.2 Study, data sets, and preprocessing

The study investigates the wound reaction of *Arabidopsis thaliana* (ecotype Columbia-0) wild type (wt) and jasmonate-deficient *dde2-2* mutant plants (Malek et al. 2002) in a time course (control, 0.5 hours post wounding (hpw), 2 hpw) and comprises four metabolomics LC/MS and one transcriptomics DNA microarray data set generated from the same biological samples (see Table 1). The study was performed as described in (Mosblech et al. 2008; Meinicke et al. 2008; Kaefer et al. 2012). For each of six experimental conditions (wt: control, wt: 0.5 hpw, wt: 2 hpw, *dde2-2*: control, *dde2-2*: 0.5 hpw, *dde2-2*: 2 hpw), three biological replicate samples were analyzed with two platforms: The four metabolomics data sets derive from Ultra Performance Liquid Chromatography (UPLC) coupled to a Time-Of-Flight (TOF) MS analysis of the non-polar and

Table 1 Overview on data sets used for the integrative metabolome and transcriptome study of wild type and jasmonate-deficient *dde2-2* mutant plants in a time course of 0, 0.5, and 2 hours post wounding (6 conditions)

Data set label	Platform	Conditions/samples per condition	Extraction phase	Ionization mode	Features	Filtered features
M1	UPLC TOF-MS	6/3 ^a	Non-polar	Negative	2,272	316
M2	UPLC TOF-MS	6/3 ^a	Non-polar	Positive	5,980	313
M3	UPLC TOF-MS	6/3 ^a	Polar	Negative	4,023	161
M4	UPLC TOF-MS	6/3 ^a	Polar	Positive	10,421	234
T1	DNA microarray	6/3	–	–	38,825	2,809

The number of data set features/variables corresponds to the number of different ion species detected in MS analysis and the number of microarray spots (after discarding spots which were not assigned to a gene), respectively. The last column shows the number of retained features after signal-to-noise filtering ($FDR < 0.05$ in random permutation test, see Sect. 2.3)

^a The metabolomics data sets comprise two technical replicates per sample.

polar extraction phases in positive and negative ionization mode, respectively (see Table 1). For each sample, two UPLC TOF-MS runs (technical replicates) were performed, which resulted in six replicates per experimental condition. Data processing of the raw UPLC TOF-MS data (peak picking, peak alignment, and deisotoping) was performed with the MarkerLynx Application Manager for the MassLynx software (Waters Corporation, Milford, USA). For the DNA microarray analysis, the Agilent-021169 Arabidopsis 4 Oligo Microarray (V4) platform was used. Spots without gene assignment were left out and the expression values were quantile-normalized.

2.3 MarVis filter: data import, adduct correction, signal-to-noise filtering, and combination of data sets

The metabolomics and transcriptomics data sets were consecutively imported in MarVis-Filter (Kaefer et al. 2012; see also Raw data import function in the MarVis-Suite handbook) and processed (see Table 1; Fig. 1). In order to calculate accurate monoisotopic masses for all ion features in the MS data sets, the m/z values were corrected in MarVis-Filter based on different sets of rules for positive and negative ionization mode (mass tolerance 0.01 Da, rt tolerance 0.05 min) as described in (Kaefer et al. 2012). The features of each of the five data sets were filtered according to a signal-to-noise ratio (SNR) (He and Zhou 2008) in combination with 1000 random permutations of sample labels (assignments of samples to conditions) and a false discovery rate (FDR) (Benjamini and Hochberg 1995) threshold of 0.05 (see MarVis-Suite handbook), similar to the Significance Analysis of Microarrays (SAM) method introduced by Tusher et al. (2001). As part of the SNR calculation for each feature, the signal was defined as difference between the maximum and minimum average condition-specific intensity and the noise term was calculated as pooled sample standard deviation of intensity

values over all conditions. For the metabolomics data sets, which contain two technical replicates per biological sample, the FDRs were estimated by randomly permuting only the biological samples (see labeling of dependent replicates in the MarVis-Suite handbook). The technical replicates were always assigned to the condition label of the corresponding sample. This procedure allows to utilize the technical variation in the SNR score calculation without assuming independence of technical replicates, which usually show a high dependence. The intensities are not assumed to follow a specific distribution, e.g. the normal or log-normal distribution, which considerably extends the range of application and allows to filter heterogeneous data sets, e.g. metabolome and transcriptome data, within the same framework. Table 1 gives an overview on the number of features after filtering. For a customized SNR (see MarVis-Suite handbook), the signal may also be defined as the difference of the maximum/minimum/mean of average intensities for two subsets of conditions, e.g. comparing the maximum of condition 2 and 3 with the maximum over all other (control) conditions. Each filtered data set was stored in the MarVis-Filter clipboard. Finally, all filtered metabolomics and transcriptomics data sets were combined by concatenating the corresponding data tables (see MarVis-Suite documentation).

2.4 MarVis-Cluster: clustering, visualization, selection, and labeling of data set features

The combined data set was clustered and visualized in MarVis-Cluster (Kaefer et al. 2009; see also Goto MarVis-Cluster function in the MarVis-Suite handbook) using 30 prototypes/clusters for the training of the 1D-SOM. For clustering, the replicate intensities per condition and feature were averaged (arithmetic mean) and the resulting profile was normalized to unit Euclidean length. For each cluster, the proportion of metabolomics and transcriptomics features was visualized (see the Label barplot function in the

MarVis-Suite handbook). In order to label features which show higher intensities in the wt wounding-specific conditions compared to *dde2-2*, all features were selected and the selection was reduced by means of a customized SNR (see Sect. 2.3 and the MarVis-Suite handbook). For this purpose, the signal was defined as the difference between the maximum of the average intensities for condition 2 and 3 (wt: 0.5 hpw and wt: 2 hpw) and the maximum of all other conditions (wt: control and all conditions associated with *dde2-2*). The selection was reduced to all features with a ratio higher than 2 (1506 features) and labeled ('wt'). For the functional analysis in MarVis-Pathway, all features (labeled and unlabeled) were then selected (see Goto MarVis-Pathway function in the MarVis-Suite handbook).

2.5 MarVis-Pathway: database query, pathway enrichment, and meta-analysis

2.5.1 Pathway databases and feature mapping

MarVis-Pathway implements pathway databases from the KEGG and the BioCyc collection (Kanehisa et al. (2012); Caspi et al. (2012); see also Fig. 1). The included KEGG collection (KEGG FTP Release Dec 9, 2013, <http://www.kegg.jp>) contains one reference and about 3,000 organism-specific databases. The included BioCyc collection (biocyc-17.5, <http://biocyc.org>) provides about the same number of organism-specific databases and one reference database (MetaCyc). Each KEGG reference pathway is associated with a number of compound, EC (Enzyme Commission), and KO (KEGG ORTHOLOGY) IDs and names. Each MetaCyc reference pathway variant is associated with a number of compound and EC IDs/names. For all compounds in the databases, the monoisotopic masses were calculated based on the molecular formula. In case of the organism-specific databases, the pathways are associated with compound IDs, names, and masses and gene IDs/names instead of the EC and KO numbers. Additionally, customized databases may be loaded from comma separated values (CSV) files (see the MarVis-Suite handbook for details).

The features of the combined data set were mapped to metabolite and gene entries in the *A. thaliana*-specific pathways from KEGG and AraCyc (Mueller et al. 2003), which is part of the BioCyc database collection. The mapping of the features from the metabolomics data sets to metabolite entries was based on the corrected accurate masses (see Sect. 2.3) and a tolerance of 0.01 Dalton. The transcriptomics features were mapped to gene entries using the corresponding IDs.

2.5.2 Pathway enrichment analysis

For statistical analysis of pathways with matched entries, MarVis-Pathway provides an extensive framework for

(Gene/Metabolite) Set Enrichment Analysis (SEA) (Subramanian et al. 2005; Xia and Wishart 2010; Huang et al. 2009). The SEA framework in MarVis-Pathway offers three different types of enrichment analysis: Entry-based, marker/feature-based, and sample-based analysis. In the first case, the number of entries in a pathway matched by the selected features (in MarVis-Filter or MarVis-Cluster) in comparison to the number of entries which could be matched over all pathways is evaluated based on a hypergeometric distribution, similar to the ORA approach (Khatri et al. 2012) introduced by Draghici et al. (2003) and Hosack et al. (2003). When analyzing MS data sets, the metabolite entries are clustered according to their mass before performing the hypergeometric test in order to reduce the systematic dependence of database entries. In case of the marker/feature-based SEA, the analysis is based on the ranks of features (as calculated in MarVis-Filter) which match entries in a particular pathway, assuming independence of features. For statistical evaluation, a static or iterative hypergeometric test (Breitling et al. 2004), a rank-sum, or a Kolmogorov-Smirnov test is utilized. The method is able to incorporate information from adduct and isotope corrections performed in MarVis-Filter. In case of the sample-based SEA, the analysis is based on the ranks of features and a rank-sum or Kolmogorov-Smirnov test statistic which is recalculated for a large number of random permutations of sample condition labels, similar to the original GSEA method (Subramanian et al. 2005). For (re-)ranking, the SNR function is used. This method does not depend on the assumption of independent features or independent database entries but requires a sufficiently high number of replicate samples and considerably more computing time in comparison to the first two methods. As for the SNR permutation test, the labels of technical replicates of the same biological sample may be permuted together. The introduced methods for marker/feature-based and sample-based enrichment analysis use concepts of the Functional Class Scoring (FCS) approaches (Khatri et al. 2012). A detailed description of the implemented types of enrichment analysis can be found in the MarVis-Suite handbook.

2.5.3 Meta-analysis of multiple data sets

MarVis-Pathway offers a framework for the joint (entry, marker/feature, or sample-based) SEA of combined data sets. For this purpose, the pathway-specific *p*-values are first calculated for each data set separately in order to account for data set-specific properties, such as the number of features. Then, the *p*-values are merged per pathway in a meta-analysis (Kaever et al. 2014; Shen and Tseng 2010; Whitlock 2005) using Fisher's (Fisher 1925) or Stouffer's method (Stouffer et al. 1949) for independent data sets. In

case a sample-based enrichment analysis is performed, biological samples in different data sets may be linked and the condition labels are permuted together, e.g. a particular sample is always assigned the same condition label in all linked data sets. The linking option may also be combined with technical replicates belonging to independent biological samples. Finally, the FDRs are calculated (Benjamini and Hochberg 1995) based on the meta- p -values. In case a random permutation test is performed, the observed meta- p -value for a particular pathway is compared to the meta- p -values obtained for all pathways and all random permutations and the corresponding FDR is estimated (Tusher et al. 2001).

In order to identify relevant pathways in the study, entry, marker/feature, and sample-based enrichment analyses were performed. Global pathways with more than 500 associated entries, such as KEGG's unspecific metabolic pathways map, were left out in this analysis. In case of the entry and marker/feature-based analysis, the p -values were calculated based on a hypergeometric test and the initial filtering of the data sets (see Sect. 2.3). For meta-analysis, Fisher's method was used. In case of the sample-based analysis, a Kolmogorov-Smirnov test in combination with Fisher's method was used. The obtained meta- p -values were recalculated for 1,000 random permutations of sample labels, linking technical replicates within the data sets M1 to M4 and samples over all data sets.

2.6 MS/MS analysis

For unequivocal identification of metabolites, MS/MS spectra of MS features mapped to jasmonic acid (JA), jasmonoyl isoleucine (JA-Ile), 11/12-Hydroxy-JA, 12-Hydroxy-JA-Ile, and 12-Carboxy-JA-Ile were obtained by LC 1290 Infinity (Agilent Technologies, Santa Clara, CA, USA) coupled with a 6540 UHD Accurate-Mass Q-TOF-MS instrument (Agilent Technologies, Santa Clara, CA, USA) with Dual Jet Stream Technology as electrospray ionization (ESI) source (see Supplementary material 4). The analysis was performed in the negative ESI mode with minor modifications as described by Floerl et al. (2012).

3 Results and discussion

The plant's response to wounding is part of the defense against insects and is mainly regulated by the isoleucine conjugate of jasmonic acid JA-Ile (Howe and Jander 2008; Mosblech et al. 2009; Wasternack and Hause 2007; Wu and Baldwin 2010). During recent years, the corresponding defense pathway has been analyzed in detail in *Arabidopsis* and Tobacco. In the model plant *Arabidopsis* so far the

focus was on transcriptomics and proteomics experiments, comparing wounded wild type plants with JA-Ile biosynthesis or perception mutants (Stintzi et al. 2001; Reymond et al. 2004; Gfeller et al. 2011). Therefore, we used the JA-Ile-dependent wound response of *Arabidopsis* as an ideal experimental background to evaluate the functionality of MarVis-Pathway and the new MarVis-Suite.

3.1 Intensity profile clustering and visualization provides a convenient overview for combined cross-omics data set

The filtered transcriptomics and four metabolomics data sets were combined in MarVis-Filter (see Sects. 2.2, 2.3) and analyzed in MarVis-Cluster (see Sect. 2.4 and workflow in Fig. 1). Figure 2 shows the heatmap of prototypes (average cluster profiles) and the proportion of metabolomics and transcriptomics features within each cluster. The upper prototype plot provides a convenient overview on prominent intensity patterns and allows to interactively browse the clusters and select features. The first block of clusters (prototype 1–6) represents metabolomics and transcriptomics features with a profile specific for the wound response in wt plants. These features are therefore dependent on the biosynthesis of the signal molecule JA-Ile. However, a closer inspection revealed that also other clusters (e.g. cluster 7 and 8, see Fig. 2) harbor additional features being JA-Ile-dependent and showing a less prominent but significant difference. In order to mark all these wt-specific features in further analysis, they were labeled utilizing a customized SNR (see Sect. 2.4).

An important issue in the context of integrative analysis of metabolomics and transcriptomics time series data is the possible time lag between the different omics levels (Takahashi et al. 2011; Gibon et al. 2006). For example, transcripts may not be translated for a couple of hours resulting in a time shift of corresponding metabolite products. The heatmap visualization (see Fig. 2) supports the interactive analysis of different time frames, e.g. by means of the identification of blocks of clusters representing an early or late wound response (see cluster 1 and 2 or 3–6). However, the introduced functions focus on the visualization and interactive analysis of time-dependent intensity patterns and not on the calculation of time lags between different omics levels.

3.2 MarVis-Pathway facilitates the reconstruction and interactive analysis of metabolic pathways

For a functional interpretation, all metabolomics and transcriptomics features were selected and used for analysis in MarVis-Pathway (see Sect. 2.5). Based on the corrected monoisotopic masses (see Sect. 2.3) and gene

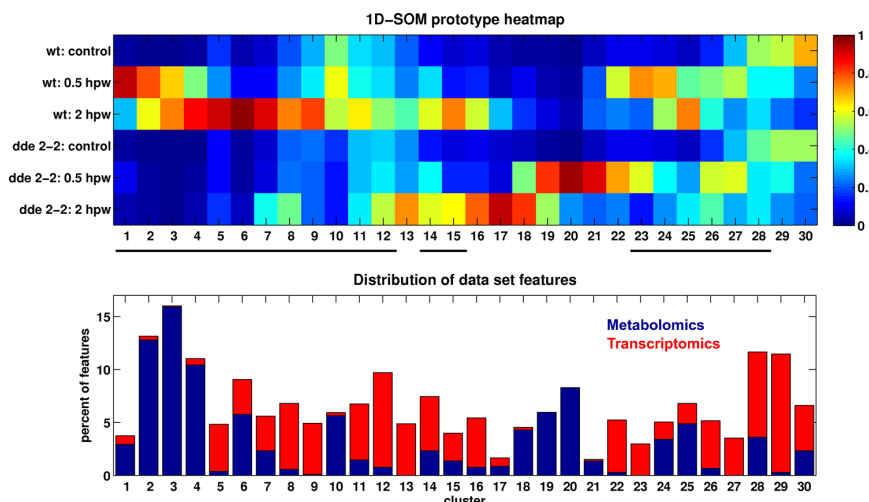


Fig. 2 Heatmap of ordered prototype profiles (average cluster profiles) from 1D-SOM clustering (*upper region*) and stacked bar plot of the distribution of data set features (*lower region*) for the combined metabolomics and transcriptomics data set. Blue bars in the lower plot indicate the percentage of features from the metabolomics

data sets (ion species) found in the corresponding cluster. Red bars show the percentage of transcriptomics features (*microarray spots*), respectively. Black lines between the prototype and bar plot mark clusters that contain features which were labeled as wt-specific by means of a customized SNR (see Sect. 2.4)

IDs, the features were mapped to entries in the *A. thaliana*-specific pathways from KEGG and AraCyc (Mueller et al. 2003).

Figure 3 shows a screenshot of MarVis-Pathway after database query together with a short description of the interactive user interface. Pathways which contain matched metabolites or genes can be interactively inspected and selected. For the selected pathway, the averaged and normalized intensity profiles of associated features (see Sect. 2.4) are visualized in a heatmap sorted according to the 1D-SOM order, which allows a convenient overview on intensity patterns. Interesting profiles can be interactively selected and mapped pathway entries inspected. Metabolite and gene entries associated with particular intensity profiles may be marked in a specific color, either by individual selection or based on previously defined labels of mapped data set features. The online resources associated with the selected pathway, e.g. the colored organism-specific KEGG pathway map, and the selected entry can be directly accessed in an additional browser window. In contrast to platforms focused on web-based interfaces (Kessler et al. 2013; Xia et al. 2012; Kastenmüller et al. 2011; Wägele et al. 2012), this approach splits the workflow into the exploratory analysis of multivariate intensity profiles by means of highly interactive desktop applications and the knowledge-based interpretation of results by means of the interconnected online resources of the KEGG and BioCyc databases. The central objective of MarVis-Pathway is the rapid detection of affected pathways that can be used as working hypotheses. This first reconstruction may be

followed by a more detailed network analysis of detected pathways using specialized tools (Gao et al. 2010; Landesfeind et al. 2014; Posma et al. 2014), e.g. by means of the visualization and expansion of pathway-specific subnetworks in the Metscape software (Gao et al. 2010). In contrast to the visualization of condition-specific network snapshots, MarVis-Pathway focuses on the pathway-specific heatmap visualization of multivariate intensity profiles, which allows a convenient overview on associated intensity patterns.

3.3 Enrichment analysis of metabolomics data sets identifies highly relevant pathways

In order to identify the most relevant pathways affected after wounding in wt and JA-deficient *dde2-2* plants, an enrichment analysis was performed in MarVis-Pathway. First, only the four metabolomics data sets (M1–M4, see Table 1) were used for analysis. Table 2A shows the top-ranked pathways and the FDRs calculated in the entry (E-SEA), marker/feature (M-SEA), and sample-based analysis (S-SEA) (see Sects. 2.5.2 and 2.5.3). The five top-ranked pathways (see Table 2A) are highly relevant in the context of plant wounding. The jasmonic acid biosynthesis (AraCyc, rank 2) and the alpha-linolenic acid metabolism (KEGG, rank 4) pathways describe the biosynthesis of JA-Ile. Additionally, pathways associated with glucosinolate biosynthesis, which is at least in major parts regulated by JA-Ile (Sønderby et al. 2010) and which constitutes a central defense reaction of *A. thaliana* plants

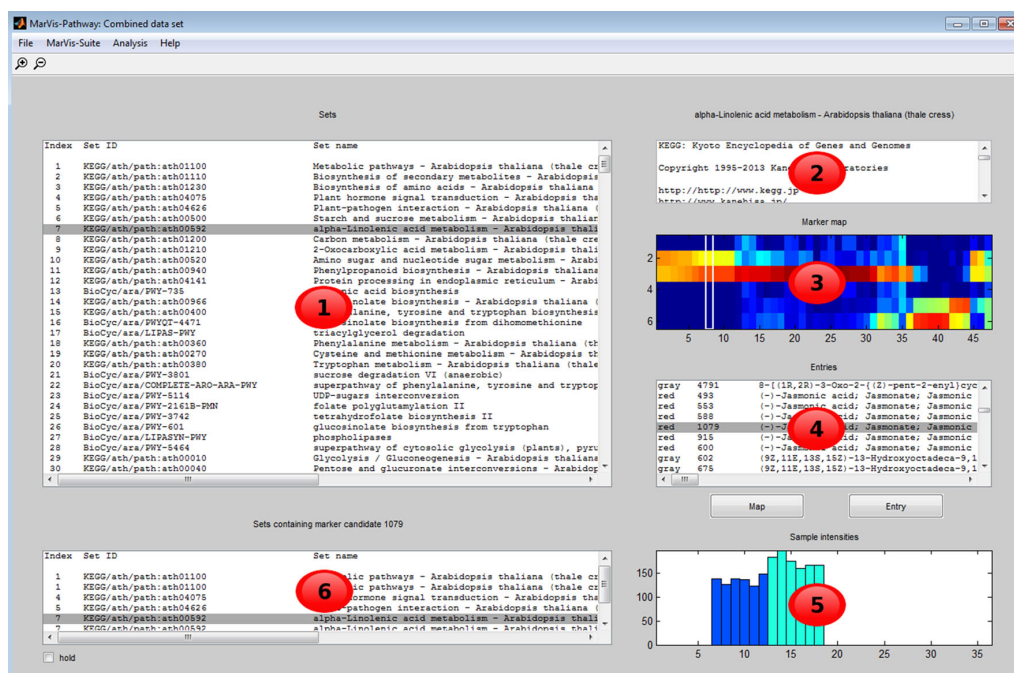


Fig. 3 Screenshot of the MarVis-Pathway interface after database query. The pathway list box (area 1) contains all matched pathways. The pathway information box (2) contains additional information about the flat files used for database construction. The marker profile map (3) shows the heatmap of feature profiles which could be mapped to the selected pathway. The entry assignment list box (4) contains the assignments of features to entries in the selected pathway. The marker profile plot (5) displays the raw intensity profile of the currently

selected feature. The related pathways list box (6) shows all pathways that contain entries mapped to the currently selected data set feature. Pathways, profiles, and entry assignments can be interactively inspected and selected. Via the Map and Entry button below the assignment list box (4), the online resources of the queried databases can be accessed, the marker color of particular entries may be interactively or automatically specified (only for KEGG pathways)

upon wounding, can be found in this list. For the relevant pathways, the FDRs calculated based on the M-SEA and E-SEA are much lower compared to the S-SEA. This is a direct result of the less conservative test assumptions (see Sect. 2.5.2). The data set features, e.g. different adducts of the same metabolite, or database entries, e.g. metabolites in the same pathway, are expected to show a systematic dependence (Subramanian et al. 2005; Barry et al. 2005). Nonetheless, the M-SEA is useful in order to identify pathways which contain entries that are matched by many significant features (see the jasmonic acid biosynthesis and alpha-linolenic acid metabolism pathways), indicating a correct adduct detection in preprocessing of non-targeted LC/MS data. However, this method also highlights pathways with a very low number of matched entries. The plant-pathogen interaction pathway (rank 3), that contains only one matched metabolite, JA, is an example for this case. The M-SEA and E-SEA methods require considerably less computing time in comparison to the random permutation-based S-SEA and can also be performed in case only a low number of replicate samples are available.

On the other hand, the S-SEA method allows to link dependent technical replicates and samples in the random permutation test and can therefore account for dependent data sets comprising measurements for the same samples (Kaefer et al. 2014). In case of the S-SEA based only on the metabolomics data sets, the estimated FDR for the important alpha-linolenic acid metabolism pathway is very high (0.807, rank 4). For most of the pathways, only a relatively small number of metabolites are matched by data set features.

3.4 Transcriptomics data significantly support the pathway analysis

The pathway enrichment analysis was repeated for the metabolomics (M1–M4) in combination with the transcriptomics (T1) data set. For the S-SEA, the sample labels in the metabolomics and transcriptomics data sets were linked (see Sect. 2.5.3). The enrichment analysis (see Table 2B) results in much lower estimated FDRs compared to the case where only the metabolomics data sets were used (see

Table 2 Top-ranked pathways from enrichment analysis based only on filtered/raw metabolomics data sets (part A), the combined metabolomics and transcriptomics data sets (B), and selected metabolomics and transcriptomics features showing a wt-constitutive intensity profile (C)

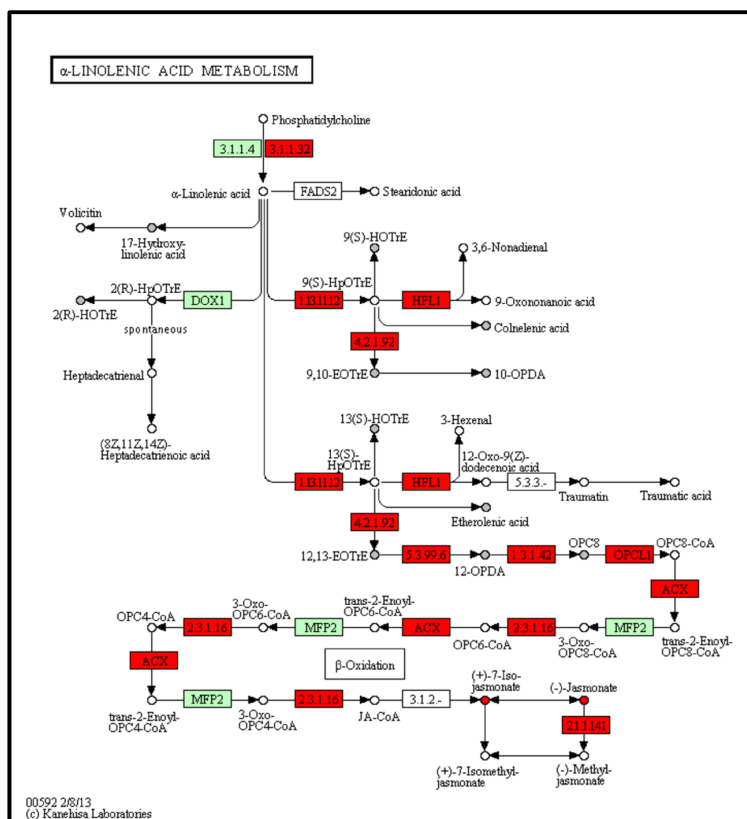
	DB	Pathway	F	M	G	M-SEA	E-SEA	S-SEA
(A) Pathway enrichment analysis of metabolomics data only								
1	KEGG	Plant hormone signal transduction	17	3	0	2.549e−06	0.005071	0.2475
2	AraCyc	Jasmonic acid biosynthesis	20	5	0	8.816e−08	0.09175	0.2475
3	KEGG	Plant–pathogen interaction	6	1	0	0.0004678	0.6159	0.357
4	KEGG	Alpha-Linolenic acid metabolism	20	13	0	2.479e−05	0.04789	0.807
5	AraCyc	Indole glucosinolate breakdown	9	4	0	0.1805	0.5675	0.8436
6	AraCyc	Heptaprenyl diphosphate biosynthesis	2	1	0	1	0.6825	0.8436
7	KEGG	Terpenoid backbone biosynthesis	2	1	0	1	1	0.8436
8	AraCyc	Glucosinolate biosynthesis from tryptophan	5	5	0	1	0.1445	0.8969
9	AraCyc	Glucosinolate biosynthesis from trihomomethionine	4	2	0	0.6825	0.9679	0.8969
10	KEGG	Sulfur relay system	2	1	0	0.2618	0.9679	0.8969
(B) Pathway enrichment analysis of metabolomics and transcriptomics data								
1	KEGG	Plant hormone signal transduction	55	3	34	2.18e−07	0.0001173	0.091
2	KEGG	Alpha-Linolenic acid metabolism	47	13	16	4.161e−18	6.228e−09	0.1363
3	KEGG	Plant–pathogen interaction	48	1	36	2.395e−09	1.13e−05	0.1363
4	AraCyc	Jasmonic acid biosynthesis	43	5	14	1.515e−15	0.0002319	0.1363
5	KEGG	Glucosinolate biosynthesis	24	12	9	0.0008703	1.001e−05	0.1578
6	KEGG	Fatty acid elongation	11	0	9	0.02568	0.01554	0.2113
7	AraCyc	Hydroxyjasmonate sulfate biosynthesis	3	0	2	0.01398	0.1264	0.2113
8	KEGG	Carotenoid biosynthesis	9	1	6	0.3106	0.341	0.3406
9	AraCyc	traumatoin and (Z)-3-hexen-1-yl acetate biosynthesis	13	0	6	2.783e−06	0.08044	0.4552
10	AraCyc	Glucosinolate biosynthesis from tryptophan	15	5	9	0.01398	0.0005411	0.5308
(C) Pathway enrichment analysis for selected wt-constitutive features								
1	AraCyc	Glucosinolate biosynthesis from tryptophan	5	5	0	0.02223	0.001129	–
2	AraCyc	Sulfate activation for sulfonation	2	0	2	0.002438	0.006558	–
3	KEGG	Tryptophan metabolism	5	3	1	0.1026	0.01164	–
4	KEGG	Glucosinolate biosynthesis	5	5	0	0.1088	0.02593	–
5	KEGG	Sulfur metabolism	2	0	2	0.01791	0.02744	–
6	KEGG	2-Oxocarboxylic acid metabolism	5	5	0	0.3276	0.1019	–
7	KEGG	Purine metabolism	2	0	2	0.1026	0.2686	–
8	AraCyc	Glucosinolate biosynthesis from homomethionine	2	1	1	0.3276	0.41	–
9	AraCyc	Glucosinolate breakdown	1	0	1	0.1672	0.4173	–
10	KEGG	Stilbenoid, diarylheptanoid and gingerol biosynthesis	2	2	0	0.4627	0.4173	–

The 4th, 5th, and 6th column contain the number of filtered/selected features over all data sets (F) which could be assigned to an entry in the corresponding pathway, the number of matched metabolites (M) in the corresponding pathway, and the number of matched genes (G). The last columns contain the estimated false discovery rates (FDRs) based on a marker/feature-based SEA (M-SEA), entry-based SEA (E-SEA), and sample-based SEA (S-SEA). The pathways are sorted according to the S-SEA (A, B) or E-SEA FDRs (C), respectively

Table 2A). Especially the E-SEA method is highly sensitive to the higher coverage of database entries due to the assigned transcript features (see alpha-linolenic acid metabolism pathway, rank 2). The alpha-linolenic acid metabolism pathway is also associated with a much lower FDR for the S-SEA method (0.1363) compared to the FDR estimated without the microarray data set (0.807). Figure 4a shows the corresponding colored KEGG pathway map. Entries (metabolites and genes) mapped to data set

features which were labeled as specific for the wounding of wt plants are marked in red. Entries mapped to features which are not associated with a wt-specific intensity profile are marked in gray. This pathway, which describes the jasmonate biosynthesis and contains the allene oxide synthase (AOS) enzyme (EC 4.2.1.92) that is missing in the *dde2-2* mutant, should be highly enriched in features showing significant differences between the experimental conditions and especially features with a wt-specific

(a) KEGG database query



(b) Query of custom database

JA metabolism (Göbel and Feussner 2009)	Oxidized Galactolipids (Ibrahim et al 2011)
JA	OPDA/dnOPDA-MGDG (Ara-A)
JA-Ile	OPDA/OPDA-MGDG (Ara-B)
OPDA	OPDA/dnOPDA-DGDG (Ara-C)
dnOPDA	OPDA/OPDA-DGDG (Ara-D)
OPC-4	OPDA/dnOPDA-MGDG-OPDA (Ara-E)
11/12-Hydroxy-JA	18:3/dnOPDA-MGDG (Ara-F)
11/12-Hydroxy-JA-Ile	OPDA/OPDA-MGDG-OPDA (Ara-G)
12-Carboxy-JA-Ile	OPDA/dnOPDA-MGDG-16:3
	...

Fig. 4 Results from database query in MarVis-Pathway. **a** The KEGG alpha-linolenic acid metabolism pathway with entries mapped to features from the filtered metabolomics and transcriptomics data sets. Entries exclusively mapped to labeled features, which are specific for the wounding of wt plants, are marked in red. Entries mapped to features which are not associated with a wt-specific intensity profile, e.g. because of the mapping of isomers with different intensity patterns to the same metabolite, are marked in gray. Green color indicates enzymes associated with *A. thaliana* genes which could not be mapped to features from the filtered transcriptomics data set. **b** Wt-specific feature hits from the query of a custom database

containing metabolites from the jasmonic acid (JA) metabolism and oxidized galactolipids described in literature. *10-OPDA* 10-oxo-11,15-phytydienoic acid, *12-OPDA* 12-oxo-10,15-phytydienoic acid, *9,10-EOTrE* 9,10-epoxyoctadecatrienoic acid, *12,13-EOTrE* 12,13-epoxyoctadecatrienoic acid, *OPC-8:0* 3-oxo-2-(pent-2'-enyl)-cyclopentane-1-octanoic acid, *9(S)-HOTrE* 9-hydroxyoctadecatri-10,12,15-enoic acid, *13(S)-HOTrE* 13-hydroxyoctadeca-9,11,15-trienoic acid, *2(R)-HOTrE* 2-hydroxyoctadecatri-9,12,15-enoic acid, *JA-Ile* jasmonoyl isoleucine, *dnOPDA* 10-oxo-8,13-dinor-phytydienoic acid, *OPC-4* 3-oxo-2-(pent-2'-enyl)-cyclopentane-1-butanoic acid, *DGDG* digalactosyl diacylglycerol, *MGDG* monogalactosyl diacylglycerol

profile. From the metabolomics point of view, only the jasmonate is clearly associated with wt-specific ion features. All other matched metabolites (gray points) are not exclusively associated with labeled features due to isomers and ambiguous mass matching (see mapping table in Supplementary material 1). By means of the mapping of the filtered microarray data set, the coverage of pathway entries is significantly increased, as represented by much lower FDRs in enrichment analysis, and the wt-specific enzymatic steps towards the biosynthesis of jasmonate are clearly highlighted (see the lower branch of the pathway). Notably, all but two mapped transcript features are labeled as wt-specific (see mapping table in Supplementary material 1).

The integration of the transcriptomics data set has a strong effect on the estimated FDRs. However, the microarray data do not bias the overall pathway ranking. In both cases, when analyzing only the metabolomics (see Table 2A) or transcriptomics data (see Supplementary material 7), the highly relevant alpha-linolenic acid metabolism, the plant hormone signal transduction, and glucosinolate-related pathways can be found in the list of top-ranked candidates. In addition, the introduced methods for integrative enrichment and meta-analysis do not depend on the estimation of a time lag between data from different omics platforms (Takahashi et al. 2011). The introduced analysis is based on the ranking of data set features according to general differences between the experimental conditions or the selection of features associated with particular intensity patterns.

3.5 Custom databases expand pathway analysis

The analysis based on KEGG and AraCyc pathways resulted in a relatively small number of metabolite annotations (see Table 2A) because many precursors and derivatives of jasmonic acid as well as related compound classes, such as oxidized galactolipids, are not yet represented in these databases. In order to integrate expert and literature knowledge, MarVis-Pathway provides an interface to import custom databases in CSV format, containing additional entries (e.g. metabolites, genes, or enzymes) and assignments to pathways or arbitrary sets/groups of related entries, such as compound classes (see workflow in Fig. 1 and MarVis-Suite handbook). For data analysis in this study on plant wounding, a custom database containing previously described metabolites (Göbel and Feussner 2009; Ibrahim et al. 2011) was created (see custom database in Supplementary material 2). This database was used for annotating additional metabolic features based on the corrected masses (see Fig. 4b and the table of additional metabolite hits in Supplementary material 3). By this means, 22 highly context-related metabolites could be

assigned to features which exclusively accumulated in wt plants after wounding. These JA-Ile-dependent wound-induced features are represented by prototypes 1 to 6 after clustering by 1D-SOM (see Fig. 2). As proof of concept, five putative metabolite hits, including JA and JA-Ile as well as the JA-derivatives described as degradation products or transport forms, 12-hydroxy-JA, 12-hydroxy-JA-Ile, and 12-carboxy-JA-Ile, were confirmed by MS/MS analysis (see MS/MS spectra in Supplementary material 4).

In the following, we will describe two further examples how the new MarVis-Suite tools support the exploratory analysis and context-related identification of data set features.

3.6 Pathway analysis of selected clusters identifies glucosinolates as JA-Ile-dependent metabolites with wt-constitutive intensity pattern

The prototype heatmap for the combined cross-omics data set (see Fig. 2) shows a number of other interesting intensity patterns. For example, cluster 10 contains features with a wt-constitutive pattern characterized by very small differences between the wt conditions and zero or very low average intensities for the mutant-associated conditions. For further analysis, the cluster was selected in MarVis-Cluster and only the associated features were imported and analyzed in MarVis-Pathway. Table 2C shows the results of marker and entry-based enrichment analysis. Interestingly, most of the top-ranked pathways are associated with glucosinolate biosynthesis (see mapping table in Supplementary material 5). Though, only a small number of features match entries in these pathways.

3.7 Customized SNR ranking detects *dde2-2*-constitutive intensity profiles

In contrast to the wt-constitutive intensity pattern, the prototype heatmap (see Fig. 2) does not reveal intensity profiles with a *dde2-2*-constitutive pattern. However, there may be a small number of corresponding features hidden in one of the more prominent clusters. Therefore, the whole cross-omics data set was re-ranked in MarVis-Filter utilizing a signal-to-noise ratio with customized signal term (see Sect. 2.3), the difference between the minimum over the average intensities of the *dde2-2*-associated conditions 4–6 and the maximum over the average intensities of the wt-associated conditions 1–3. Interestingly, only two of the 2,809 filtered transcriptomics data set features, ambiguously associated with At1g53490 and At1g53480, could be found with a ratio greater than 2 (see expression profiles in Supplementary material 6). These two microarray spots show high expression levels for the *dde2-2*-associated conditions independent of the wounding and may be an

interesting starting point for further studies on the *dde2-2* mutant.

4 Concluding remarks

The MarVis-Suite combines a statistical framework with highly interactive interfaces for exploratory data analysis. Data sets from different omics platforms can be filtered, combined, clustered, and visualized. By means of the new MarVis-Pathway interface, filtered or selected data set features may be annotated in the context of organism-specific pathway databases or custom pathway/entry set definitions which represent expert knowledge. The signal-to-noise ratio allows the ranking and filtering of heterogeneous data sets within a common framework and can easily be customized for the search for particular intensity patterns. The framework allows many other options, including alternative ratios, e.g. the signal-to-level ratio, or moderation/shrinkage of the noise term (Smyth 2004; Allison et al. 2006). By means of the enrichment analysis, annotated pathways can be statistically evaluated based on different assumptions, e.g. independence of features, database entries, or samples. Additionally, MarVis-Pathway provides functions for the meta-analysis of pathway enrichment for multiple data sets. The tools were successfully applied in a cross-omics study on plant wounding. The integration of transcriptomics data significantly supported the analysis of the non-targeted metabolomics data sets. Additionally, proteomics data can be integrated for a more comprehensive analysis.

Acknowledgments This work was supported by the German Federal Ministry of Education and Research (BioFung 0315595A) and the German Research Foundation (FL3 INST186/822-1, He3424/2). We thank Lennart Opitz, Claudia Pommerenke, and Gabriela Salinas-Riester (DNA Microarray and Deep-Sequencing Facility Göttingen) for microarray analysis and Pia Meyer for excellent assistance.

Conflict of interest The authors declare that they have no conflict of interest.

Compliance with Ethical Standards This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10, 47.
- Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422, 198–207.
- Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1), 55–65.
- Arakawa, K., Kono, N., Yamada, Y., Mori, H., & Tomita, M. (2005). KEGG-based pathway visualization tool for complex omics data. *Silico Biology*, 5(4), 419–423.
- Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, 21(9), 1943–1949.
- Bellaire, A., Ischebeck, T., Staedler, Y., Weinhaeuser, I., Mair, A., Parameswaran, S., et al. (2013). Metabolism and development-integration of micro computed tomography data and metabolite profiling reveals metabolic reprogramming from floral initiation to silique development. *New Phytologist*, 202, 322–335.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1), 289–300.
- Breitling, R., Amtmann, A., & Herzyk, P. (2004). Iterative group analysis (iga): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5(1), 34.
- Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C., Francis-McIntyre, S., et al. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134(7), 1322–1332.
- Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., et al. (2011). Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics*, 27(8), 1108–1112.
- Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21, 33–37.
- Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., et al. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1), D742–D753.
- Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., & Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, 31(1), 19–20.
- Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78.
- Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., et al. (2011). Metabolic priming by a secreted fungal effector. *Nature*, 478(7369), 395–398.
- Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., & Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, 81(2), 98–104.
- Draper, J., Enot, D., Parker, D., Beckmann, M., Snowdon, S., Lin, W., et al. (2009). Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics*, 10, 227.
- Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1), 111–140.
- Dunn, W. B., Bailey, N. J., & Johnson, H. E. (2005). Measuring the metabolome: Current analytical technologies. *Analyst*, 130(5), 606–625.
- Dunn, W. B., Erban, A., Weber, R. J., Creek, D. J., Brown, M., Breitling, R., et al. (2013). Mass appeal: Metabolite

- identification in mass spectrometry-focused metabolomics. *Metabolomics*, 9(1), 44–66.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25), 14,863–14,868.
- Fiehn, O. (2002). Metabolomics-the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Floerl, S., Majcherzyk, A., Possienke, M., Feussner, K., Tappe, H., Gatz, C., et al. (2012). *Verticillium longisporum* infection affects the leaf apoplastic proteome, metabolome, and cell wall properties in *Arabidopsis thaliana*. *PLoS One*, 7(2), e31435.
- Gao, J., Tarcea, V. G., Karnovsky, S. I., Baliga, B. R., Weymouth, T. E., Beecher, C. W., et al. (2010). Metscape: A Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics*, 26(7), 971–973.
- Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., et al. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7, S56–S68.
- Gerke, J., Bayram, Ö., Feussner, K., Landesfeind, M., Shelest, E., Feussner, I., et al. (2012). Breaking the silence: Protein stabilization uncovers silenced biosynthetic gene clusters in the fungus *Aspergillus nidulans*. *Applied and Environmental Microbiology*, 78(23), 8234–8244.
- Gfeller, A., Baerenfaller, K., Loscos, J., Chételat, A., Baginsky, S., & Farmer, E. E. (2011). Jasmonate controls polypeptide patterning in undamaged tissue in wounded arabidopsis leaves. *Plant Physiology*, 156(4), 1797–1807.
- Gibon, Y., Usadel, B., Blaessing, O. E., Kamlage, B., Hoehne, M., Trethewey, R., et al. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. *Genome Biology*, 7(8), R76.
- Göbel, C., & Feussner, I. (2009). Methods for the analysis of oxylipins in plants. *Phytochemistry*, 70(13–14), 1485–1503.
- Gürdeniz, G., Hansen, L., Rasmussen, M. A., Acar, E., Olsen, A., Christensen, J., et al. (2013). Patterns of time since last meal revealed by sparse PCA in an observational LC-MS based metabolomics study. *Metabolomics*, 9(5), 1073–1081.
- He, Z., & Zhou, J. (2008). Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. *Applied and Environmental Microbiology*, 74(10), 2957–2966.
- de Hoon, M. J., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9), 1453–1454.
- Hosack, D. A., Dennis, G. Jr, Sherman, B. T., Lane, H. C., & Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10), R70.
- Howe, G., & Jander, G. (2008). Plant immunity to insect herbivores. *Annual Review of Plant Biology*, 59, 41–66.
- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), 1–13.
- Huang, S. M., Toh, W., Benke, P. I., Tan, C. S., & Ong, C. N. (2014). MetaboNexus: An interactive platform for integrated metabolomics analysis. *Metabolomics*. doi:10.1007/s11306-014-0648-8.
- Ibrahim, A., Schütz, A., Galano, J., Herrfurth, C., Feussner, K., Durand, T., et al. (2011). The alphabet of galactolipids in *Arabidopsis thaliana*. *Frontiers in Plant Physiology*, 2, 95.
- Junker, B. H., Klukas, C., & Schreiber, F. (2006). VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1), 109.
- Kaever, A., Lingner, T., Feussner, K., Göbel, C., Feussner, I., & Meinicke, P. (2009). MarVis: A tool for clustering and visualization of metabolic biomarkers. *BMC Bioinformatics*, 10, 92.
- Kaever, A., Landesfeind, M., Possienke, M., Feussner, K., Feussner, I., & Meinicke, P. (2012). MarVis-Filter: Ranking, filtering, adduct and isotope correction of mass spectrometry data. *Journal of Biomedicine and Biotechnology*. doi:10.1155/2012/263910.
- Kaever, A., Landesfeind, M., Feussner, K., Morgenstern, B., Feussner, I., & Meinicke, P. (2014). Meta-analysis of pathway enrichment: Combining independent and dependent omics data sets. *PLoS One*, 9(2), e89297.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1), D109–D114.
- Kastenmüller, G., Römisch-Margl, W., Wägele, B., Altmaier, E., & Suhre, K. (2011). metaP-Server: A web-based metabolomics data analysis tool. *Journal of Biomedicine and Biotechnology*. doi:10.1155/2011/839862.
- Kessler, N., Neuweger, H., Bonte, A., Langenkämper, G., Niehaus, K., Nattkemper, T. W., et al. (2013). MeltDB 2.0-advances of the metabolomics software system. *Bioinformatics*, 29(19), 2452–2459.
- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2), e1002375.
- König, S., Feussner, K., Schwarz, M., Kaever, A., Iven, T., Landesfeind, M., et al. (2012). *Arabidopsis* mutants of sphingolipid fatty acid α -hydroxylases accumulate ceramides and salicylates. *New Phytologist*, 196(4), 1086–1097.
- König, S., Feussner, K., Kaever, A., Landesfeind, M., Thurow, C., Karlovsky, P., et al. (2014). Soluble phenylpropanoids are involved in the defense response of *Arabidopsis* against *Verticillium longisporum*. *New Phytologist*, 202(3), 823–837.
- Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., & Neumann, S. (2012). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1), 283–289.
- Landesfeind, M., Kaever, A., Feussner, K., Thurow, C., Gatz, C., Feussner, I., et al. (2014). Integrative study of *Arabidopsis thaliana* metabolomic and transcriptomic data with the interactive MarVis-Graph software. *PeerJ*, 2(e239).
- Lee, T. S., Ho, Y. S., Yeo, H. C., Lin, J. P. Y., & Lee, D. Y. (2013). Precursor mass prediction by clustering ionization products in LC-MS-based metabolomics. *Metabolomics*, 9(6), 1301–1310.
- von Malek, B., van der Graaff, E., Schneitz, K., & Keller, B. (2002). The *Arabidopsis* male-sterile mutant *dde2-2* is defective in the ALLENE OXIDE SYNTHASE gene encoding one of the key enzymes of the jasmonic acid biosynthesis pathway. *Planta*, 216(1), 187–192.
- Meinicke, P., Lingner, T., Kaever, A., Feussner, K., Göbel, C., Feussner, I., et al. (2008). Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology*, 3, 9.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628.
- Mosblech, A., König, S., Stenzel, I., Grzeganeck, P., Feussner, I., & Heilmann, I. (2008). Phosphoinositide and inositolpolyphosphate signalling in defense responses of *Arabidopsis thaliana* challenged by mechanical wounding. *Molecular Plant*, 1(2), 249–261.
- Mosblech, A., Feussner, I., & Heilmann, I. (2009). Oxylipins: Structurally diverse metabolites from fatty acid oxidation. *Plant Physiology and Biochemistry*, 47(6), 511–517.
- Mueller, L. A., Zhang, P., & Rhee, S. Y. (2003). AraCyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiology*, 132(2), 453–460.
- Nahlik, K., Dumkow, M., Bayram, Ö., Helmstaedt, K., Busch, S., Valerius, O., et al. (2010). The COP9 signalosome mediates

- transcriptional and metabolic response for hormones, oxidative stress protection and cell wall rearrangement during fungal development. *Molecular Microbiology*, 78, 964–979.
- Neuweger, H., Persicke, M., Albaum, S. P., Bekel, T., Dondrup, M., Hüser, A. T., et al. (2009). Visualizing post genomics data-sets on customized pathway maps by ProMeTra-aeration-dependent gene expression and metabolism of *Corynebacterium glutamicum* as an example. *BMC Systems Biology*, 3(1), 82.
- Persicke, M., Rückert, C., Plassmeier, J., Stutz, L. J., Kessler, N., Kalinowski, J., et al. (2012). MSEA: metabolite set enrichment analysis in the MeltDB metabolomics software platform: Metabolic profiling of *Corynebacterium glutamicum* as an example. *Metabolomics*, 8(2), 310–322.
- Posma, J. M., Robinette, S. L., Holmes, E., & Nicholson, J. K. (2014). MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG. *Bioinformatics*, 30(6), 893–895.
- Reymond, P., Bodenhausen, N., Van Poecke, R. M., Krishnamurthy, V., Dicke, M., & Farmer, E. E. (2004). A conserved transcript pattern in response to a specialist and a generalist herbivore. *The Plant Cell*, 16(11), 3132–3147.
- Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: A free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2), 374–378.
- Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., et al. (2006). TM4 microarray software suite. *Methods in Enzymology*, 411, 134–193.
- Saldanha, A. J. (2004). Java Treeview-extensible visualization of microarray data. *Bioinformatics*, 20(17), 3246–3248.
- Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B. S., van Ommen, B., et al. (2009). Mass-spectrometry-based metabolomics: Limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5(4), 435–458.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
- Shen, K., & Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10), 1316–1323.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 3.
- Sønderby, I. E., Geu-Flores, F., & Halkier, B. A. (2010). Biosynthesis of glucosinolates—gene discovery and beyond. *Trends in Plant Science*, 15(5), 283–290.
- Stintzi, A., Weber, H., Reymond, P., & Farmer, E. E. (2001). Plant defense in the absence of jasmonic acid: The role of cyclopentenones. *PNAS*, 98(22), 12,837–12,842.
- Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). *The American soldier: Adjustment during army life*. Princeton, NJ: Princeton University Press.
- Sturn, A., Quackenbush, J., & Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1), 207–208.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43), 15,545–15,550.
- Sugimoto, M., Kawakami, M., Robert, M., Soga, T., & Tomita, M. (2012). Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Current Bioinformatics*, 7(1), 96–108.
- Suhre, K., & Schmitt-Kopplin, P. (2008). MassTRIX: Mass translator into pathways. *Nucleic Acids Research*, 36(suppl 2), W481–W484.
- Takahashi, H., Morioka, R., Ito, R., Oshima, T., Altaf-Ul-Amin, M., Ogasawara, N., et al. (2011). Dynamics of time-lagged gene-to-metabolite networks of *Escherichia coli* elucidated by integrative omics approach. *OMICS*, 15(1–2), 15–23.
- Tarpley, L., Duran, A., Kebrom, T., & Sumner, L. (2005). Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period. *BMC Plant Biology*, 5, 8.
- Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6), 914–939.
- Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9), 5116–5121.
- Wägele, B., Witting, M., Schmitt-Kopplin, P., & Suhre, K. (2012). MassTRIX reloaded: Combined analysis and visualization of transcriptome and metabolome data. *PLoS One*, 7(7), e39,860.
- Wasternack, C., & Hause, B. (2013). Jasmonates: Biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in *Annals of Botany*. *Annals of Botany*, 111(6), 1021–1058.
- Watanabe, M., Balazadeh, S., Tohge, T., Erban, A., Giavalisco, P., Kopka, J., et al. (2013). Comprehensive dissection of spatio-temporal metabolic shifts in primary, secondary, and lipid metabolism during developmental senescence in *Arabidopsis*. *Plant Physiology*, 162(3), 1290–1310.
- Whitlock, M. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, 18(5), 1368–1373.
- Wijetunge, C. D., Li, Z., Saeed, I., Bowne, J., Hsu, A. L., Roessner, U., et al. (2013). Exploratory analysis of high-throughput metabolomic data. *Metabolomics*, 9(6), 1311–1320.
- Wu, J., & Baldwin, I. T. (2010). New insights into plant responses to the attack from insect herbivores. *Annual Review of Genetics*, 44, 1–24.
- Xia, J., & Wishart, D. S. (2010). MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(suppl 2), W71–W77.
- Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Research*, 40(W1), W127–W133.

Applications of the MarVis-Suite

The tools of the MarVis-Suite have been applied in many experimental studies. In the following, selected applications are summarized.

In a book chapter by Alexander Kaefer et al. [107], the principle of clustering of ion features by means of 1D-SOMs was reviewed in the context of two previous applications [12, 108] and a detailed protocol for data analysis in MarVis-Cluster [41] was described. Furthermore, the combination of data sets derived from MS analysis of polar and non-polar extracts measured in positive and negative ionization mode was introduced. The section *Applications of the Technology* and the corresponding tables were drafted by Alexander Kaefer, Dr. Kirstin Feussner, and Prof. Dr. Ivo Feussner and all coauthors critically revised the chapter. The figures are conceptually based on graphics from the corresponding publications and figure 9 was created by Dr. Kirstin Feussner.

In [109], metabolic fingerprints of *Arabidopsis thaliana* mutants of sphingolipid fatty acid α -hydroxylases were compared to wild type plants using MarVis-Filter, MarVis-Cluster, and a prototype version of MarVis-Pathway. In [110], the MarVis-Suite tools were applied in order to analyze non-targeted MS data in the context of priming of *A. thaliana* against the fungal pathogen *Plectosphaerella cucumerina*. In [111] and [112], data sets in the context of infection of *Arabidopsis* with the plant-pathogenic fungus *Verticillium longisporum* were analyzed. In the latter case, the data sets were derived from RNA-seq Transcriptomics and MS-based Proteomics measurements. The protein profiles were filtered in MarVis-Filter based on a signal-to-level ratio (see MarVis-Suite handbook 11.1). The transcript candidates were ranked and filtered based on a combination of the edgeR [113] tool, the DESeq [114] package, and a moderated Chi-squared test, which is based on the same shrinkage method used for the moderated signal-to-noise ratio (see MarVis-Suite handbook 11.1). Alexander Kaefer implemented the moderated Chi-squared test and conducted the analysis of RNA-seq data using the implemented test, the edgeR, and DESeq packages (see Methods S1 in the corresponding Supporting

Information). The respective description (see sections *RNA-Seq data analysis* and *Moderated Chi-squared test* in Methods S1) was drafted by Alexander Kaever. In the context of all described applications, Alexander Kaever provided the MarVis-Suite tools and hands-on training, supported data analysis, and revised the manuscripts.

In the evaluation and application of the recently published MarVis-Graph software [115], MarVis-Filter was used for adduct/isotope correction and filtering of Metabolomics and Transcriptomics data sets as preprocessing for the graph-based analysis of metabolic reaction-chains in the context of *Arabidopsis* wounding experiments. Alexander Kaever performed the preprocessing of data with MarVis-Filter, contributed conceptually to the development of the permutation test for detected sub-networks, and revised the manuscript.

Discussion

In this work, a statistical framework for the analysis of non-targeted Metabolomics MS data sets in combination with data from other omics platforms was introduced. The proposed methods and tools were implemented in the MarVis-Suite, which provides interactive user interfaces for exploratory analysis of omics data. The original MarVis-Cluster tool [41] was extended in order to support the visualization of combined cross-omics data sets and the labeling of selected data set features (see chapter 6). The MarVis-Filter interface (see chapter 4 and 6) includes functions for the adduct and isotope correction of LC/MS data, molecular formula prediction, statistical ranking, filtering, and combination of multiple data sets. The MarVis-Pathway tool (see chapter 6) allows the mapping of data set features to pathway databases and provides functions for enrichment analysis in order to identify highly relevant pathways in the experimental context. An important function is the combination of multiple cross-omics data sets in a meta-analysis (see chapter 5 and 6). The meta-analysis of pathway enrichment thereby forms the top level of data analysis in the proposed framework.

The tools of the MarVis-Suite were evaluated in the context of a cross-omics case study on wounding of *Arabidopsis thaliana* (see section 3.4) and successfully applied in many studies for the detection of relevant data set features and associated metabolites and metabolic pathways (see chapter 7). In the studies on plant wounding, the integration of Transcriptomics data significantly supported the analysis and interpretation of the non-targeted Metabolomics data. In the first application in this context (see chapter 5), non-targeted LC/MS data sets were combined with publicly available DNA microarray data sets, which were obtained from independent biological samples. In the second application (see chapter 6), Metabolomics data were integrated with a Transcriptomics data set obtained from the same biological samples, which allowed the cluster analysis of correlated metabolite and transcript features.

8.1 Adduct correction and annotation of LC/MS data

A central requirement for the functional annotation of ion features in LC/MS data is the adduct and isotope correction, which calculates potential accurate masses for the feature m/z ratios. Based on these masses, the features can be mapped to metabolites in public or custom databases or molecular formulas can be predicted (see section 3.1). The correction of m/z ratios thereby provides the basis for putative identification and hypothesis generation. The annotated metabolites may then be confirmed by means of the coelution with authentic standards or MS/MS fragmentation patterns [6].

The adduct and isotope correction in MarVis-Filter (see chapter 4) is based on predefined ionization rules $[xm + y]^{z[+/-]}$ [21, 42, 43]. These rules describe the building blocks of an ionization product, e.g. the number of target molecules (x) or charges (z) and the addition of other molecules (y), and allow the calculation of potential masses (m). Additionally, MarVis-Filter detects carbon-13 isotopes and corrects for the difference between carbon-13 and carbon-12 to calculate accurate monoisotopic feature masses (see section 3.3). For each ion feature and hypothetical combination of ionization rule and number of included carbon-13 isotopes (feature annotation), the introduced algorithm calculates a score based on other features in the data set that support this hypothesis. A particular feature annotation is thereby supported by another ion feature if, for any allowed annotation of the latter feature, both calculated masses are equal within a given tolerance. Then, the annotation with the highest score is chosen for each feature and used for the prediction of the corresponding mass. The score for a particular annotation is calculated as the sum of cosine similarities between the intensity profile of the feature to be annotated and the supporting candidates. Additionally, supporting candidates are restricted to features within a predefined rt tolerance. This concept is based on the observation that in LC/MS analysis a metabolite species is usually represented by several ion features with similar rt and intensity profile, but different m/z ratios, representing the different possibilities of ionization and number of isotopes (see section 3.1).

Besides machine and vendor-specific tools, such as the deisotoping function in the MassLynx software (Waters Corporation), which removes common isotopologues in a data set, and the Molecular Feature Extractor in the MassHunter software (Agilent Technologies Corporation), many free software packages and workflows for the correction of MS features in LC/MS data have been published in recent years [24, 46, 47, 116]. Similar to the adduct and isotope correction in MarVis-Filter, these methods utilize the exact mass difference between adduct rules, a retention time tolerance, and a similarity measure between intensity profiles in order to annotate ion features. But in contrast to MarVis-Filter, most of these methods also group

features that putatively derive from the same metabolite. For this purpose, the CAMERA package [47] also uses the correlation coefficients between chromatographic peak shapes of related ion features, which have to be extracted from the raw data and which are not available in the MarVis-Suite. The clustering of potentially related MS features has the advantage that the data set size can be significantly reduced [46] and that the systematic dependence of features may thereby be reduced. Furthermore, the number of distinct metabolites represented in a particular data set can be estimated more directly. On the other hand, the clustering of MS features may be erroneous, grouping features representing the same metabolite species into different clusters, and result in an incorrect mass prediction. For clustering and visualization of the intensity profiles, the merging of feature profiles in the same group can be problematic because the resulting profiles may significantly differ from the original feature profiles. One of the key principles in the design of the original MarVis-Cluster tool was the convenient clustering and visualization of intensity profiles with as little as possible preprocessing, still allowing the manual inspection of adduct formations in particular clusters and visualization as rt - m/z plot [12, 41]. Therefore, the annotation and not the merging of data set features is the central objective of the adduct and isotope correction in MarVis-Filter.

For comparison of the results obtained by means of the adduct correction in MarVis-Filter, the M1 data set from chapter 6, containing UPLC TOF-MS features for the non-polar extraction phase analyzed in negative ionization mode, was preprocessed with the XCMS software [23], using recommended parameter settings from [117], and annotated using the CAMERA and the AStream [46] packages (results not shown). The CAMERA package requires peak detection and sample alignment by XCMS and the AStream package expects data containing isotopologues. In comparison to the deisotoped M1 data set, which contains about 2000 ion features, the XCMS preprocessing resulted in a data set of about 3000 features, including isotopologues. The 1D-SOM clustering of both data sets revealed overall similar intensity profiles.

After annotation of the XCMS data set with CAMERA and AStream using default and modified parameter settings, the data set was searched for the exact masses of central metabolites (see section 3.4 and chapter 6 figure 4). In case of CAMERA and default options for negative ionization mode, the exact mass of jasmonic acid was not correctly predicted, despite the presence of related adducts. When using the fixed set of three ionization rules applied in chapter 4, 5, and 6 (see table 2 in chapter 4), the features representing jasmonic acid were correctly annotated by CAMERA. However, the accurate mass of 12-OPDA was not correctly predicted using the default options or the three described rules, despite the presence of the corresponding adduct peaks in the XCMS data set. When deactivating the default peak shape correlation option, the mass of 12-OPDA and related adducts were correctly annotated. Apart from that,

the query of the custom database described in chapter 6 with the corrected feature masses based on the CAMERA annotations showed a large overlap of hits with the results from the MarVis-Filter-based processing of data set M1.

The AStream package seems to have been tested only on LC/MS data from positive ionization mode because the specification of the $[m + H]^+$ rule is mandatory for adduct search. After extending the corresponding function, allowing $[m - H]^-$ as reference rule for negative ionization mode, the package was applied to the XCMS data set. In case of the default options as well as modified parameters, the mass of 12-OPDA was not correctly annotated and the query of the custom database resulted in much less hits.

All in all, the three tested packages XCMS, CAMERA, and AStream provide powerful and highly specialized methods for automatic processing of LC/MS data but also include many (hyper-)parameters, e.g. a peak shape correlation threshold in CAMERA, that significantly influence the results. In contrast, MarVis-Filter provides a less complex function for adduct and isotope correction embedded in a highly interactive user interface. By means of the visualization capabilities of MarVis-Cluster, e.g. the cluster-specific rt - m/z plot, errors in the adduct and isotope correction may be detected by means of the user's expert knowledge.

Nonetheless, data sets preprocessed with the described packages may also be imported in MarVis-Filter or MarVis-Cluster, using the predicted masses instead of the original m/z values.

Instead of the mass-based mapping of features from non-targeted MS data sets to metabolites, a more sophisticated method may be used. After calculating potential accurate masses (first step), molecular formulas can be predicted for each mass (second step) and used for database query [15, 24]. In order to reduce the number of potential formulas, especially for large masses, heuristic chemical rules and information about the isotope distribution can be utilized [48]. This method is able to reduce the number of false-positive database hits, e.g. if the number of carbon atoms predicted for the observed isotopic pattern does not match the corresponding number for a particular compound. On the other hand, the prediction of formulas and numbers of included atoms can be erroneous. Especially for large masses, the prediction results in multiple formulas [6] and the automatic filtering can be problematic. Therefore, the MarVis-Suite tools provide a function for molecular formula prediction only for single selected data set features (see chapter 4 and MarVis-Suite handbook 11.1), e.g. for features associated with highly interesting intensity profiles that could not be mapped to any metabolite in the available databases.

Based on the mass correction in MarVis-Filter described in chapter 4, 5, and 6, metabolic features specific for the wounding of wild type in comparison to jasmonate-deficient *dde2-2*

mutant plants [101] could be associated with oxylipins in the alpha-linolenic acid metabolism pathway, mono- and di-galactosyldiacylglycerols, and glucosinolates. Additionally, further jasmonic acid derivatives, hormones, and poly-hydroxy fatty acids could be identified in [115]. Ion features with accumulating intensity for *Arabidopsis* mutants of sphingolipid fatty acid α -hydroxylases could be identified as ceramides, glucosylceramides, salicylic acid (SA), and SA-derived metabolites [109]. In the context of priming of *Arabidopsis* against the fungal pathogen *Plectosphaerella cucumerina* [110], compounds from the primary metabolism of sugars were detected. Additionally, three central metabolites in the priming fingerprint could be identified as indole-3-carboxylic acid, hypoxanthine, and galacturonic acid. In the context of the defense response of *Arabidopsis* against the fungus *Verticillium longisporum* [111], metabolites derived from the phenylpropanoid pathway were identified as infection-induced and confirmed in targeted analysis.

In these applications, the calculation of accurate masses based on different sets of ionization rules for positive and negative ionization mode was an important step for metabolite identification and also for combining the data sets. This combination of data sets obtained from different extraction phases analyzed in positive or negative ionization mode is essential since many identified metabolites could only be detected in one of the corresponding data sets.

8.2 Statistical ranking and filtering of intensity profiles

In order to detect data set features which represent actual differences between the experimental conditions instead of noise, the associated intensity profiles are ranked and filtered by means of statistical methods. The ranking of features thereby allows to focus the following data analysis on high-ranked candidates without losing the context of other features in the data set. The strict filtering of features can simplify the user-driven interactive data analysis, e.g. by reducing the data set size by a factor of more than 10 (see chapter 6).

For this purpose, several tests, e.g. the Kruskal-Wallis and ANOVA test [25], in combination with methods for multiple testing correction were implemented in MarVis-Filter (see chapter 4) and applied in various studies (see chapter 7). In all these studies, context-related metabolites could be detected based on high-ranked feature intensity profiles. The filtering by means of an error rate threshold, e.g. the false discovery rate (FDR) [30] or familywise error rate (FWER) [29], allowed to significantly reduce the data set sizes for initial analysis and hypothesis generation. Besides filtering, the ranking of whole data sets could be used for global pathway enrichment analysis (see chapter 5 and 6). While the strict filtering of data sets allows the rapid interactive data analysis, e.g. in MarVis-Cluster, the ranking of whole data sets and

the following enrichment analysis facilitates a broader and more thorough statistical analysis (see section 8.4).

In many applications, the non-parametric Kruskal-Wallis test was used for ranking and filtering of MS data set features [13, 108, 110, 111]. This test utilizes the ranks of intensities for a given feature profile in order to detect differences between the experimental conditions and does not assume a particular distribution of intensities, e.g. the normal or log-normal distribution. Therefore, it is well suited for testing profiles containing not normally-distributed outlier intensities or replaced missing values, which can result from an erroneous peak detection and alignment of LC/MS data [22]. However, the rank-based test is less powerful compared to the parametric ANOVA [25, 75], assuming normally or log-normally distributed intensities. Thus, ANOVA filtering was applied in a number of studies [109, 118], too.

Especially for the statistical analysis of large DNA microarray data sets, the Kruskal-Wallis test is too conservative and results, after correction for multiple testing, in a very low number of retained features. For example, performing the Kruskal-Wallis test on the Transcriptomics data from chapter 6 results in no features below an FDR threshold of 0.05 (results not shown).

In order to replace the Kruskal-Wallis test with a more sensitive method that does not assume a particular distribution of intensities, a signal-to-noise ratio (SNR) [119] based framework was developed (see chapter 5 and 6). The SNR here is defined as the ratio of the difference between condition-specific average intensities (signal) and the pooled sample standard deviation of intensities within the conditions (noise), similar to Hedges's effect size estimator [120]. In analogy to the significance analysis of microarrays (SAM) [27], the SNR is calculated for each feature, the observed ratios are compared to values obtained in random permutations of sample labels (assignments of samples to conditions), and the FDRs or FWERs are estimated per feature. Since the random permutation test makes no assumptions about the intensity/expression level distribution, it has proven to be useful for combining heterogeneous omics data (see chapter 6).

The SNR definition can be easily customized, e.g. in the form of the signal-to-level ratio (see MarVis-Suite handbook 11.1) or for selecting condition-specific intensity patterns (see labeling of wt-specific features or search for mutant-specific constitutive intensity profiles in chapter 6). In case, a data set contains intensity measurements for technical/analytical replicates of biological samples [14, 121, 122], which usually show a high systematic dependence, these replicates may be permuted together and not independently. This allows to utilize the information from analytical replicates (technical variation) within the ratio calculation without extending the test assumptions, which are independent biological samples and random assignments of condition

labels. In case only measurements for a small number of biological samples are available, the labels of technical replicates may also be permuted independently.

In comparison to the SAM method, the implemented default SNR definition does not include a fudge factor, a small constant which is added to the noise term. This factor is meant to stabilize the ratio calculation for very small noise values. By adding a small constant to the noise term in the denominator of the ratio, profiles containing very small intensities are automatically scored with relatively low SNRs. However, the fudge factor value has to be estimated and depends on the technological platform used for data acquisition [27, 75, 123]. A very small value would not have much effect while a large factor would strongly discriminate low intensity profiles, which may represent important but not highly abundant metabolites/transcripts. Especially in the context of RNA-seq analysis, the problem of detecting differences for low-abundant or short transcripts has been discussed [124, 125]. Therefore and because MarVis-Filter provides general methods for ranking and filtering of intensity profiles, the introduced SNR framework does not utilize a fudge factor by default. In case an extremely small or even zero noise term is calculated for a particular profile with a non-zero signal, the ratio is set to infinity and the corresponding feature is top-ranked. However, this phenomenon was not observed for the analyzed MS and microarray data sets.

In case the noise term should be stabilized, the framework provides a function for noise moderation (see MarVis-Suite handbook 11.1) based on a shrinkage method [74, 126, 127]. Furthermore, a constant fudge factor could be easily added to a customized SNR definition (see SNR macro definition in 11.1) or profiles containing only very low intensities could be filtered out (see intensity-level based ranking in 11.1).

The SNR ranking and filtering described in chapter 6 was performed on raw MS intensities and (quantile-normalized) expression values. This is in accordance with the default clustering and heatmap visualization in MarVis-Cluster based on averaged and linearly scaled raw intensity profiles [41]. MarVis-Filter also offers the option to apply a logarithm function to all intensities (log-transformation), which is performed by default in most tools for microarray data analysis [127], before calculating the SNRs. The log-transformation can thereby significantly reduce the noise for large intensities but also discriminates small values. Overall, the introduced tests and options for ranking of intensity profiles show very similar results (average feature rank correlation coefficient above 0.9, results not shown). This corresponds to the observation that parametric and non-parametric tests result in a large overlap of significant features [25] and that the feature-specific scoring/ranking method does not have a large effect on the following enrichment analysis [83].

8.3 Combination of multi-omics data sets, clustering, and visualization

In order to extract and detect as many metabolites as possible in non-targeted LC/MS experiments, different classes of metabolites are extracted separately and analyzed in positive and negative ionization mode (see section 3.1). This results in multiple data sets comprising measurements for the same samples and experimental conditions. The identification of metabolites associated with detected ion features represented in these data sets is a major challenge [6], which becomes manifest in the ambiguous matching of accurate masses or predicted formulas to metabolites (see section 3.3). Additionally, many context-specific metabolites are not yet represented in public databases (see chapter 6). The integration of data from other omics platforms, such as DNA microarrays, which allow a more reliable mapping of microarray spot IDs to organism-specific gene IDs, is a promising approach to cope with this challenge (see chapter 5). After preprocessing, data sets from both platforms can be represented in the form of a matrix, which comprises the intensity measurements for different features, representing ion species or microarray spots (see chapter 6).

MarVis-Filter features a simple but powerful interface for the combination of data sets, which can be used to combine MS-derived data sets, e.g. analyzed in positive and negative ionization mode (see chapter 4), or Metabolomics and Transcriptomics data (see chapter 6). The corresponding data matrices are concatenated (see MarVis-Suite handbook 11.1), retaining all original feature intensity profiles and annotations. The assignment of each feature to the original data set is stored in additional fields, which are utilized for example in meta-analysis of pathway enrichment (see section 8.5). In case the data sets contain measurements for the same experimental conditions, the features of the combined data set, e.g. ion features from positive and negative ionization mode and transcript features, can be used for 1D-SOM clustering based on the averaged condition-specific intensity profiles. In combination with a bar plot of the distribution of original data set features per cluster, this results in a highly convenient visualization of the combined data set (see figure 2 in chapter 6). In interactive data analysis, information/annotations about transcripts/genes may be used in order to interpret metabolic features in the same or neighboring clusters (see chapter 6), assuming that the clustering of similar intensity/expression profiles may represent related functional groups [31].

For the integrative analysis of multi-omics data sets, principal component, correlation, and hierarchical cluster analysis as well as related methods have been introduced and applied [71, 128, 129, 130, 131, 132, 133, 134]. In the context of LC/MS data analysis, the 1D-SOM

clustering resulted in a more convenient data interpretation in comparison to the loading plot from principal component analysis and a more robust clustering and visualization compared to hierarchical clustering combined with the K-means algorithm [12] (see also section 3.2). In a study on nutritional stresses in *Arabidopsis* [135], Transcriptomics and Metabolomics data sets were clustered separately by means of 2D-SOMs and identified intensity patterns were compared. By means of the *omeSOM tool [136], Transcriptomics and Metabolomics data can be clustered and visualized together based on 2D-SOMs. For each node in the 2D-grid visualization, the marker color indicates the association of only transcript features, only metabolite features, or the combination of both, while the marker size indicates the number of associated features. In comparison, a distinct advantage of the 1D-SOM prototype visualization in MarVis-Cluster is the direct overview on complex multivariate intensity profiles (see upper plot in figure 2 in chapter 6) in combination with a bar plot of the distribution of data set features (see lower plot).

Another very popular approach to visualization and analysis of cross-omics data [70] is the projection onto networks or sub-networks [69] and many tools have been developed for this purpose [137, 138, 139]. In this context, the vertices and edges of the corresponding graph may be annotated and colored according to the intensity profiles of mapped transcript or metabolite features. The mapping of data set features to metabolic pathway maps [140, 141, 142] (see following section) can be seen as sub-category of this type of analysis, though genome-scale networks [143] also comprise the links between different pathways and can be used for the analysis of metabolic fluxes [144] and complex protein interactions [145]. However, for interactive and user-driven data analysis, large networks can be complex to overview and the analysis of individual, simplified pathway maps may be better suited. The recently published MarVis-Graph tool [115] (see also chapter 7) provides a convenient interface for the analysis of Metabolomics and Transcriptomics data, e.g. after preprocessing in MarVis-Filter and MarVis-Cluster, based on a graph/network representation instead of the focus on separate pathways in MarVis-Pathway.

8.4 Pathway annotation, visualization, and enrichment analysis

In order to cope with the challenge of ambiguous mass matching of ion features to metabolites, the biochemical context, e.g. corresponding metabolic pathways, plays an important role (see section 3.3). The mapping of multiple features to different metabolites in the same pathway

thereby provides more confidence in the identification of corresponding metabolites. This context can be further backed up by the mapping of data set features from other omics platforms, e.g. DNA microarray-based Transcriptomics, to corresponding pathway entries (see chapter 5 and 6).

In this context, the introduced MarVis-Pathway tool provides functions for the mapping of selected data set features to entries of metabolic pathways or arbitrary sets defined in custom databases (see chapter 6). The mapping is either based on accurate masses, e.g. obtained from the adduct and isotope correction of LC/MS data in MarVis-Filter, or feature IDs, e.g. in the form of gene identifiers for DNA microarray data. The mapping of features to metabolic pathways, which describe the relationships between metabolites, metabolic reactions, associated enzymes, transcripts, and genes, thereby allows to link information from many omics fields, e.g. Metabolomics, Transcriptomics, and Proteomics. Therefore, the visualization concept based on the annotation and coloring of pathway entries, especially for the KEGG database [50], has been implemented in various tools [62, 87]. MarVis-Pathway provides interfaces for the coloring/marketing of entries in a similar way. A major difference to other tools is that MarVis-Pathways allows to interactively color single entries, e.g. associated with a specific intensity profile, or perform the coloring based on feature labels, e.g. assigned in MarVis-Cluster based on intensity patterns (see figure 4 in chapter 6).

In order to identify highly relevant pathways for a particular experiment, methods for the statistical evaluation of pathways based on the concept of (gene/metabolite) set enrichment analysis (G/M SEA) [81, 85] were implemented and applied. In chapter 5 and 6, the enrichment analysis resulted in the top-ranking of well-known pathways in the context of plant wounding. The implemented methods are either based on the assumption of independent data set features or database entries, similar to the concept of overrepresentation analysis [79, 80, 83], or on the assumption of independent samples, as introduced in the original GSEA method [81]. In case of the marker/feature-based SEA, the pathway-specific p-values and corresponding FDRs or FWERs are expected to be biased, since features representing the same biological entity, e.g. different ionization products representing the same metabolite species, show a systematic correlation [42, 47], independent of the experimental context. A similar dependence is expected for entries associated with the same pathway [81, 146, 147]. For this reason, the pathway-specific p-values in the application described in chapter 5 were restandardized based on the sample mean and standard deviation of transformed p-values (normal deviates, see figure 1 in chapter 5). This is a conservative procedure, since the observed normal deviates also include values for pathways that should be detected in enrichment analysis. In addition, the restandard-

ization procedure depends on a sufficiently high number of matched pathways for the mean and standard deviation estimation. The mapping of strictly filtered data sets or even single clusters though results in only few entry hits (see chapter 6) and therefore also a relatively low number of matched pathways. In this case, the restandardization of p-values is not recommended.

However, the ranking of pathways based on the FDRs from entry or marker/feature-based SEA, which requires considerably less computing time compared to the sample-based approach, is still very useful in order to identify relevant pathways. Especially the marker/feature-based SEA, which is able to integrate information from the adduct and isotope correction in MarVis-Filter, focuses on pathways associated with metabolites that are matched by multiple ion features representing different ionization products. These multiple hits indicate a correct adduct detection and increase the confidence in the predicted mass [47]. Additionally, these methods can also be applied if the number of available samples is very small.

In contrast, the sample-based approach [81] is based only on the assumption of independent samples (see chapter 6) and can be used for a more stringent analysis of unfiltered data sets, requiring considerably more computing time. As for the SNR random permutation test (see section 8.2), the labels of technical/analytical replicates may be permuted together and in this case only biological replicate samples are assumed to be independent. In the context of the random permutation-based SEA of DNA microarray data, highly specialized and powerful methods have been developed [82, 147]. In contrast, the implemented rank-based methods are more general and can be applied to arbitrary data set rankings or feature selections. In this context, the results on simulated data set correlation (see chapter 5) are also transferable to other methods for the ranking of data set features.

As extension to the analysis of pathways represented as simplified sets of entries, a network-assisted approach to enrichment analysis was recently introduced [148]. The EnrichNet approach utilizes the connections between genes or proteins within an interaction network [149] in order to allow a more comprehensive enrichment analysis, e.g. by taking into account differentially expressed genes which are not directly associated with a particular pathway but are in close network proximity to associated genes. This concept was partially integrated into the MarVis-Graph tool [115].

The reconstruction of metabolic pathways by means of MarVis-Pathway facilitated the interpretation of non-targeted Metabolomics experiments and the identification of corresponding metabolites in several applications. In chapter 5 and 6, highly relevant pathways for the wound response of *Arabidopsis* plants were investigated. The phenylpropanoid biosynthesis was identified as a central pathway induced upon infection with *Verticillium longisporum* [111]. Metabolic pathways associated with the primary metabolism of sugars were detected as

overrepresented in the context of priming of *Arabidopsis* against *Plectosphaerella cucumerina* [110].

8.5 Meta-analysis of pathway enrichment

In chapter 5 and 6, different methods for the meta-analysis of pathway enrichment for multiple data sets were introduced and applied. In MS-based Metabolomics experiments, multiple data sets arise from the extraction of different classes of metabolites, e.g. from the polar or non-polar extraction phase, and analysis in positive or negative ionization mode (see section 3.1). Furthermore, the integration with data from other omics platforms increases the number of data sets and the combination of results becomes an essential step in data interpretation.

In the context of MS-based non-targeted Metabolomics, the metaXCMS package [150] and a detailed protocol for the summary of overlapping features in multiple comparisons of conditions [117] were introduced. For the statistical evaluation of results from related studies, methods for meta-analysis [88, 151] have been developed and applied to the analysis of independent DNA microarray data [90, 91, 92, 93] in order to extract genes which are differentially expressed under particular conditions combining the evidence from multiple data sets. In contrast to the metaXCMS approach, most methods for meta-analysis first summarize the test statistics across multiple studies before performing a combined test, which results in a list of significant genes. By this means, also genes with weak differential expression may be detected if the corresponding expression pattern can be observed in multiple studies.

Fisher's method [152] was used to identify significant genes in multiple studies on prostate cancer and these genes were used to query the KEGG pathway database [153]. Meta-analysis on gene level has also been combined with enrichment or overrepresentation analysis in the context of cancer studies [154, 155] and gene set scores from enrichment analysis of different microarray data sets were combined in a meta-analysis based on hierarchical clustering [156]. Furthermore, it was pointed out that the testing of gene sets improves the comparability of different microarray data sets in comparison to gene-wise analysis [157].

For the meta-analysis of pathway enrichment, a framework was described in [94] and [158], specialized on the analysis of independent microarray data sets. The authors introduced two different approaches: MAPE_G, which summarizes gene-specific test statistics and then performs an enrichment analysis based on the summarized values [154, 155], and MAPE_P, which combines the test statistics from enrichment analysis on the pathway level. In the context of combining multi-omics data sets, where single features represent ion species or microarray spots and cannot be directly summarized across data sets, a MAPE_P-like approach, which

combines the data sets on the pathway level, is required (see chapter 5 and 6). However, methods assuming independent data sets, such as MAPE_P, are not directly applicable because the MS-derived Metabolomics data sets are expected to be dependent (see section 3.1). Furthermore, the focus on DNA microarray data does not allow the integration of special characteristics of the MS-derived data sets, such as the mapping of multiple ion species to a particular metabolite.

Therefore, a general framework for the estimation of data set correlations and meta-analysis of multi-omics data based on the results from pathway enrichment analysis was developed (see chapter 5). After correlation estimation, which indicated a relatively low dependence of Metabolomics data sets, the pathway-specific p-values were combined using an extended version of Fisher's [159] or Stouffer's (normal) method [160].

In the context of combining results from pathway enrichment analysis using general rank or selection-based tests (see section 8.4), p-value-based methods for meta-analysis [90, 161, 162], such as Fisher's or Stouffer's method, provide a more general framework in comparison to specialized effect size-based approaches [88, 89]. In contrast to meta-analysis using the minimum or n^{th} smallest observed p-value [163, 164], Fisher's and Stouffer's method integrate the p-values for all data sets into the summary statistic and were therefore implemented in the framework described in chapter 5.

In the application, the extended version of Fisher's method (Brown's method) seemed to be more powerful for pathways associated with low p-values for only some of the data sets. In contrast, Stouffer's method resulted in lower FDRs for pathways associated with comparably small p-values in case of all data sets. This is in agreement with the results presented in [162]. For the analysis of non-targeted Metabolomics data, e.g. containing measurements for metabolites that can be extracted either in the polar or non-polar phase or that can be detected either in the positive or negative ionization mode (see chapter 4), Fisher's/Brown's method seems to be the better choice. However, this depends on the objective of the meta-analysis, detection of pathways showing a consensus enrichment for all or most data sets or also pathways which are enriched for only a proportion of the data sets [161, 165].

Parallel to the described framework for meta-analysis of pathway enrichment (see chapter 5), the iPEAP [166] platform has been developed. iPEAP allows to combine the results from different methods for pathway enrichment analysis and data from Transcriptomics, Proteomics, and Metabolomics studies. The platform focuses on the evaluation and aggregation of pathway rankings obtained for different methods and data sets. Importantly, Metabolomics data set features have to be associated with compound IDs for the pathway mapping (a requirement

for many published tools). Especially in non-targeted LC/MS-based studies, the mapping of ion features to distinct metabolite IDs is a central challenge and cannot easily be solved (see section 3.3 and 8.1). In order to cope with this challenge, the highly interactive MarVis-Suite was developed, which covers many preprocessing steps before pathway enrichment analysis and allows to integrate the user's expert knowledge into the process of data analysis.

In chapter 6, an alternative method for the meta-analysis of dependent data sets, whose dependence arises from analysis of the same biological samples, based on the sample-based SEA was introduced. In this approach, the observed meta-p-values are calculated per pathway using Fisher's or Stouffer's method and recalculated for a large number of random permutations of sample labels, similar to the MAPE_P approach. But as essential part of the meta-analysis, the labels of samples in different data sets are linked during random permutations. By this means, a particular biological sample is always assigned the same condition label in all data sets. This approach is computationally much more expensive compared to the direct calculation of meta-p-values. On the other hand, the p-values do not have to be restandardized and no data set correlations have to be estimated (see chapter 5). In both applications (chapter 5 and 6), the integration of Transcriptomics data significantly supported the analysis and interpretation of non-targeted Metabolomics data, which is indicated by much lower FDRs and a higher entry coverage for relevant pathways (see figure 4 in chapter 6).

As recommendation, the marker/feature-based and entry-based SEA (see chapter 6) should be used in combination with Fisher's method for fast ranking of pathways and hypothesis generation in case the data sets were filtered or a small proportion of features were selected, e.g. based on the clustering in MarVis-Cluster. The marker/feature-based SEA thereby focuses on pathways associated with multiple significant feature profiles, e.g. representing different adducts of the same metabolite, while the entry-based analysis primarily detects pathways showing a high entry coverage. The sample-based SEA in combination with the linking option or the marker/feature-based SEA in combination with p-value restandardization and data set correlation estimation (see chapter 5) should be used for a more thorough statistical analysis of whole data sets. The sample-based SEA is thereby computationally much more expensive compared to the marker/feature-based SEA but allows to omit the conservative restandardization and the correlation estimation, which both require a sufficiently high number of annotated pathways. The marker/feature-based, entry-based, and sample-based SEA methods are complementary and should be combined in a comprehensive data analysis (see chapter 6).

8.6 Workflow and platforms

In this work, the MarVis-Suite, which provides highly interactive user interfaces enclosed in a statistical framework for the analysis of non-targeted Metabolomics data, was developed. The introduced workflow (see figure 1 in chapter 6) is straightforward and allows to easily integrate data from other omics platforms. For this purpose, generally applicable and robust methods, e.g. based on the ranking of data set features, were developed.

In recent years, similar workflows were implemented in web-based platforms [62, 63, 64, 65]. Most of these platforms utilize different R-packages [167] and provide tools for data analysis without the need to install additional software on the local machine. However, the data sets have to be uploaded to the corresponding server. In comparison, the MarVis-Suite is locally installed and allows a much more interactive exploratory data analysis (see chapter 6 and MarVis-Suite handbook 11.1). Furthermore, the general methods for ranking, filtering, clustering, and enrichment analysis facilitate the analysis and integration of data from other omics platforms besides MS-based Metabolomics.

In the context of non-targeted Metabolomics, other specialized workflows based on the combination of R-packages, such as XCMS or CAMERA, were introduced [168, 169]. The corresponding packages allow the implementation of a programmable workflow, e.g. in the form of an R-script for each individual data analysis, but are not directly applicable for scientists without appropriate programming skills and do not allow interactive data analysis. In future studies, the user-driven MarVis-Suite workflow for fast hypothesis generation should be combined with an adapted R-based workflow, featuring more sophisticated tools for metabolite identification, e.g. based on the annotation of in-source fragmentations [168]. Such a parallel workflow would also facilitate the integration of powerful R-packages specialized on the enrichment analysis of microarray data sets, such as the GSA package [147].

In order to cope with the challenge of metabolite identification, the development of the MarVis-Suite was focused on the integration of other omics platforms, e.g. Transcriptomics data from DNA microarray analysis (see chapter 5 and 6). In this context, the MarVis-Suite was also applied to the evaluation of RNA-seq Transcriptomics and MS-based Proteomics data [112] in the BMBF BioFung project (“The plant-pathogenic fungus *Verticillium longisporum* and the interaction with its host *Brassica napus*”). In future projects, also MS/MS and GC/MS data in combination with more specialized databases [58, 59, 60, 61] should be integrated into the MarVis-Suite or a parallel workflow.

References

- [1] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18:1157–1161, 2000.
- [2] W. Weckwerth. Metabolomics in systems biology. *Annual Review of Plant Biology*, 54(1):669–689, 2003.
- [3] R. J. Bino, R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. J. Nikolau, P. Mendes, U. Roessner-Tunali, M. H. Beale, R. N. Trethewey, B. M. Lange, E. S. Wurtele, and L. W. Sumner. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, 9(9):418–425, 2004.
- [4] O. Fiehn. Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, 2002.
- [5] W. Lu, B. D. Bennett, and J. D. Rabinowitz. Analytical strategies for LC–MS-based targeted metabolomics. *Journal of Chromatography B*, 871(2):236–242, 2008.
- [6] W. B. Dunn, A. Erban, R. J. Weber, D. J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, and M. R. Viant. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9(1):44–66, 2013.
- [7] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007.
- [8] C. A. Sellick, D. Knight, A. S. Croxford, A. R. Maqsood, G. M. Stephens, R. Goodacre, and A. J. Dickson. Evaluation of extraction processes for intracellular metabolite profiling of mammalian cells: matching extraction approaches to cell type and metabolite targets. *Metabolomics*, 6(3):427–438, 2010.
- [9] V. Matyash, G. Liebisch, T. V. Kurzchalia, A. Shevchenko, and D. Schwudke. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of Lipid Research*, 49(5):1137–1146, 2008.

- [10] M. Possienke. *Signals and metabolic consequences during the interaction of Brassicaceae and Verticillium longisporum*. PhD thesis, Georg-August-University Göttingen, 2011.
- [11] C. Göbel and I. Feussner. Methods for the analysis of oxylipins in plants. *Phytochemistry*, 70(13-14):1485–1503, 2009.
- [12] P. Meinicke, T. Lingner, A. Kaever, K. Feussner, C. Göbel, I. Feussner, P. Karlovsky, and B. Morgenstern. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology*, 3: 9, 2008.
- [13] F. Kaplan, J. Kopka, D. W. Haskell, W. Zhao, K. C. Schiller, N. Gatzke, D. Y. Sung, and C. L. Guy. Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiology*, 136(4):4159–4168, 2004.
- [14] W. B. Dunn, N. J. Bailey, and H. E. Johnson. Measuring the metabolome: current analytical technologies. *Analyst*, 130(5):606–625, 2005.
- [15] W. B. Dunn, D. I. Broadhurst, H. J. Atherton, R. Goodacre, and J. L. Griffin. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, 40(1):387–426, 2011.
- [16] J. Schripsema. Application of NMR in plant metabolomics: techniques, problems and prospects. *Phytochemical Analysis*, 21(1):14–21, 2010.
- [17] T. M. Annesley. Ion suppression in mass spectrometry. *Clinical Chemistry*, 49(7):1041–1044, 2003.
- [18] J. Lisec, N. Schauer, J. Kopka, L. Willmitzer, and A. R. Fernie. Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature Protocols*, 1(1):387–396, 2006.
- [19] W. Niessen. State-of-the-art in liquid chromatography–mass spectrometry. *Journal of Chromatography A*, 856(1):179–197, 1999.
- [20] J. W. Allwood and R. Goodacre. An introduction to liquid chromatography–mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis*, 21(1):33–47, 2010.

- [21] M. Brown, W. B. Dunn, P. Dobson, Y. Patel, C. Winder, S. Francis-McIntyre, P. Beggley, K. Carroll, D. Broadhurst, A. Tseng, N. Swainston, I. Spasic, R. Goodacre, and D. B. Kell. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134(7):1322–1332, 2009.
- [22] E. Lange, R. Tautenhahn, S. Neumann, and C. Gröpl. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, 9(1):375, 2008.
- [23] C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.
- [24] M. Brown, D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, L. C. Kenny, M. A. Mamas, L. Neyses, and W. B. Dunn. Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics*, 27(8):1108–1112, 2011.
- [25] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, 2(4):775–795, 2012.
- [26] M. Sugimoto, M. Kawakami, M. Robert, T. Soga, and M. Tomita. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Current Bioinformatics*, 7(1):96–108, 2012.
- [27] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121, 2001.
- [28] S. Wright. Adjusted p-values for simultaneous inference. *Biometrics*, 48(4):1005–1013, 1992.
- [29] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [30] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- [31] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

- [32] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
- [33] L. Tarpley, A. Duran, T. Kebrom, and L. Sumner. Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period. *BMC Plant Biology*, 5:8, 2005.
- [34] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
- [35] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS letters*, 451(2):142–146, 1999.
- [36] P. Jonsson, J. Gullberg, A. Nordström, M. Kusano, M. Kowalczyk, M. Sjöström, and T. Moritz. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry*, 76(6):1738–1745, 2004.
- [37] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky. Analysis of metabolomic data using support vector machines. *Analytical Chemistry*, 80(19):7562–7570, 2008.
- [38] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [39] Y. Truong, X. Lin, and C. Beecher. Learning a complex metabolomic dataset using random forests and support vector machines. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 835–840. ACM, 2004.
- [40] X. Lin, Q. Wang, P. Yin, L. Tang, Y. Tan, H. Li, K. Yan, and G. Xu. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, 7(4):549–558, 2011.
- [41] A. Kaefer, T. Lingner, K. Feussner, C. Göbel, I. Feussner, and P. Meinicke. MarVis: a Tool for Clustering and Visualization of Metabolic Biomarkers. *BMC Bioinformatics*, 10:92, 2009.
- [42] J. Draper, D. Enot, D. Parker, M. Beckmann, S. Snowdon, W. Lin, and H. Zubair. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics*, 10:227, 2009.

- [43] R. Tautenhahn, C. Böttcher, and S. Neumann. Annotation of LC/ESI-MS mass signals. *Bioinformatics Research and Development*, 4414:371–380, 2007.
- [44] A. Kaefer and M. Landesfeind. Identifikation und Analyse von relevanten Metabolit-Gruppen in Massenspektrometrie-Daten. Project thesis, Georg-August-University Göttingen, 2009.
- [45] A. Kaefer. Entwicklung und Evaluation von Methoden zur Erkennung von Metabolit-Markern in Massenspektrometrie-Daten. Master’s thesis, Georg-August-University Göttingen, 2010.
- [46] A. Alonso, A. Julià, A. Beltran, M. Vinaixa, M. Díaz, L. Ibañez, X. Correig, and S. Marsal. AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics*, 27(9):1339–1340, 2011.
- [47] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, and S. Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2012.
- [48] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8(1):105, 2007.
- [49] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.
- [50] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.
- [51] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1):D742–D753, 2012.
- [52] K. Suhre and P. Schmitt-Kopplin. MassTRIX: mass translator into pathways. *Nucleic Acids Research*, 36(suppl 2):W481–W484, 2008.
- [53] L. A. Mueller, P. Zhang, and S. Y. Rhee. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology*, 132(2):453–460, 2003.

- [54] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhtudinov, L. Li, H. J. Vogel, and I. Forsythe. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(suppl 1): D603–D610, 2009.
- [55] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. Raetz, D. W. Russell, and S. Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic Acids Research*, 35(suppl 1):D527–D532, 2007.
- [56] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya. KNApSACk: a comprehensive species-metabolite relationship database. In *Plant Metabolomics*, volume 57 of *Biotechnology in Agriculture and Forestry*, pages 165–181. Springer, 2006.
- [57] Y. Wang, J. Xiao, T. Suzek, J. Zhang, J. Wang, and S. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(suppl 2):W623–W633, 2009.
- [58] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148, 2010.
- [59] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010.
- [60] C. A. Smith, G. O’Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, 27(6):747–751, 2005.
- [61] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser.

- GMD@CSB.DB: the Golm metabolome database. *Bioinformatics*, 21(8):1635–1638, 2005.
- [62] B. Wägele, M. Witting, P. Schmitt-Kopplin, and K. Suhre. MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. *PLoS ONE*, 7(7):e39860, 2012.
- [63] N. Kessler, H. Neuweger, A. Bonte, G. Langenkämper, K. Niehaus, T. W. Nattkemper, and A. Goesmann. MeltDB 2.0—advances of the metabolomics software system. *Bioinformatics*, 29(19):2452–2459, 2013.
- [64] J. Xia, R. Mandal, I. V. Sinelnikov, D. Broadhurst, and D. S. Wishart. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Research*, 40(W1):W127–W133, 2012.
- [65] G. Kastenmüller, W. Römisch-Margl, B. Wägele, E. Altmaier, and K. Suhre. metaP-Server: A Web-Based Metabolomics Data Analysis Tool. *Journal of Biomedicine and Biotechnology*, 2011, 2011. doi: 10.1155/2011/839862.
- [66] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.
- [67] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.
- [68] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.
- [69] A. R. Joyce and B. Ø. Palsson. The model organism as a system: integrating ‘omics’ data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006.
- [70] N. Gehlenborg, S. I. O’Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A. C. Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7:S56–S68, 2010.
- [71] W. Weckwerth, K. Wenzel, and O. Fiehn. Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, 4(1):78–83, 2004.
- [72] V. Shulaev, D. Cortes, G. Miller, and R. Mittler. Metabolomics for plant stress response. *Physiologia Plantarum*, 132(2):199–208, 2008.

- [73] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–140, 2002.
- [74] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):3, 2004.
- [75] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210, 2003.
- [76] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.
- [77] A. J. Saldanha. Java Treeview-extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–3248, 2004.
- [78] M. J. de Hoon, S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.
- [79] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.
- [80] D. A. Hosack, G. Dennis Jr, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):R70, 2003.
- [81] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.
- [82] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.
- [83] M. Ackermann and K. Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47, 2009.
- [84] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and

- G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
- [85] J. Xia and D. S. Wishart. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(suppl 2):W71–W77, 2010.
- [86] M. Persicke, C. Rückert, J. Plassmeier, L. J. Stutz, N. Kessler, J. Kalinowski, A. Goemann, and H. Neuweger. MSEA: metabolite set enrichment analysis in the MeltDB metabolomics software platform: metabolic profiling of *Corynebacterium glutamicum* as an example. *Metabolomics*, 8(2):310–322, 2012.
- [87] K. Arakawa, N. Kono, Y. Yamada, H. Mori, and M. Tomita. KEGG-based pathway visualization tool for complex omics data. *In Silico Biology*, 5(4):419–423, 2005.
- [88] S.-L. T. Normand. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359, 1999.
- [89] L. V. Hedges. *Statistical Methodology in Meta-Analysis*. ERIC, Princeton, 1982.
- [90] Y. Moreau, S. Aerts, B. D. Moor, B. D. Strooper, and M. Dabrowski. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, 19(10):570–577, 2003.
- [91] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine*, 5(9):e184, 2008.
- [92] F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, 2008.
- [93] G. C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, 40(9):3785–3799, 2012.
- [94] K. Shen and G. C. Tseng. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323, 2010.
- [95] A. Stintzi, H. Weber, P. Reymond, and E. E. Farmer. Plant defense in the absence of jasmonic acid: the role of cyclopentenones. *PNAS*, 98(22):12837–12842, 2001.

- [96] P. Reymond, N. Bodenhausen, R. M. Van Poecke, V. Krishnamurthy, M. Dicke, and E. E. Farmer. A conserved transcript pattern in response to a specialist and a generalist herbivore. *The Plant Cell*, 16(11):3132–3147, 2004.
- [97] A. Gfeller, K. Baerenfaller, J. Loscos, A. Chételat, S. Baginsky, and E. E. Farmer. Jasmonate controls polypeptide patterning in undamaged tissue in wounded arabidopsis leaves. *Plant Physiology*, 156(4):1797–1807, 2011.
- [98] G. Howe and G. Jander. Plant immunity to insect herbivores. *Annual Review of Plant Biology*, 59:41–66, 2008.
- [99] A. Mosblech, I. Feussner, and I. Heilmann. Oxylipins: structurally diverse metabolites from fatty acid oxidation. *Plant Physiology and Biochemistry*, 47(6):511–517, 2009.
- [100] C. Wasternack and B. Hause. Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in *Annals of Botany*. *Annals of Botany*, 111(6):1021–1058, 2013.
- [101] B. von Malek, E. van der Graaff, K. Schneitz, and B. Keller. The Arabidopsis malesterile mutant *dde2-2* is defective in the ALLENE OXIDE SYNTHASE gene encoding one of the key enzymes of the jasmonic acid biosynthesis pathway. *Planta*, 216(1):187–192, 2002.
- [102] Y. Yan, S. Stolz, A. Chételat, P. Reymond, M. Pagni, L. Dubugnon, and E. E. Farmer. A downstream mediator in the growth repression limb of the jasmonate pathway. *The Plant Cell*, 19(8):2470–2483, 2007.
- [103] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, and S. A. Sansone. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, 2003.
- [104] A. Kaefer, M. Landesfeind, M. Possienke, K. Feussner, I. Feussner, and P. Meinicke. MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data. *Journal of Biomedicine and Biotechnology*, 2012, 2012. doi: 10.1155/2012/263910.
- [105] A. Kaefer, M. Landesfeind, K. Feussner, B. Morgenstern, I. Feussner, and P. Meinicke. Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets. *PLoS ONE*, 9(2):e89297, 2014.

- [106] A. Kaefer, M. Landesfeind, K. Feussner, A. Mosblech, I. Heilmann, B. Morgenstern, I. Feussner, and P. Meinicke. MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics*, 2014. doi: 10.1007/s11306-014-0734-y.
- [107] A. Kaefer, M. Landesfeind, K. Feussner, I. Feussner, and P. Meinicke. Metabolite clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. In *The Handbook of Plant Metabolomics*, pages 273–287. Wiley-Blackwell, 2013.
- [108] K. Nahlik, M. Dumkow, Ö. Bayram, K. Helmstaedt, S. Busch, O. Valerius, J. Gerke, M. Hoppert, E. Schwier, L. Opitz, M. Westermann, S. Grond, K. Feussner, C. Göbel, A. Kaefer, P. Meinicke, I. Feussner, and G. H. Braus. The COP9 signalosome mediates transcriptional and metabolic response for hormones, oxidative stress protection and cell wall rearrangement during fungal development. *Molecular Microbiology*, 78:964–979, 2010.
- [109] S. König, K. Feussner, M. Schwarz, A. Kaefer, T. Iven, M. Landesfeind, P. Ternes, P. Karlovsky, V. Lipka, and I. Feussner. Arabidopsis mutants of sphingolipid fatty acid α -hydroxylases accumulate ceramides and salicylates. *New Phytologist*, 196(4):1086–1097, 2012.
- [110] J. Gamir, V. Pastor, A. Kaefer, M. Cerezo, and V. Flors. Targeting novel chemical and constitutive primed metabolites against *Plectosphaerella cucumerina*. *The Plant Journal*, 78(2):227–240, 2014.
- [111] S. König, K. Feussner, A. Kaefer, M. Landesfeind, C. Thurow, P. Karlovsky, C. Gatz, A. Polle, and I. Feussner. Soluble phenylpropanoids are involved in the defense response of Arabidopsis against *Verticillium longisporum*. *New Phytologist*, 202(3):823–837, 2014.
- [112] V.-T. Tran, S. A. Braus-Stromeyer, H. Kusch, M. Reusche, A. Kaefer, A. Kühn, O. Valerius, M. Landesfeind, K. Aßhauer, M. Tech, K. Hoff, T. Pena-Centeno, M. Stanke, V. Lipka, and G. H. Braus. *Verticillium* transcription activator of adhesion *vta2* suppresses microsclerotia formation and is required for systemic infection of plant roots. *New Phytologist*, 202(2):565–581, 2014.
- [113] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.

- [114] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.
- [115] M. Landesfeind, A. Kaefer, K. Feussner, C. Thurow, C. Gatz, I. Feussner, and P. Meinicke. Integrative study of *Arabidopsis thaliana* metabolomic and transcriptomic data with the interactive MarVis-Graph software. *PeerJ*, 2(e239), 2014.
- [116] T. S. Lee, Y. S. Ho, H. C. Yeo, J. P. Y. Lin, and D.-Y. Lee. Precursor mass prediction by clustering ionization products in LC-MS-based metabolomics. *Metabolomics*, 9(6):1301–1310, 2013.
- [117] G. J. Patti, R. Tautenhahn, and G. Siuzdak. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nature Protocols*, 7(3):508–516, 2012.
- [118] T. Kind, V. Tolstikov, O. Fiehn, and R. H. Weiss. A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical Biochemistry*, 363(2):185–195, 2007.
- [119] Z. He and J. Zhou. Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. *Applied and Environmental Microbiology*, 74(10):2957–2966, 2008.
- [120] L. V. Hedges. Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2):107–128, 1981.
- [121] G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32:490–495, 2002.
- [122] Y. H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 3(8):579–588, 2002.
- [123] B. Efron, R. Tibshirani, V. Goss, and G. Chu. Microarrays and their use in a comparative experiment. *Technical Report, Stanford University*, 2000.
- [124] A. Oshlack and M. J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4:14, 2009.
- [125] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):R14, 2010.
- [126] I. Lönnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 12(1):31–46, 2002.

- [127] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65, 2006.
- [128] W. Zhang, F. Li, and L. Nie. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156(2):287–301, 2010.
- [129] F. Kaplan, J. Kopka, D. Y. Sung, W. Zhao, M. Popp, R. Porat, and C. L. Guy. Transcript and metabolite profiling during cold acclimation of *Arabidopsis* reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *The Plant Journal*, 50(6):967–981, 2007.
- [130] F. Carrari, C. Baxter, B. Usadel, E. Urbanczyk-Wochniak, M. I. Zanor, A. Nunes-Nesi, V. Nikiforova, D. Centero, A. Ratzka, M. Pauly, L. J. Sweetlove, and A. R. Fernie. Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiology*, 142(4):1380–1396, 2006.
- [131] S. Osorio, R. Alba, C. M. Damasceno, G. Lopez-Casado, M. Lohse, M. I. Zanor, T. Tohge, B. Usadel, J. K. Rose, Z. Fei, J. J. Giovannoni, and A. R. Fernie. Systems biology of tomato fruit development: combined transcript, protein, and metabolite analysis of tomato transcription factor (*nor*, *rin*) and ethylene receptor (*nr*) mutants reveals novel regulatory interactions. *Plant Physiology*, 157(1):405–425, 2011.
- [132] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, and J. Trygg. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52(6):1181–1191, 2007.
- [133] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.
- [134] I. González, K.-A. Lê Cao, M. J. Davis, and S. Déjean. Visualising associations between paired 'omics' data sets. *BioData Mining*, 5(1):1–23, 2012.
- [135] M. Y. Hirai, M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *PNAS*, 101(27):10205–10210, 2004.
- [136] D. H. Milone, G. S. Stegmayer, L. Kamenetzky, M. López, J. M. Lee, J. J. Giovannoni, and F. Carrari. *omeSOM: a software for clustering and visualization of transcriptional

- and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinformatics*, 11(1):438, 2010.
- [137] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, 2(10):2366–2382, 2007.
- [138] B. H. Junker, C. Klukas, and F. Schreiber. VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1):109, 2006.
- [139] Z. Hu, J. Mellor, J. Wu, T. Yamada, D. Holloway, and C. DeLisi. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Research*, 33(suppl 2):W352–W357, 2005.
- [140] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, 31(1):19–20, 2002.
- [141] O. Thimm, O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, P. Krüger, J. Selbig, L. A. Müller, S. Y. Rhee, and M. Stitt. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6):914–939, 2004.
- [142] T. Tokimatsu, N. Sakurai, H. Suzuki, H. Ohta, K. Nishitani, T. Koyama, T. Umezawa, N. Misawa, K. Saito, and D. Shibata. KaPPA-View. A web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiology*, 138(3):1289–1300, 2005.
- [143] J. Förster, I. Famili, P. Fu, B. Ø. Palsson, and J. Nielsen. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13(2):244–253, 2003.
- [144] A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, 14(2):301–312, 2004.

- [145] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000.
- [146] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.
- [147] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.
- [148] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, and A. Valencia. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457, 2012.
- [149] C. Von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261, 2003.
- [150] R. Tautenhahn, G. J. Patti, E. Kalisiak, T. Miyamoto, M. Schmidt, F. Y. Lo, J. McBee, N. S. Baliga, and G. Siuzdak. metaXCMS: second-order analysis of untargeted metabolomics data. *Analytical Chemistry*, 83(3):696–700, 2010.
- [151] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, San Diego, 1985.
- [152] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 1925.
- [153] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62(15):4427–4433, 2002.
- [154] M. C. Alles, M. Gardiner-Garden, D. J. Nott, Y. Wang, J. A. Foekens, R. L. Sutherland, E. A. Musgrove, and C. J. Ormandy. Meta-analysis and gene set enrichment relative to ER status reveal elevated activity of MYC and E2F in the "basal" breast cancer subgroup. *PLoS ONE*, 4(3):e4710, 2009.
- [155] S. R. Setlur, T. E. Royce, A. Sboner, J.-M. Mosquera, F. Demichelis, M. D. Hofer, K. D. Mertz, M. Gerstein, and M. A. Rubin. Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer. *Cancer Research*, 67(21):10296–10303, 2007.

- [156] P. Martini, D. Risso, G. Sales, C. Romualdi, G. Lanfranchi, and S. Cagnin. Statistical Test of Expression Pattern (STEPath): a new strategy to integrate gene expression data with genomic information in individual and meta-analysis studies. *BMC Bioinformatics*, 12(1):92, 2011.
- [157] T. Manoli, N. Gretz, H.-J. Gröne, M. Kenzelmann, R. Eils, and B. Brors. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22(20):2500–2506, 2006.
- [158] X. Wang, D. D. Kang, K. Shen, C. Song, S. Lu, L.-C. Chang, S. G. Liao, Z. Huo, S. Tang, Y. Ding, N. Kaminski, S. Etienne, Y. Lin, J. Li, and G. C. Tseng. An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, 28(19):2534–2536, 2012.
- [159] M. B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975.
- [160] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. *The American soldier: adjustment during army life*. Princeton University Press, Princeton, 1949.
- [161] M. Whitlock. Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *Journal of Evolutionary Biology*, 18(5):1368–1373, 2005.
- [162] T. M. Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47(3):467–485, 2004.
- [163] L. H. C. Tippett. *The Methods of Statistics*. Williams & Norgate Ltd., 1931.
- [164] B. Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156, 1951.
- [165] W. R. Rice. A consensus combined p-value test and the family-wide significance of component tests. *Biometrics*, 46(2):303–308, 1990.
- [166] H. Sun, H. Wang, R. Zhu, K. Tang, Q. Gong, J. Cui, Z. Cao, and Q. Liu. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics*, 30(5):737–739, 2014.
- [167] The R Project for Statistical Computing, 2014. URL <http://www.r-project.org>.

- [168] E. Gaquerel, C. Kuhl, and S. Neumann. Computational annotation of plant metabolomics profiles via a novel network-assisted approach. *Metabolomics*, 9(4):1–15, 2013.
- [169] F. Fernández-Albert, R. Llorach, C. Andrés-Lacueva, and A. Perera. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics*, 2014. doi: 10.1093/bioinformatics/btu136.

Acknowledgments

I am deeply grateful to Dr. Kirstin Feussner and Prof. Dr. Ivo Feussner for conceptual input to the MarVis-Suite workflow and interfaces, providing data sets, support in interpretation of results, and careful revision of manuscripts. Furthermore, I would like to thank Prof. Dr. Burkhard Morgenstern and Dr. Peter Meinicke for supervision and Volkhard Kaefer, Manuel Landesfeind, Kathrin Aßhauer, and Prof. Dr. Helmut Grubmüller for support and fruitful discussions. The people from the GGNB office did a great job providing fast replies to all organizational questions. And, of course, thanks to all fellows of the Biomolecules program for having a good time.

This work was funded by the German Federal Ministry of Education and Research (BMBF BioFung project 0315595A).

Supplementary Material

11.1 MarVis-Suite handbook

This handbook describes all methods and user interfaces implemented in the MarVis-Suite version 2.0. Previous versions of the handbook/documentation for the original MarVis tool [41] and the MarVis-Suite 1.0 [104] can be found on the project homepage <http://marvis.gobics.de>.

The MarVis-Suite workflow and interfaces were conceptually designed in collaboration with Manuel Landesfeind and Dr. Peter Meinicke (Department of Bioinformatics) and Dr. Kirstin Feussner and Prof. Dr. Ivo Feussner (Department of Plant Biochemistry). The MarVis-Cluster software is based on the original MarVis tool [41]. The algorithm for training of one-dimensional self-organizing maps [12] and the function for principal component analysis were implemented by Dr. Peter Meinicke. The interface for setting of figure properties was developed by Lars Söder under supervision of Alexander Kaefer and Dr. Peter Meinicke. The MarVis-Suite handbook includes the documentation of the original MarVis tool, which was revised by Dr. Peter Meinicke and Dr. Thomas Lingner (Department of Bioinformatics).

Handbook

**The MarVis-Suite:
Marker Visualization, Filtering,
Clustering, and Functional Annotation**

marvis@gobics.de

Version 2.0

Contents

1	Introduction	1
2	Installation and Requirements	2
3	MarVis-Filter	3
3.1	File import and export	3
3.2	Data transformation and normalization	7
3.3	Ranking	8
3.3.1	ANOVA/t-test and Kruskal-Wallis/ranksum test	8
3.3.2	Signal-to-noise/level ranking	8
3.3.3	Fold-change ranking	13
3.3.4	Intensity level-based ranking	14
3.3.5	Additional ranking methods	14
3.4	Visualization, filtering, and data analysis	17
3.4.1	Main window	17
3.4.2	Selection and re-filtering	19
3.4.3	Sample-based analysis and visualization	20
3.5	Adduct and isotope correction	21
3.6	Data exchange with MarVis-Cluster and MarVis-Pathway	24
4	MarVis-Cluster	26
4.1	File import	26
4.2	File export	27
4.3	Clustering	28
4.4	Visualization and data analysis	29
4.4.1	Main window	31
4.4.2	Selection of marker candidates	36
4.4.3	General visualization properties	39
4.5	Data exchange with MarVis-Pathway	40

4.6	Example data sets	40
4.6.1	The wound response of <i>Arabidopsis thaliana</i>	40
4.6.2	The yeast cell cycle	43
5	MarVis-Pathway	45
5.1	Data import	45
5.2	Database selection and query	45
5.2.1	Database selection	46
5.2.2	Entry mapping	47
5.2.3	Scoring of pathways	48
5.3	Visualization and data analysis	51
5.3.1	Main window	51
5.3.2	Pathway and entry search	53
5.4	Set Enrichment Analysis	54
5.4.1	Entry-based enrichment analysis	54
5.4.2	Marker/feature-based enrichment analysis	56
5.4.3	Sample-based enrichment analysis	57
5.4.4	Meta-analysis of multiple data sets	58
5.5	Data export	59
5.6	Data exchange with MarVis-Cluster	59
6	General functions in the MarVis-Suite	60
6.1	Toolbar	60
6.2	Save and load projects	61
6.3	Search for marker candidates	61
6.4	General visualization functions	63
6.4.1	Standard deviation, box, and scatter plots	63
6.4.2	Export of graphics	63
6.5	MarVis-Suite log function	64
6.6	Molecular formula calculation	64

Introduction

The MarVis-Suite is a toolbox for interactive ranking, filtering, combination, clustering, visualization, and functional annotation of data sets containing intensity-based profile vectors (data set features, marker candidates, or markers) as obtained e.g. from mass spectrometry (MS), microarray, or RNA-seq experiments. The clustering algorithm is based on a realization of one-dimensional self-organizing maps (1D-SOMs) [1]. Additionally, the MarVis-Suite includes specialized functions for analysis of MS data in the context of untargeted Metabolomics studies, such as adduct and isotope correction and molecular formula calculation.

This documentation covers all features implemented in the MarVis-Suite version 2.0. The MarVis-Filter interface (see section 3) provides functions for import, preprocessing, filtering, and combination of raw data files, while the MarVis-Cluster interface (see section 4) was designed for high-level visualization and cluster analysis. The MarVis-Pathway interface (see section 5) is used for functional annotation of filtered/combined data sets or selected clusters in the context of reference or organism-specific pathway maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and BioCyc databases [2, 3, 4]. For statistical analysis of combined data sets from different omics platforms, MarVis-Pathway provides an extensive framework for (Gene/Metabolite) Set Enrichment Analysis [5, 6] and meta-analysis [7]. Within the MarVis-Suite, selected data can be easily exchanged between the different interfaces. Nonetheless, the interfaces can also be utilized as independent tool as described in [8, 9].

The MarVis-Suite was developed at the Department of Bioinformatics in collaboration with the Department of Plant Biochemistry at the Georg-August-University Göttingen. Installation packages for Microsoft Windows XP/Vista/7/8 and Linux x86 (only MarVis-Cluster) can be obtained from <http://marvis.gobics.de>. Please send questions, bug reports, and feature requests to marvis@gobics.de.

Installation and Requirements

The MarVis-Suite is implemented in the MATLAB[®] (MathWorks Corporation) programming language and compiled for Microsoft Windows XP/Vista/7/8 and Linux x86 (only MarVis-Cluster). The install packages can be downloaded from <http://marvis.gobics.de>. For installation, follow the steps in the README-file in the package. MarVis was originally developed for Windows and then ported to Linux. Therefore we recommend the Windows version. When installing and testing MarVis-Cluster on Linux, a number of additional packages had to be installed and minor bugs regarding the GUI-elements occurred. For experienced users we have added the corresponding references to the README-file.

The MarVis-Suite software is free for academic use. It comes with no guarantee or warranty at all. Use it at your own risk.

MarVis-Filter

3.1 File import and export

For data import and export, MarVis uses the CSV (Comma Separated Values) and Microsoft Excel file format, which can easily be processed by statistical analysis software and spreadsheet applications. Note that on systems without Microsoft Excel installed, the Excel import function is error-prone. In case an error occurs, the conversion and import in CSV format is recommended.

MarVis-Filter supports the import of files in two different ways: Via the entry `Import MarVis data` in the `File` menu, data in a MarVis-specific spreadsheet format can be imported. In case of a CSV file, each line of the file corresponds to a row of data fields separated by a delimiter character, which can freely be chosen (e.g. a comma). The file `dataset1.csv` (in the `examples` directory¹) contains an exemplary data set of the metabolomic case study from [1]. The first rows and columns can be used for comments. In this example, the first four rows and two columns are used for this purpose. The comment rows and columns are followed by the regular data starting with a header row (line 5 and column 3 in the example). The header contains customizable column labels, which are displayed in MarVis. Each of the succeeding rows represents a feature/marker candidate. The first regular column must contain identifiers for all candidates. They are interpreted and displayed as text. The second and third regular column are reserved for x and y numerical values, which are displayed by MarVis-Cluster as two-dimensional scatter plot (retention time vs. mass-to-charge-ratio in the example). The first three regular columns are followed by the numerical intensity values. They must be ordered according to replicate measurements and experimental conditions. The example data set contains intensities for eight different conditions, which are represented by nine replicate measurements, respectively. The intensity columns can be followed by additional user-specific data

¹The examples directory is located within the MarVis-Suite program directory

columns. Values in these columns are displayed by MarVis as marker-specific text, which can be helpful for further interpretation. After selecting the import file, the delimiter character (a comma in this example), the start row and column of the header (5 and 3), the number of conditions (8) and the number of replicate measurements/samples per condition (9) can be specified in the `Import dialog` (see figure 3.1). In case of an Excel import file, the delimiter character is ignored. If the number of replicate samples per condition varies, the different numbers have to be given in order of the conditions separated by space characters (e.g. `6 6 6 6 9 9 9 9`).

Data in a more general format can be imported via the entry `Import raw data` in the `File` menu. This file format conforms to the MarVis format but allows variable positions for the data columns. The files `wound_neg_raw.csv` and `wound_pos_raw.csv` (the files can be found in the `examples` directory or downloaded from the project home page) contain exemplary data sets of the metabolomic case study from [9] as generated by the MarkerLynx™ Application Manager of MassLynx™ (Waters Corporation) for the negative and positive ionization mode. After selecting the input file, the column positions (relative to the start column, e.g. 1 for the start column itself) have to be specified in the `Import dialog` (see figure 3.2). In this dialog, the position of the column containing the marker candidate IDs (`ID column`), the positions of the columns containing `x` and `y`-values (`x column` and `y column`), and the corresponding labels which should be displayed in MarVis have to be specified. Marker candidate IDs are automatically generated as ascending numbers if the checkbox `Generate IDs` is activated. If no `x` or `y`-columns are specified, MarVis inserts zeros, signal-to-level ratios (`y`) and log-levels (`x`) (see section 3.3.2), or parses the values from the marker IDs. In the latter case, the user has to specify the pattern which encodes the `x` and `y`-values within the ID strings (e.g. “`y_x`” for IDs which contain the `y` and `x`-values separated by “`_`” characters). In the field `Condition identifier`, distinct identifiers for the experimental conditions have to be specified separated by comma. MarVis searches the header row (first row of regular data) for columns which contain these identifiers as substrings of their labels and groups them as replicate samples into different conditions. The identifiers are interpreted as case-insensitive regular expressions (see MATLAB® `regexp` function for details). White spaces are ignored in the corresponding text field. In addition to the regular data columns, other columns may be imported by specifying the (relative) positions in the `Additional columns` text field separated by space characters and corresponding labels in the `Additional labels` field separated by comma. If the `Additional labels` field contains more labels (separated by comma) than positions in the `Additional columns` field (separated by

space characters), MarVis-Filter tries to fill the missing column positions by searching the column headers for the extra labels.

If the import file is a tab-separated values file (using a tabulator as delimiter), this can be specified by `\t` in the `Delimiter` textfield.

Ranked and filtered data sets can be exported in the MarVis CSV file format using the `Export MarVis data` entry in the `File` menu. After selecting an output file, MarVis opens the `Export` dialog (see figure 3.3). A delimiter character for CSV export has to be specified in the `Delimiter` textfield. If the `Filter data set` checkbox is activated, MarVis exports only the marker candidates below the selected threshold (see section 3.4). The current filter criterion (e.g. `p-value`) can be exported in addition to the regular marker data by selecting the `Export filter criterion` checkbox. The rows in the output file can also be sorted according to the filter criterion (`Sort data set by filter criterion` checkbox).

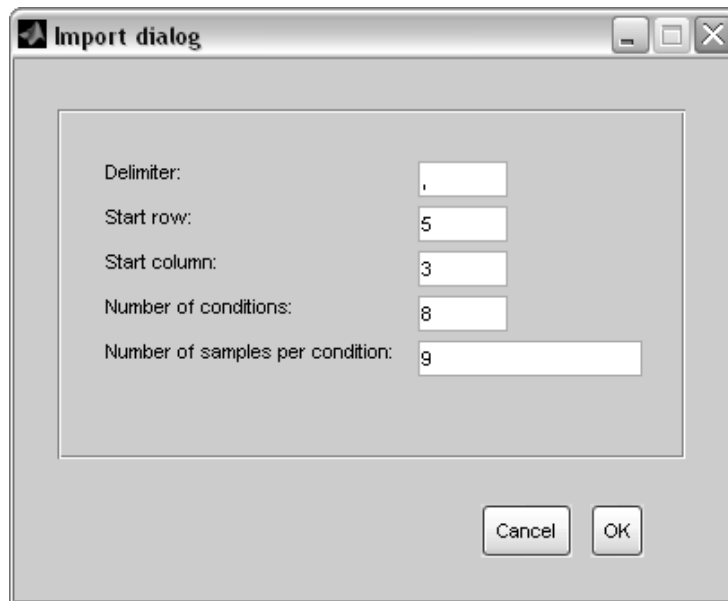


Figure 3.1: Import dialog for `dataset1.csv`

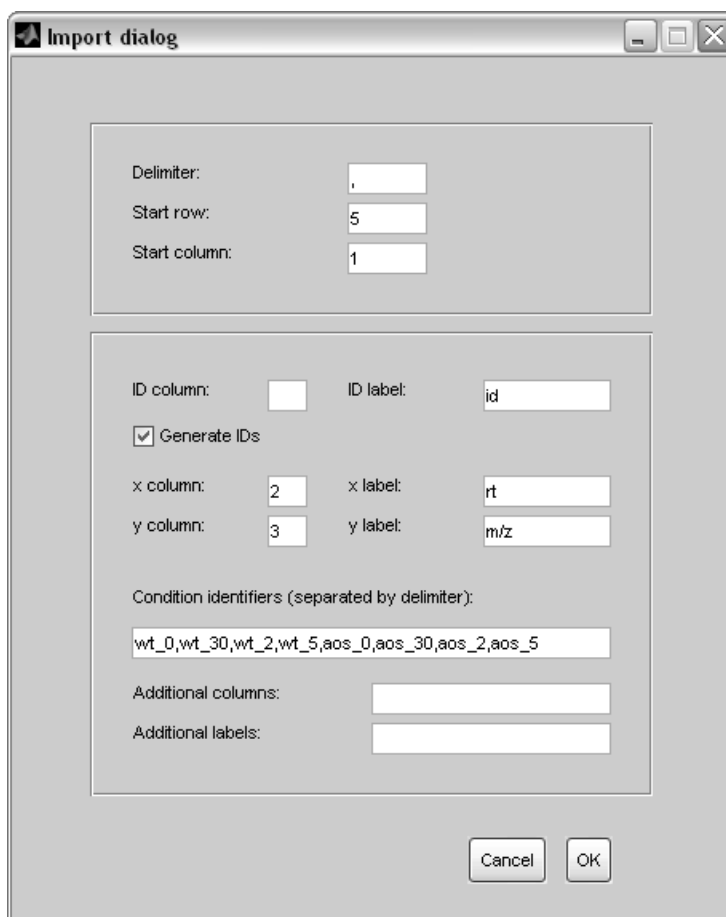


Figure 3.2: Raw import dialog for wound_neg_raw.csv

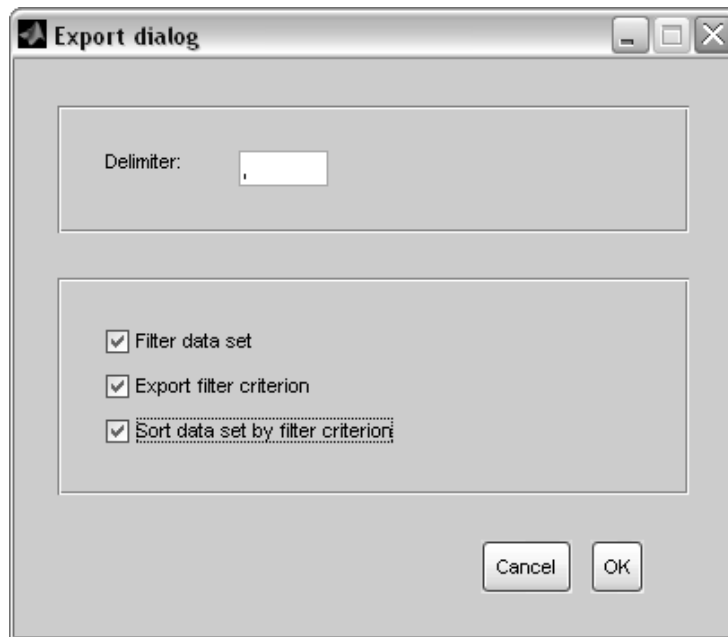


Figure 3.3: Export dialog in MarVis-Filter

3.2 Data transformation and normalization

After import and before ranking, a simple log-transformation ($\log_2(x)$) and a sample-based normalization of the intensity data can be performed.

In case the log-transformation should be performed and the data set contains negative intensities, the global minimum is subtracted from all values, previously. If the data set contains positive intensities smaller than 1, a pseudo count of 1 is added to all values before applying the log-transformation.

In case of the sample-based normalization, all intensities in a sample are scaled according to the respective sum of absolute intensities and multiplied by 1000, divided by the respective mean absolute intensity, divided by the median (considering only non-zero absolute intensities), divided by a size factor as introduced in [10] (considering only profiles with all intensities greater than zero), or normalized based on a quantile-normalization. The sample-based normalization is useful when comparing samples with different average intensity levels.

3.3 Ranking

After data import, a method for ranking has to be selected in the `Filter` dialog (see figure 3.4). Along with a title, a subset of conditions which should be used for testing/ranking can be specified in the `Use conditions` textfield. The condition indices have to be given separated by space characters or colons (1 2 3 or 1:3, for example, selects the first three conditions for testing/ranking). By default (blank field), all conditions are used. Additionally, the method for ranking has to be selected. This method is then applied to rank the marker candidates according to the relevance of their intensity profiles. If the `New window` checkbox is selected, MarVis displays the results of ranking in a new instead of the current window. This option can be useful when different methods should be compared.

3.3.1 ANOVA/t-test and Kruskal-Wallis/ranksum test

MarVis-Filter provides functions for the well-known one-way ANOVA and nonparametric Kruskal-Wallis test [11, 12], which are used to identify candidates with significant intensity differences between the experimental conditions. After calculation of p-values, MarVis opens a dialog for multiple testing adjustment/correction [13] (see figure 3.5). The p-values may be adjusted according to the Bonferroni or Holm-Bonferroni method [14], which controls the familywise error rate (FWER), or the Benjamini-Hochberg method [15], which controls the false discovery rate (FDR).

If the data set contains only two experimental conditions, a two-sample student's t-test or Wilcoxon ranksum test is performed instead of the ANOVA/Kruskal-Wallis test, respectively. The ANOVA test may be performed on log-transformed intensities. In this case and in contrast to the general transformation of intensities after data import (see section 3.2), the transformed intensities are used only for this test and not for further processing.

3.3.2 Signal-to-noise/level ranking

Similar to the Significance Analysis of Microarrays (SAM) [16], MarVis-Filter provides a signal-to-noise ratio (SNR) [17] based method for ranking and filtering. This method can be used for an undirected or directed comparison of different subsets of conditions (e.g. higher intensities in samples of condition one or two compared to samples of condition three and four). For each marker candidate intensity profile, the signal-to-noise ratio (option

Calculate signal-to-noise ratio) is calculated as

$$R_{SN} = \frac{S}{N} \quad (3.1)$$

while the signal-to-level ratio (SLR) (option Calculate signal-to-level ratio) is calculated as

$$R_{SL} = \frac{S - N}{L} . \quad (3.2)$$

The noise level N is defined as the square root of the pooled unbiased sample variance over all c conditions:

$$N = \sqrt{\frac{\sum_c (n_c - 1) s_c^2}{\sum_c (n_c - 1)}} , \quad (3.3)$$

where n_c denotes the number of samples in condition c and s_c^2 the unbiased sample variance for condition c .

The signal is calculated either as the maximum average (arithmetic mean) intensity over all conditions $S = \max_c (\bar{x}_c)$ (Signal type option Maximum condition) or as the difference of the maximum and minimum average intensities over all conditions $S = \max_c (\bar{x}_c) - \min_c (\bar{x}_c)$ (option Difference between conditions). The first option is useful when searching for candidates with a good signal-to-noise/level ratio but not necessarily a large difference between two or more conditions. The second definition is similar to Hedges's effect size estimator [18]. For the signal-to-level calculation, the level is defined as the maximum absolute average intensity over all conditions: $L = \max_c (|\bar{x}_c|)$. In contrast to the signal-to-noise ratio, the signal-to-level ranking is robust against very small noise levels and focuses on high signals relative to the level.

In case the difference is treated as signal, the user may customize the formula for the difference calculation. First, the conditions whose aggregated average intensities should be added in the signal calculation can be specified as indices (e.g. 1 2 3 or 1:3 for conditions one to three). Following this dialog, the method for aggregation of the selected average condition intensities has to be chosen (Maximum over selected conditions, Minimum over selected conditions, or Mean over selected conditions). Second, the indices of conditions whose average intensities should be subtracted in the signal calculation and a method for aggregation can

be specified. The signal is then calculated as

$$S = \max | \min | \text{mean}_d(\bar{x}_d) - \max | \min | \text{mean}_e(\bar{x}_e) \quad (3.4)$$

over all condition indices d which were selected in the first part and all conditions e which were selected in the second part. A negative signal is set to minus-infinity. If no condition indices are specified (default option) the signal is calculated as described at the start of this section.

The customized difference calculation is useful when searching for candidates which show high intensities in one or more conditions in comparison to the remaining conditions (e.g. controls).

For the final ranking, the user can choose between three different options: The first option (`Calculate ratio/score`) is to calculate and display the raw signal-to-noise/level ratios. The second option (`Calculate FWER using random permutations`) is to calculate the familywise error rates based on random permutations. The third option (`Calculate FDR using random permutations`) is the calculation of false discovery rates based on random permutations [16]. In the second and third case, the user can choose between the permutation of sample condition labels (`Permute sample labels`), if enough independent samples per condition are available, and the permutation of intensity values per sample (`Permute intensities per sample`), which is only recommended if not enough samples are available. In either case, the number of random permutations and the labels of dependent samples have to be specified. The dependency labels have to be entered as integers separated by space characters (e.g. 1 2 3 for three independent samples or 1 1 2 for three samples and sample one and two dependent). The number of dependency labels has to be in accordance with the overall number of samples (over all conditions). The order of labels corresponds to the order of samples in the current data set (see function `Show sample names` in the `Selection` menu). Samples which are assigned to different conditions cannot be defined as dependent. By default, all samples are assumed to be independent. For each random permutation, dependent samples are treated together. In case of the `Permute sample labels` option, dependent samples are always assigned the same condition label and, in case of the `Permute intensities per sample` option, intensity values of dependent samples are always permuted using the same permutation.

The concept of dependent samples is especially useful when dealing with biological

and technical replicates (different measurements of the same biological sample). The dependency labels 1 1 2 2 3 3 4 4, for example, could be used to indicate two dependent technical replicates for each of four independent biological samples. Note that these dependency labels usually differ from the condition labels (e.g. 1 1 1 1 2 2 2 2 for two conditions in the example).

The signal-to-noise/level ratios may be calculated based on log-transformed intensities. In this case and in contrast to the general transformation of intensities after data import (see section 3.2), the transformed intensities are used only for this calculation and not for further processing.

Noise moderation

In the SAM method [16], a fudge factor, which represents a small positive value that is added to the denominator of the ratio statistic, is used to stabilize the calculation for very small denominators. In a moderated t-test [19], a global variance estimate is used to stabilize the gene-wise variance estimation. This principle is also known as shrinkage and has become very popular in microarray analysis [20]. In this context, the calculation of the raw SNR (see equation 3.1) might lead to extremely high ratios for very small noise terms. Additionally, the feature-wise noise estimation may not be very reliable in case of small sample size. In order to stabilize this estimation, a moderated noise term based on shrinkage can be used:

$$N = \frac{aN_1 + bN_0}{a + b} \quad (3.5)$$

N_1 corresponds to the definition in 3.3 and represents the local marker candidate-specific noise estimate, N_0 corresponds to a global estimate based on all available profiles. Instead of using a constant N_0 value for all features in a data set, N_0 is calculated based on the relationship of the average intensity observed for the particular candidate and the average noise observed for candidates with similar average intensities. This approach takes into account that small intensities are often associated with relatively high noise levels in contrast to large intensities which often show good reproducibility. Instead of modeling the relationship between average intensity and noise based on a predefined function class, which is difficult for heterogeneous data sets and which would impose additional assumptions about the intensity distribution, the global noise estimate is obtained by binning the average intensities per condition, calculation of median bin-specific noise values, and linear interpolation between the bins. In the following, the procedure for binning is described in detail:

1. Calculate the average intensities (arithmetic mean) for all conditions and feature profiles in the current data set
2. Leave out means which equal zero
3. Calculate the corresponding coefficients of variation (sample standard deviation divided by mean)
4. Order the value pairs into K bins (by default $K = 10$)
5. For each bin, calculate the median mean intensity (bin average) and median coefficient of variation (bin coefficient)
6. Conservatively correct the bin coefficients: If a bin coefficient is smaller than the following coefficient (for a larger bin average), assign the larger value

Step 6 ensures a monotonically decreasing function. In many applications, very small intensities are associated with purely technical noise, which is considerably smaller than the variation of biological signals. Without the correction, small intensities without biological meaning might be backed up by the moderation procedure.

The following procedure describes the lookup of the global noise estimate N_0 for a particular feature profile:

1. Calculate the mean intensities for all conditions and take the maximum
2. If the maximum mean is larger than the largest bin average: Select the bin coefficient for the largest bin average
3. Else if the maximum mean is smaller than the smallest bin average: Select the bin coefficient for the smallest bin average
4. Else: Linearly interpolate the coefficient based on the neighbor bins
5. Multiply the coefficient with the maximum mean (see step 1) in order to obtain the absolute noise estimate N_0 (see equation 3.5)

The intensity profile might contain very small intensities for some of the experimental conditions, e.g. representing only technical variation without any biological signal. Therefore, the lookup procedure utilizes the maximum condition-specific mean intensity instead of the mean over all intensities of the profile. After looking up N_0 , the noise term of the ratio can be calculated according to equation 3.5. By default, the weights a and b are set to 0.5, balancing the influence of the local and global noise estimate.

A similar procedure was used for the analysis of RNA-seq data in order to estimate the standard deviation for a moderated Chi-squared test [21].

Macro definition

Instead of using the predefined signal-to-noise/level ratio definitions, the user may also utilize a customized macro for the ratio/score calculation (within the random permutation framework). In this case, the macro has to be defined in the MATLAB[®] programming language and can be loaded from a file. When selecting this option, a commented example macro for the calculation of the signal-to-noise/level ratio is shown in a new window and can be customized.

3.3.3 Fold-change ranking

When selecting the `Fold-change` option, MarVis-Filter ranks the marker candidates according to the ratio

$$R_1 = \frac{\bar{x}_{max}}{\bar{x}_{min}} \quad (3.6)$$

or (user's choice)

$$R_2 = \frac{\bar{x}_{min}}{\bar{x}_{max}} \quad (3.7)$$

with $\bar{x}_{max} = \max_c(\bar{x}_c)$ and $\bar{x}_{min} = \min_c(\bar{x}_c)$ over all average condition-specific intensities \bar{x}_c . In the first case, if \bar{x}_{min} equals zero and \bar{x}_{max} is greater than zero, the fold change is set to infinity. In the second case, if \bar{x}_{max} equals zero, the fold change is set to infinity, too.

Similar to the signal-to-noise/level calculation, the user can define subsets of conditions which are treated as candidates for the maximum and minimum average intensities. The corresponding condition indices have to be specified in the fields `Treat the following conditions as maximum` and `Treat the following conditions as minimum` separated by space characters (e.g. `1 2 3` or `1:3` for

conditions one to three). If a list of conditions is specified for the maximum, the maximum average intensity is calculated only over the selected d conditions: $\bar{x}_{max} = \max_d(\bar{x}_d)$. The corresponding minimum average intensity (see \bar{x}_{min} in equation 3.6 and 3.7) is calculated as maximum over the remaining (not specified) e conditions: $\bar{x}_{min} = \max_e(\bar{x}_e)$. If a list of conditions is specified for the minimum, the numerator and denominator are calculated as $\bar{x}_{min} = \min_d(\bar{x}_d)$ and $\bar{x}_{max} = \min_e(\bar{x}_e)$.

Both options may be combined and in this case the final ratio is calculated as the maximum/minimum (depending on the ratio definition) of both ratios. If no subsets are specified (default option), \bar{x}_{max} and \bar{x}_{min} are calculated over all available conditions, as described at the start of this section.

Note that if the data set contains many zero or missing intensity values, the fold-change ranking is not very robust and should only be performed as second step of filtering.

3.3.4 Intensity level-based ranking

In addition to the comparative ranking methods, MarVis-Filter provides a number of functions purely based on intensity levels (and not on the comparison of different conditions). The `Maximum level` function ranks the marker candidates according to the maximum average intensity over all conditions $\max_c(\bar{x}_c)$ (descending order). The `Minimum level` function uses the minimum average intensity $\min_c(\bar{x}_c)$ (ascending order). The `High intensity ratio` method ranks the candidates according to the number of intensity values greater than a specified threshold (e.g. zero) in relation to the overall number of samples or the number of samples per condition. In the second case, the ratio is calculated per condition and the maximum is taken over all conditions.

3.3.5 Additional ranking methods

Marker candidates can also be ranked and filtered according to values imported along with the regular data (see section 3.1) using the `Value ranking` option. In this case, the respective column label has to be specified. In case the ID column is selected, the marker candidates are sorted in alphabetical order of the IDs. If another label is selected, the candidates are sorted according to the numerical values of the selected column in ascending or descending order.

If the checkbox `None` is selected, MarVis sorts the imported candidates according to

the row order in the input file. In case of a combined data set (see section 3.4.1), they are also sorted according to the original data set index.

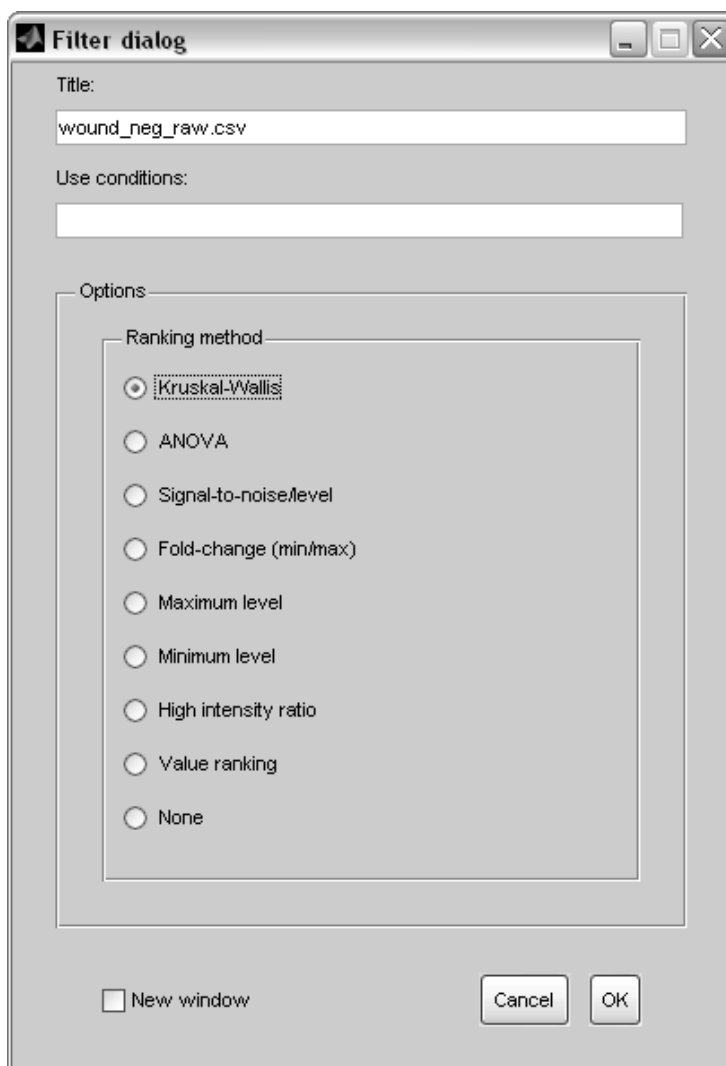


Figure 3.4: Filter dialog

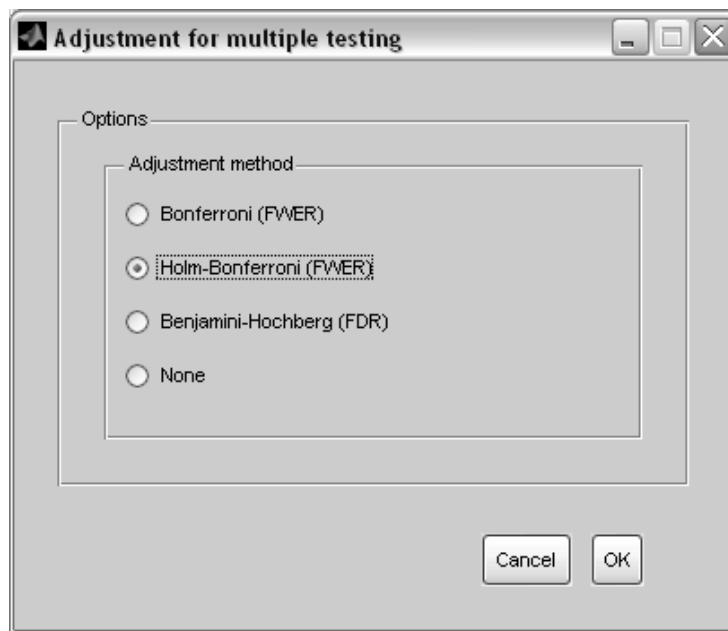


Figure 3.5: Dialog for multiple testing adjustment

3.4 Visualization, filtering, and data analysis

After import and ranking, MarVis presents the results in the MarVis-Filter main window (see figure 3.6).

3.4.1 Main window

The ranking plot

The ranking plot (area 1, see figure 3.6) displays the values used for ranking (e.g. adjusted p-values) of all marker intensity profiles in the current data set on the y-axis sorted in ascending or descending order (based on the ranking method). A marker candidate with a particular value can be selected by clicking into the plot. The red separator line indicates the currently selected value/candidate. By pressing the cursor keys, the user can slide the separator line. The data set can be interactively filtered according to a user-defined threshold by positioning the red separator line, by jumping to a predefined level (`Goto level` entry in `Selection` menu), or by going to a rank/proportion of the ranked data set (e.g. rank number 100 or best 10% using the `Goto rank` entry in `Selection` menu). The selected data (all features before and including the current candidate) can then be exported as MarVis CSV file (see section 3.1) or handed to MarVis-Cluster/MarVis-Pathway (see section 3.6). Only marker candidates associated with values on the left side of the separator line (including the current candidate) are used.

The marker profile plot

The marker profile plot (2) shows the raw intensity profile of the currently selected marker candidate as bar plot. Intensity values of replicate samples associated with the same experimental condition are marked in the same color. The intensities are sorted according to conditions and the sample order in the original file (see function `Show sample names` in the `Selection` menu for display of the sample labels/names).

The marker information box

The marker information box (3) displays information about all marker candidates in the data set ordered according to the ranking (e.g. by adjusted p-value), including ID, x-value (e.g. retention time), y-value (e.g. mass-to-charge value) and additional scores or annotations which were imported along with the data set (see section 3.1) or generated by MarVis. A candidate can be selected by clicking into the list or using the arrow keys.

The data set clipboard listbox

The data set clipboard listbox (4) shows data sets which are currently stored in the MarVis clipboard. The current data set can easily be added or removed to/from this list by clicking on the `add data set` or `remove` button. Before adding the data set to the clipboard, it is filtered according to the currently selected value. A data set in the clipboard can be selected by clicking into the listbox. Multiple data sets can be selected by holding the `Control` or `Shift` key. The data set clipboard supports an adduct and isotope correction of selected data sets in batch mode (`correct` button, see section 3.5).

Data sets in the clipboard (e.g. MS data sets corrected according to positive and negative ionization mode or from different omics platforms) may be combined into a single data set using the `combine` button. In this case, MarVis concatenates the marker candidates of selected data sets and presents the results in a new MarVis-Filter main window. In order to concatenate the intensity profiles of the different data sets (which may contain measurements for different conditions or different numbers of replicate samples), the user has to specify whether the conditions should be interlaced or stacked.

In the first case, the number of conditions in the combined data set corresponds to the maximum number of conditions within one of the selected data sets and the number of replicates for a particular condition corresponds to the maximum number of replicates for that condition and one of the data sets. The conditions of the original data sets are mapped to the combined set of conditions according to their position in the experimental setup (e.g. the first condition of an original data set is mapped to the first condition of the combined data set). If a data set does not contain the maximum number of conditions, the missing intensity values are set to NaN (“Not a Number”) as placeholder. If a single data set does not contain the maximum number of replicate samples for a particular condition, the missing intensity values are set to NaN, too.

In the second case, the conditions are not interlaced but stacked. The number of conditions in the combined data set corresponds to the sum of conditions over all selected data sets and the total number of samples to the sum of samples over all data sets. Missing intensity values (all intensities in a particular data set for conditions corresponding to another data set) are set to NaN. The stacking of conditions is useful when combining data sets with different experimental setups.

In all following steps of data analysis (e.g. calculation of average intensities), NaN values are left out or replaced by user-defined intensities.

All functions of the data set clipboard can also be accessed via the Clipboard menu.

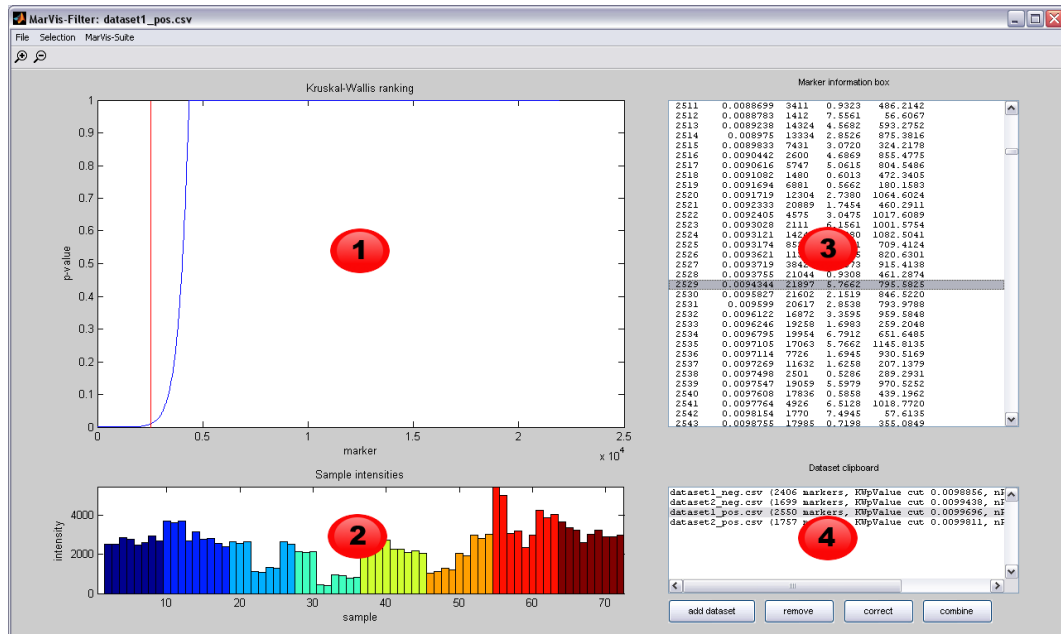


Figure 3.6: MarVis-Filter main window after import and ranking

3.4.2 Selection and re-filtering

The currently loaded data set can be ranked and filtered once again according to another criterion via the `Refilter markers ...` entry in the `Selection` menu. In this case, only the selected marker candidates (on the left side of the separator line) are used. By activating the `New window` checkbox in the `Filter` dialog the results are presented in a new MarVis-Filter main window.

Via the entries `Regroup samples` and `Remove samples`, the user can change the association of samples to experimental conditions and remove samples (e.g. outliers identified in the PCA or HCA plot, see section 3.4.3) from the data set.

In the first case, new condition labels have to be specified. The labels have to be defined as ascending integers representing the new conditions and the number of specified labels has

to correspond to the number of samples in the current data set. The labels 2 2 2 1 1 1 1 1 3 3 3 3 for example represent 12 samples with sample one to three associated with the new condition two, sample four to eight associated with condition one, and the last four samples with condition three. The current data set is then filtered (marker candidates to the left-hand side of the separator-line are retained), the samples are ordered according to the new condition labels, and the data set is ranked and displayed in a new MarVis-Filter window according to the new condition setup.

For removing of samples, the respective sample names/numbers have to be selected in a listbox dialog. Multiple samples can be selected by holding the `Control` or `Shift` key. As in the case of regrouping samples, the resulting data set is displayed in a new MarVis-Filter window.

3.4.3 Sample-based analysis and visualization

For analysis and quality control of replicate samples, MarVis-Filter provides interfaces for sample-based Principal Component Analysis (entry `Sample PCA` in the `Selection` menu) and Hierarchical Clustering Analysis (entry `Sample HCA` in the `Selection` menu).

In the first case, MarVis-Filter performs a PCA using the samples as high-dimensional intensity vectors (see [1]) and plots the eigenvalue spectrum and the scores for the first two principal components. Only the intensities of selected marker candidates are used. Before applying PCA, the selected marker intensity profiles are normalized to unit Euclidean length, respectively.

In the second case, MarVis-Filter performs a hierarchical clustering of the sample intensity vectors using different distance and linkage methods (see MATLAB[®] `linkage` and `pdist` function). In all cases, MarVis-Filter uses only the intensities of selected marker candidates. The selected marker intensity profiles may be log-scale-transformed, aggregated, or normalized before clustering. Samples corresponding to technical replicates or the same experimental condition should form distinct clusters in the PCA score or HCA dendrogram plot.

Via the entry `Show sample names` in the `Selection` menu, the names of all samples, which correspond to the respective column labels in the original data file (see section 3.1), can be shown in the current order. The samples are sorted according to experimental conditions and order in the original file. Additionally, the sample names are shown on the left-hand side of the HCA dendrogram. For the PCA score plot, the samples

names can be displayed by activating the MATLAB[®] datacursor mode (see toolbar section 6.1) and clicking on one of the scatter points.

3.5 Adduct and isotope correction

The adduct and isotope correction in MarVis-Filter takes as input values the retention times (rt, x-values), mass-to-charge ratios (m/z, y-values) and raw intensity profiles of all selected marker candidates (on the left side of the separator line) and calculates as output the putative (monoisotopic) molecular mass, ionization rule, and number of included ¹³C-isotopes for every candidate. It is based on a greedy strategy which minimizes the number of actual molecular masses and simultaneously maximizes the similarity of intensity profiles between candidates with similar retention time and actual mass. This concept follows the paradigm that in MS analysis a metabolite is usually represented by several marker candidates which show a similar retention time and intensity profile but different m/z ratios according to the various possibilities of ionization and number of included isotopes. As parameters, the function expects a list of ionization/adduct rules sorted according to relevance, the assumed maximal number of ¹³C-isotopes per marker candidate, a mass tolerance, an rt tolerance and a minimal cosine similarity. Each ionization rule is represented by a formula in the format $[xm + y]^{z[+/-]}$ [22]. x denotes the number of combined target molecules, y the mass or chemical formula of attached ions (adduct formation), and z the ionization charge (e.g. single or double charge). m (or M) is a placeholder for the target molecule. Assuming a particular formula and number of included ¹³C-isotopes, the actual (monoisotopic) mass of a marker candidate is calculated by solving the formula for m and subtracting the mass difference between ¹³C and ¹²C isotopes.

For storage of pairwise similarities between candidate profiles, the algorithm utilizes a five-dimensional matrix M . Each entry $M_{(m,a_1,i_1,a_2,i_2)}$ corresponds to the maximal cosine similarity between the intensity profile of candidate m assuming ionization rule a_1 and i_1 ¹³C-isotopes and another candidate which has a similar retention time (within tolerance) and corrected mass (within tolerance) assuming ionization rule a_2 and i_2 ¹³C-isotopes. For each candidate m , the algorithm chooses then the ionization rule and number of ¹³C-isotopes which are supported by the highest sum of cosine similarities. In the following, the algorithm is described in detail:

1. Initialize M with zeros

2. Calculate all potential masses by applying all ionization rules and number of ^{13}C -isotopes to all candidate m/z ratios
3. Consider all pairs of potential masses under the following constraints and fill M with pairwise cosine similarities of corresponding candidate profiles:
 - Consider only pairs of different marker candidates
 - Consider only pairs within mass and rt tolerance
 - Consider only pairs with at least the requested cosine similarity
 - Consider only pairs with different combinations of adduct rules and number of isotopes
 - For each entry in M , store only the maximum observed cosine similarity
4. Calculate the reduced three-dimensional matrix M^{red} with summed entries

$$M_{(m,a_1,i_1)}^{red} = \sum_{a_2,i_2} M_{(m,a_1,i_1,a_2,i_2)}$$

5. Choose for each candidate m : Adduct rule and isotope number with maximal sum of similarities $c_{max} = \max_{a_1,i_1} (M_{(m,a_1,i_1)}^{red})$. If $c_{max} = 0$, use first ionization rule and zero ^{13}C -isotopes as default
6. Calculate output masses according to chosen rules and isotope numbers

In order to avoid apparently false associations between marker candidates, negative cosine similarities are not considered. If for a given candidate different combinations of ionization rule and number of isotopes maximize the sum of cosine similarities, the ionization rule with the highest relevance and the lowest number of ^{13}C -isotopes are selected.

The adduct rules have to be specified in a text file (see files `adduct_neg.txt` and `adduct_pos.txt` in the `examples` directory). Every line in this file corresponds to an adduct/ionization rule. Each line starts with a name/description followed by a colon and the formula in the format $[xm + y]z[+/-]$. In case $x = 1$, $z = 1$, or $y = 0$, the values do not have to be specified. Every specified rule has to end with the + (positive ionization) or - (negative ionization) sign. The rules have to be sorted according to relevance (e.g. the first rule has a higher relevance than the second rule in file order). Lines starting with a % character are ignored and can be used to comment rules.

After adduct and isotope correction, the number of carbon atoms per candidate is estimated by comparing the raw intensities of marker candidates with zero predicted ^{13}C -isotopes (I_M) and the respective marker candidates including one ^{13}C -isotope (I_{M+1}) according to the formula

$$n_C = \frac{98.9 I_{M+1}}{1.1 I_M} \quad (3.8)$$

corresponding to the natural abundances of carbon isotopes. Given a pair of candidates annotated as isotopologues (M and $M + 1$) and with the same ionization rule, a robust estimation of the number of carbon atoms is obtained by calculating the median n_C over all samples which show non-zero intensities in both profiles.

The adduct and isotope correction and the estimation of the number of carbon atoms can be performed on the selected marker candidates (on the left side of the separator line) by clicking on the `Adduct and isotope correction` entry in the `Selection` menu. The user has to specify the input file for adduct/ionization rules and the additional parameters. After correction, MarVis displays the number of applied rules, the number of detected ^{13}C -isotopes, and a histogram of the maximal sum of cosine similarities per marker candidate in new windows. Note that marker candidates which could not be corrected based on supporting candidates (with similar `rt`, profile, and potential mass) are counted for the first (default) ionization rule, zero isotopes, and a maximal cosine similarity sum of zero. The marker candidates are annotated according to the results of the adduct/isotope correction and estimation of the number of carbon atoms (see marker information box 3.4.1, `ARule`: Description of the applied adduct rule, `nC13`: Number of included ^{13}C -isotopes, `CosSum`: Maximal sum of cosine similarities, `FormerY`: Original `m/z` ratio (`y` value), `nC`: Estimated number of included carbon atoms). The original `m/z` values (`y` values) are replaced by the corrected masses. A high sum of cosine similarities indicates marker candidates which were corrected based on a high number of supporting candidates and high similarities of corresponding intensity profiles.

The adduct and isotope correction should be performed on raw data sets (e.g. `wound_neg_raw.csv` in combination with the rules in `adduct_neg.txt`). Note that the filtered data set `dataset1.csv` contains only nominal instead of accurate `m/z` ratios.

The correction can be undone by clicking on the `Undo correction` entry in the

Selection menu.

3.6 Data exchange with MarVis-Cluster and MarVis-Pathway

The selected marker candidates (to the left side of the separator line) can be analyzed in MarVis-Cluster (see chapter 4) or MarVis-Pathway (5) using the `Goto MarVis-Cluster` or `Goto MarVis-Pathway` entry in the MarVis-Suite menu. A method for aggregation of replicate intensities per condition and marker candidate and for scaling of aggregated intensity profiles has to be specified (see figure 3.7). The replicate intensities can be aggregated per condition using the `mean`, `mean+std`, `median`, or `mean ranks` function. The `mean+std` method aggregates the replicate intensities per condition using the arithmetic mean and additionally adds the square root of the pooled sample variance (noise level, see section 3.3.2) to each mean value. This method is useful when visualizing noisy or unfiltered data. The `mean ranks` method corresponds to the `mean` function but replaces intensity values by ranks for every candidate profile. When selecting `none`, no aggregation is performed. The resulting intensity vectors can be scaled according to unit Euclidean (`2-norm`), Manhattan (`1-norm`), or maximum norm (`max-norm`), or by calculating z-scores (subtracting the mean value and dividing by the sample standard deviation of condition-specific average intensities). By selecting `none`, no scaling is performed. The aggregated and scaled condition-specific profile vectors are then used for clustering and visualization in MarVis-Cluster (see section 4.3) or MarVis-Pathway. After clustering in MarVis-Cluster, the marker candidates within a cluster may be sorted according to the projection onto the 1D-SOM (and thereby sorted according to similarity of the intensity profiles). This can be used for a finer visualization (see section 4.4.1 the prototype and cluster plot). In MarVis-Pathway, the profiles are automatically projected and visualized in the corresponding order (no cluster structure).

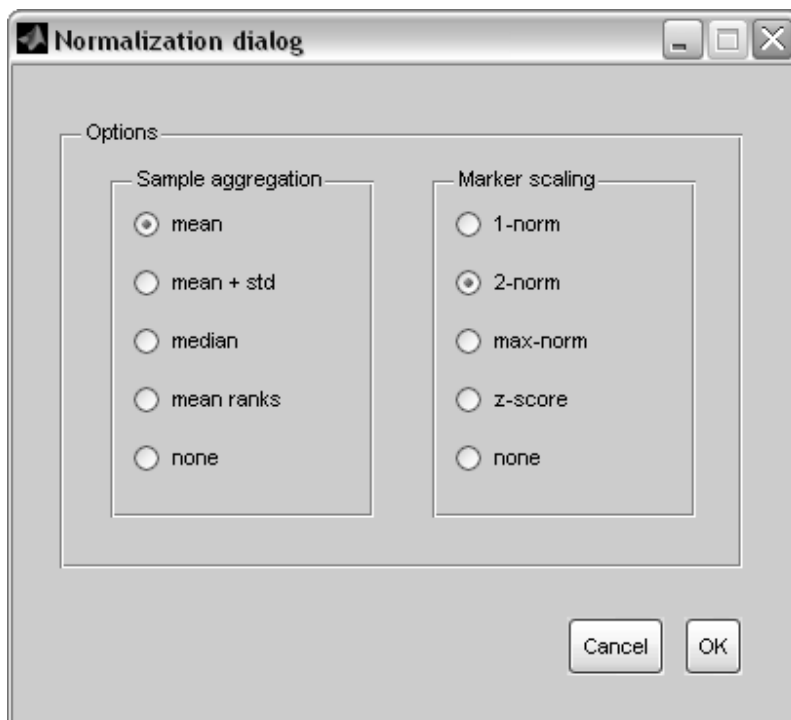


Figure 3.7: Normalization dialog in MarVis-Filter

MarVis-Cluster

4.1 File import

For import of files in MarVis CSV format (see section 3.1) open the `File` menu and select `Open for clustering`. After displaying a file browser dialog for selection of the input file, MarVis will open the `Import` dialog (see figure 4.1) with further options. Here, the delimiter character, the start row and column of the header of the regular data, the number of conditions, and the number of samples for each condition have to be specified. If all conditions have the same number of replicate measurements, this can be specified using a single number (see figure 4.1). Otherwise, the different numbers must be given in order of conditions separated by spaces (e.g. `9 9 9 9 9 9 9` for `dataset1.csv` in the `examples` directory¹). By default, MarVis performs an aggregation of replicate measurements for each condition using the corresponding mean or median value. Furthermore, the resulting intensity vectors are normalized before clustering using e.g. the Euclidean norm. If alternative intensity profiles containing intensities for each marker candidate and condition should be used for clustering, these optional values can be stored after the regular intensity values in succeeding columns. If the checkbox `Import normalized markers` is activated, MarVis tries to import these additional columns and uses them for clustering. If the internal normalization is used, the panel `Sample aggregation and Marker scaling` allows the customization of the aggregation and scaling method. MarVis-Cluster supports the aggregation of replicate measurements for each condition using the mean or median value and the scaling of aggregated marker candidates using the Manhattan norm (1-norm), the Euclidean norm (2-norm), or the z-score transformation (subtracting the mean value and dividing by the sample standard deviation). Note that MarVis displays the marker candidates of selected clusters according to the order of rows in the input file. For example, the candidates in `dataset1.csv` are sorted by retention time, which supports the identification of adducts in MS data sets.

¹The example directory is located within the MarVis-Suite program directory.

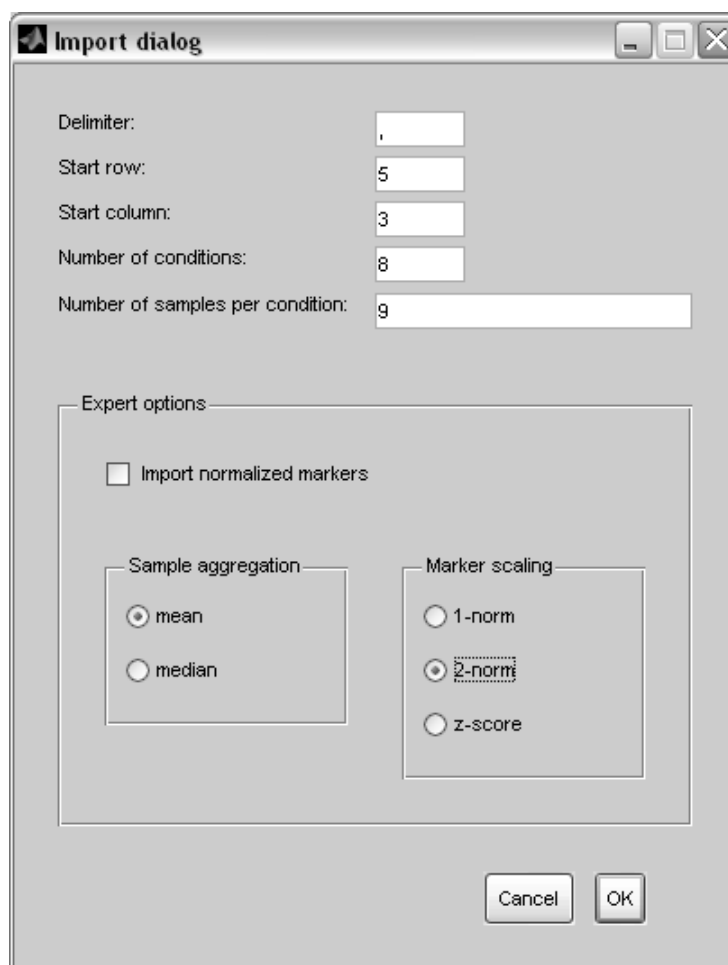


Figure 4.1: Import dialog for dataset1.csv

4.2 File export

Clustering results can be exported to a file in MarVis CSV format. In order to export all clusters, open the File menu and select the menu item `Export clustering results`. After specification of an output file, MarVis will export the marker candidate data (sorted by clusters) in CSV format with additional columns for normalized intensities, cluster number, and intensity profiles of associated prototypes (see example file

dataset1Results.csv). For export of selected candidates only, select the menu item `Export markers` in the `Selection` menu (for interactive selection of clusters and single candidates see section 4.4.2). In both cases, MarVis will use the delimiter character specified in the `Import dialog`.

4.3 Clustering

After file import (or data exchange with MarVis-Filter, or selection of marker candidates for re-clustering, see section 3.6 and 4.4.2), MarVis opens the `Clustering dialog` (see figure 4.2). Here, a data set title and the number of prototypes that should be used for clustering have to be specified. If the checkbox `New window` is activated, MarVis will display the results in a new window, which can be useful for comparison of different clustering results. Click the `OK` button to start the clustering process.

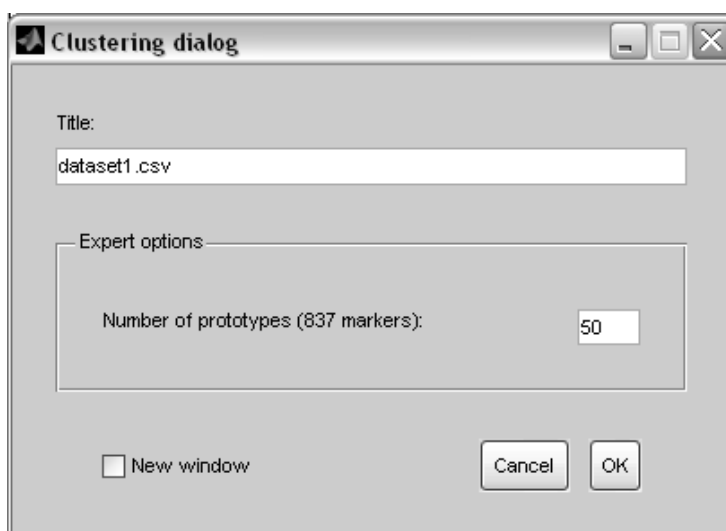


Figure 4.2: Clustering dialog for dataset1.csv

During the clustering process MarVis displays the intermediate clustering results in a new window (see figure 4.3). The red bar in the middle of the window indicates the clustering progress. The upper plot shows the intermediate prototype intensity profiles for each clustering step. The vertical axis represents the data set conditions, the horizontal axis corresponds to the prototype numbers. MarVis uses the current colormap for color-coding. By

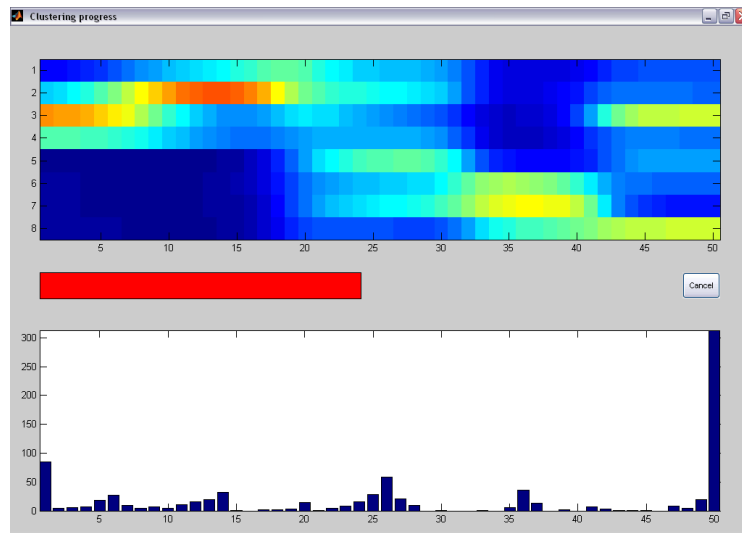


Figure 4.3: Clustering progress window for `dataset1.csv` (intermediate state).

default, the MATLAB[®] Jet map is used (e.g. red colors represent high intensities and blue colors represent low intensities). In the lower plot, the number of marker candidates that are associated with each prototype are represented as vertical bars. The clustering process can be stopped by clicking the `Cancel` button. When the clustering process is finished, MarVis displays a scrollbar instead of the progress bar (see figure 4.4). The slider (the element within the scrollbar) can be used to browse through intermediate clustering results according to different amount of smoothing over the intensity vectors. In most cases, the final clustering state with minimal smoothing is most suitable for analysis. After clicking the `OK` button, MarVis will open the main window. Here, the results according to the selected clustering state are displayed for further analysis.

4.4 Visualization and data analysis

After clustering, the MarVis-Cluster main window displays the results according to the selected clustering state (see figure 4.5). Here, a particular prototype can be selected (“activated”) by clicking on the corresponding column in the heatmap in the upper right region of the window. Afterwards, additional information about the associated cluster is presented (see figure 4.6). In the following section, the different regions of the main window are described in detail.

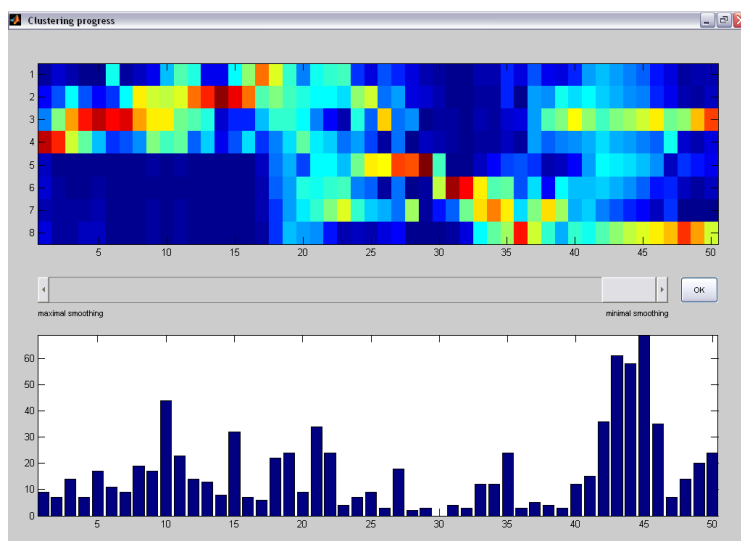


Figure 4.4: Clustering progress window for `dataset1.csv` (final state).

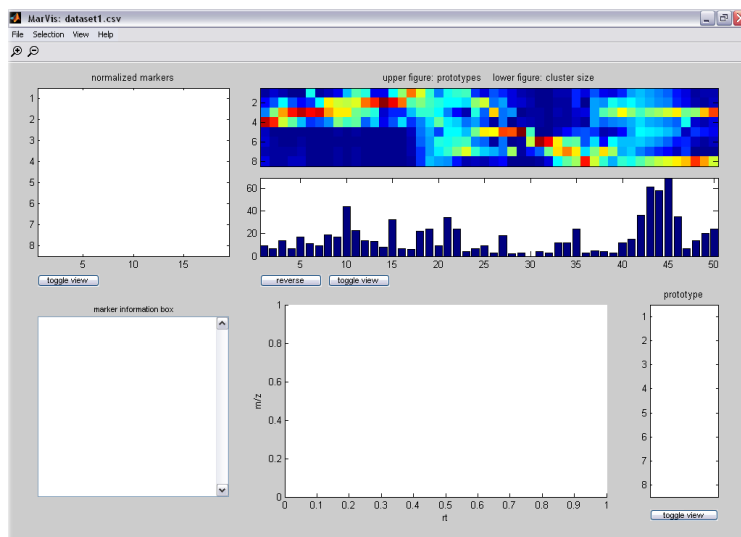


Figure 4.5: Initial MarVis-Cluster main window for `dataset1.csv`

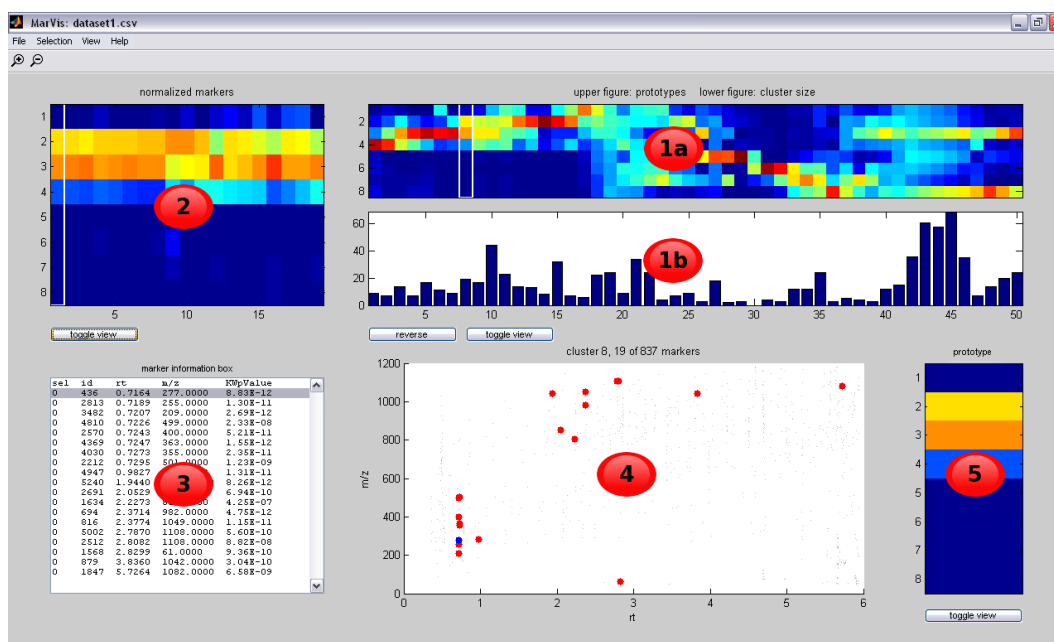


Figure 4.6: MarVis-Cluster main window for `dataset1.csv` after activation of a particular prototype.

4.4.1 Main window

The prototype plot

The prototype plot (see figure 4.6, region 1) shows the ordered prototypes in a heatmap (region 1a) according to the current colormap and additional information about associated clusters (region 1b). By default, the displayed prototype profiles are equally spaced and region 1b shows the associated cluster sizes as a bar diagram (see figure 4.7 a). By clicking on a column corresponding to a particular prototype profile or using the left and right cursor keys, the respective cluster can be activated. MarVis will display further information about the activated cluster in the other regions. A cursor (represented by a white rectangle) marks the current prototype. Clicking the `reverse` button under the left corner of region 1b causes MarVis to reverse the prototype order. This can be helpful when different clustering results should be compared. The graphical representation of the prototypes can be changed via the `toggle view` button. Besides the default view, the prototype profiles can be scaled (in width) according to cluster size, which helps to identify dominating intensity profiles. In this case, region 1b shows the normalized or original intensity profiles of the

marker candidates associated with each prototype (see figure 4.7 b and c). The title above region 1a indicates the currently activated view mode.

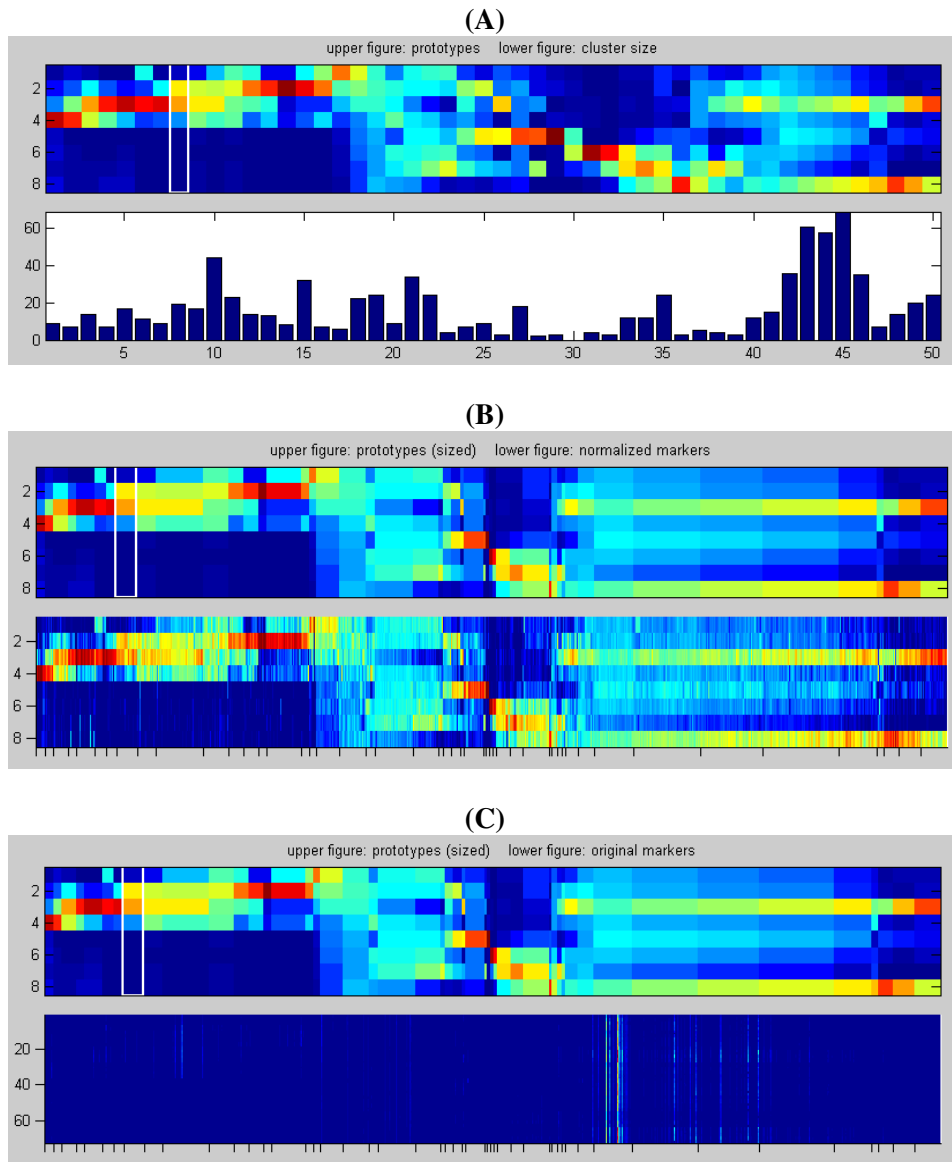


Figure 4.7: Different modes of the prototype plot for dataset1.csv: (A) Equally scaled prototypes and cluster size diagram (B) Scaled prototypes and normalized marker candidate profiles (C) Scaled prototypes and original marker candidate profiles

The cluster plot

The cluster plot (see figure 4.6, region 2) displays the intensity profiles of marker candidates in the activated cluster. Each column represents the intensity profile of a single candidate according to the current colormap. Via the `toggle view` button, the graphical representation of marker intensities can be switched between normalized (aggregated and scaled) and original intensities (see figure 4.8). By default, normalized intensities are shown. Depending on the view mode, the vertical axis corresponds to conditions or replicate measurements of conditions (see figure 4.8). After clicking into the plot, a tooltip displays the marker-specific normalized/original intensity value (m: marker candidate number, c/r: condition/replicate number, i: normalized/original intensity value). A (white) rectangle indicates the current candidate while the cursor keys can be used to browse the plot.

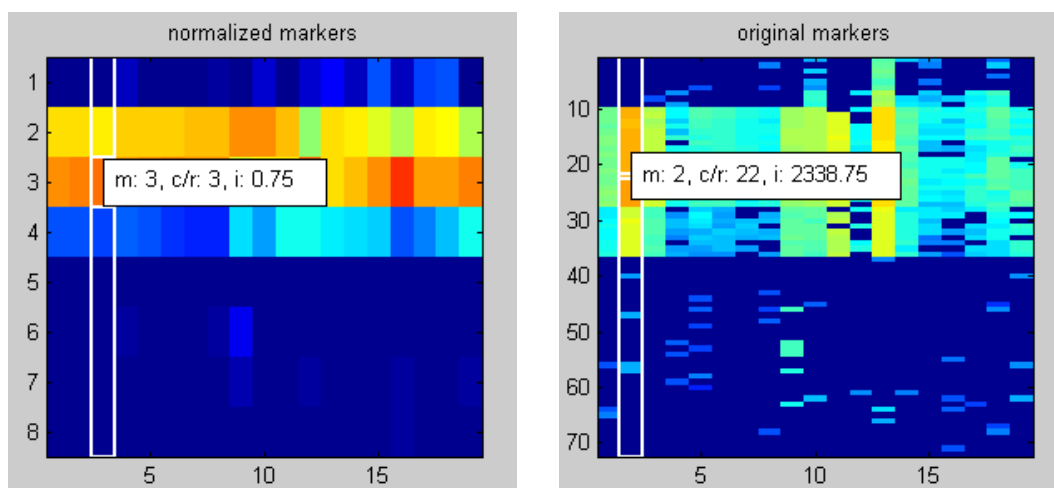
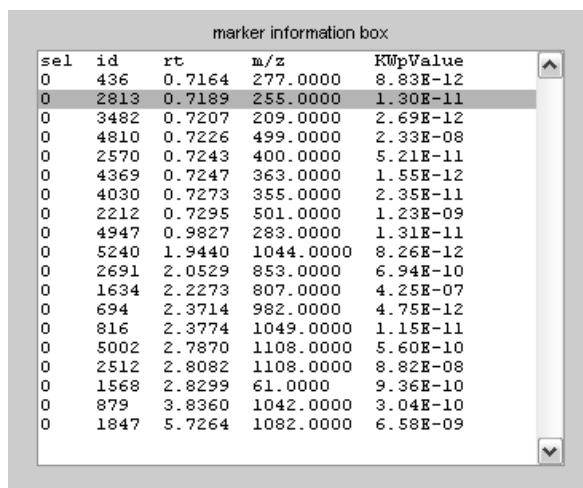


Figure 4.8: Graphical representation of marker candidates of cluster 8 for `dataset1.csv` using normalized intensities (left-hand side, 8 rows corresponding to 8 conditions) and original intensities (right-hand side, 72 rows according to 8 conditions and 9 replicates for each condition).

The marker information box

The marker information box (see figure 4.6, region 3) shows a table containing information about all marker candidates in the currently activated cluster (see figure 4.9). Each candidate is represented by a particular row. Apart from marker ID (second column), x-value (third

column), and y-value (fourth column), also the values for additional data columns (see section 3.1) and the candidate selection status (first column: 0=not selected, 1=selected) are displayed (see section 4.4.2 for details). By using the up and down cursor keys or clicking on a particular row, the corresponding marker candidate can be activated/highlighted. A cursor in the cluster plot represents the currently highlighted candidate.



sel	id	rt	m/z	KmpValue
0	436	0.7164	277.0000	8.83E-12
0	2813	0.7189	255.0000	1.30E-11
0	3482	0.7207	209.0000	2.69E-12
0	4810	0.7226	499.0000	2.33E-08
0	2570	0.7243	400.0000	5.21E-11
0	4369	0.7247	363.0000	1.55E-12
0	4030	0.7273	355.0000	2.35E-11
0	2212	0.7295	501.0000	1.23E-09
0	4947	0.9827	283.0000	1.31E-11
0	5240	1.9440	1044.0000	8.26E-12
0	2691	2.0529	853.0000	6.94E-10
0	1634	2.2273	807.0000	4.25E-07
0	694	2.3714	982.0000	4.75E-12
0	816	2.3774	1049.0000	1.15E-11
0	5002	2.7870	1108.0000	5.60E-10
0	2512	2.8082	1108.0000	8.82E-08
0	1568	2.8299	61.0000	9.36E-10
0	879	3.8360	1042.0000	3.04E-10
0	1847	5.7264	1082.0000	6.58E-09

Figure 4.9: Marker information box for cluster 8 and dataset1.csv

The marker scatter plot

The marker scatter plot (see figure 4.6, region 4) displays the x vs. y-values (e.g. retention time vs. mass, see section 3.1) of all marker candidates in the currently activated cluster using big red dots (see figure 4.10). The currently activated candidate is represented by a big blue dot. The selected candidates in the currently activated cluster are shown using big black dots (see section 4.4.2 for details). In the background, all marker candidates of the underlying data set are displayed as small gray points. By clicking into the plot, a particular candidate can be activated.

The active-prototype/marker plot

The active-prototype/marker plot (see figure 4.6, region 5) displays by default the magnified prototype of the currently activated cluster mapped to the current colormap. Via

the `toggle view` button, the graphical representation mode can be switched between the prototype view, the normalized intensity profile, and the original intensity profile of the currently activated marker candidate (see figure 4.11). By clicking into the plot and navigating with the up and down cursor keys, the original or normalized intensity values can be inspected.

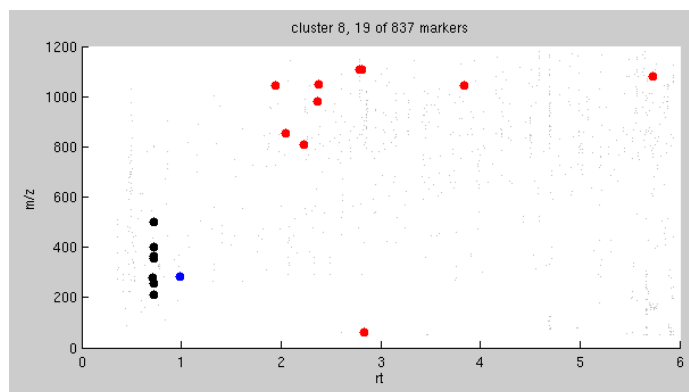


Figure 4.10: Marker scatter plot of cluster 8 for `dataset1.csv`

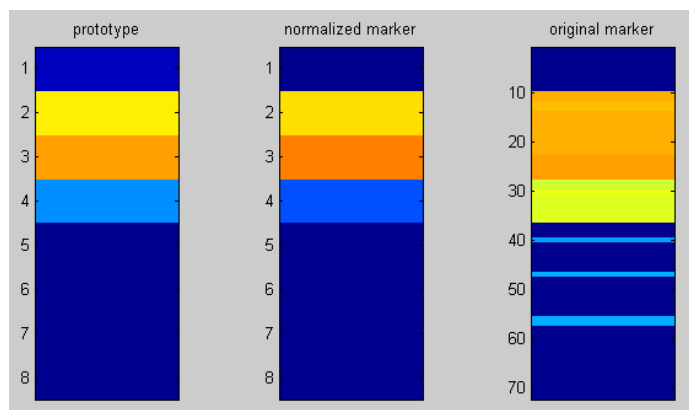


Figure 4.11: Three different views on the active-prototype/marker plot for cluster 8 and marker candidate 2 of data set `dataset1.csv`

4.4.2 Selection of marker candidates

MarVis-Cluster stores a list of selected marker candidates in memory. By pressing `c` in the main window, all candidates of the currently activated cluster are added to/removed from this list. By pressing `m`, only the currently activated candidate is affected. The menu items `Select/Deselect cluster` and `Select/Deselect marker` from the `Selection` menu can be used for this purpose, too. By choosing the entries `Select all` or `Reset all` from the `Selection` menu, all marker candidates from all clusters can be selected or deselected. By means of the `Select block` function, all marker candidates in clusters between the first selected candidate (starting on the left-hand side of the prototype plot) and the currently activated cluster are selected. This function is especially useful when selecting many small neighboring clusters. The selection status of all marker candidates can be inverted by means of the `Invert selection` entry. The function `Filter selection` can be used to filter the currently selected candidates based on one of the MarVis-Filter ranking methods (see section 3.3) and a given threshold for the corresponding ranking score (e.g. fold-change ratio). After specifying a threshold, MarVis-Cluster deselects all candidates below or above the threshold (depending on the filter method, e.g. only candidates with a fold change above 2 are kept selected). This function can be used to reduce the selection step by step starting with the selection of all candidates.

Selected candidates are highlighted in the marker information box (entry 1 in the selection indicator column) and in the scatter plot by big black dots (see figure 4.10). Clusters that contain selected marker candidates are highlighted by a black line below the corresponding prototype in the prototype plot (see figure 4.12). The list of selected candidates can be exported, re-clustered, or combined in a new window via the entries `Export markers`, `Recluster markers ...`, or `Combine markers ...` in the `Selection` menu (see also section 4.2). In the latter case, all selected candidates are presented in a new MarVis-Cluster main window within a single cluster. For re-clustering, the user can select a subset of conditions used for clustering (e.g. `1:8` or `1 2 3 4 5 6 7 8` for the first eight conditions). After re-clustering, the marker candidates within a cluster may be sorted according to the projection onto the 1D-SOM (and thereby sorted according to similarity of the intensity profiles). This can be utilized for a finer visualization (see section 4.4.1 the prototype and cluster plot).

By means of the `Set marker labels` function in the `Selection` menu, se-

lected marker candidates can be annotated with a user-defined label (e.g. “wt-specific“). In this case, another column with the header `MarkerLabel` is added to the additional data. For all selected candidates, the corresponding entries are set to the specified label. If marker labels are defined for the first time, all marker candidates which are currently not selected are assigned the label ”-“. The specified marker labels can be exported along with the data set (see section 4.2) or utilized in MarVis-Pathway for coloring of entries in pathway maps (see section 4.5 and 5.3.1). The marker labels may be removed using the function `Reset marker labels`.

Marker candidates may be selected based on discrete values/labels by means of the `Select labeled markers` function in the `Selection` menu. The column which contains these label information can be chosen from a list including the marker labels described in the previous paragraph and other additional columns which seem to contain discrete values/labels (not more than 25 distinct values, e.g. the original data set index in a combined data set). In the following listbox, the user can choose values from the previously selected column. All marker candidates which are associated with one of these values (in the respective column) will be selected. This function does not change the status of previously selected candidates.

By means of the `Label barplot` function in the `Selection` menu, a barplot which shows the distribution of particular labels/values over all clusters can be generated. This plot is similar to the cluster size diagram (see lower part of the prototype plot in figure 4.7) but with colored bars split into different groups of labels/values. It is displayed instead of the cluster size diagram in the lower part of the prototype plot (reset plot by clicking the `toggle view` button). First, a data set column containing discrete values/labels has to be specified (e.g. marker labels or data set indices). Second, the discrete values have to be grouped by selecting a single or multiple values representing the first group, selecting values representing the second group, and so on. For each cluster and each group, the marker candidates which are associated with the corresponding labels are then counted and plotted as colored proportions of the corresponding bar. The proportions are colored according to the current colormap (e.g. dark blue for the first group and dark red for the last group). Figure 4.13 shows the label plot for the data set indices of `wound_neg_raw.csv` (label 1) and `wound_pos_raw.csv` (label 2) after filtering/combination in MarVis-Filter and clustering with MarVis-Cluster. The number of features/marker candidates for each cluster may be normalized according to the overall number of features in the corresponding group. In this case, the barplot shows the percentage of features in the corresponding group

found in the respective cluster.

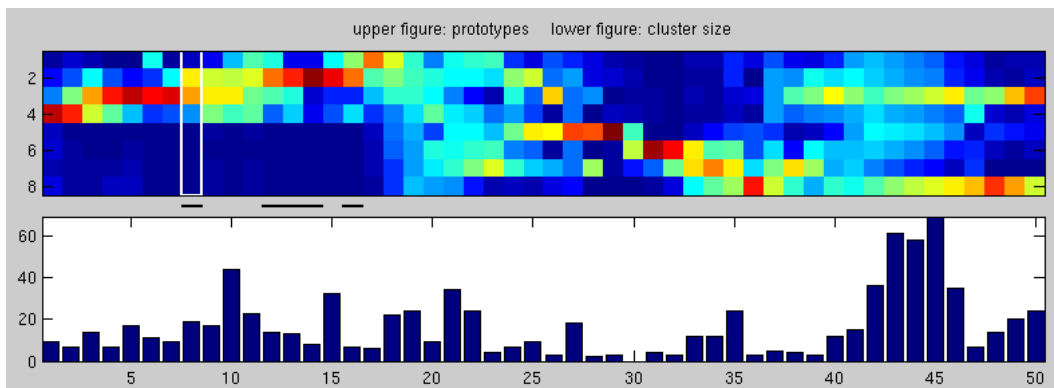


Figure 4.12: Clusters with selected marker candidates from `dataset1.csv`. Clusters that contain selected marker candidates are marked by a black line below the corresponding column in the prototype plot.

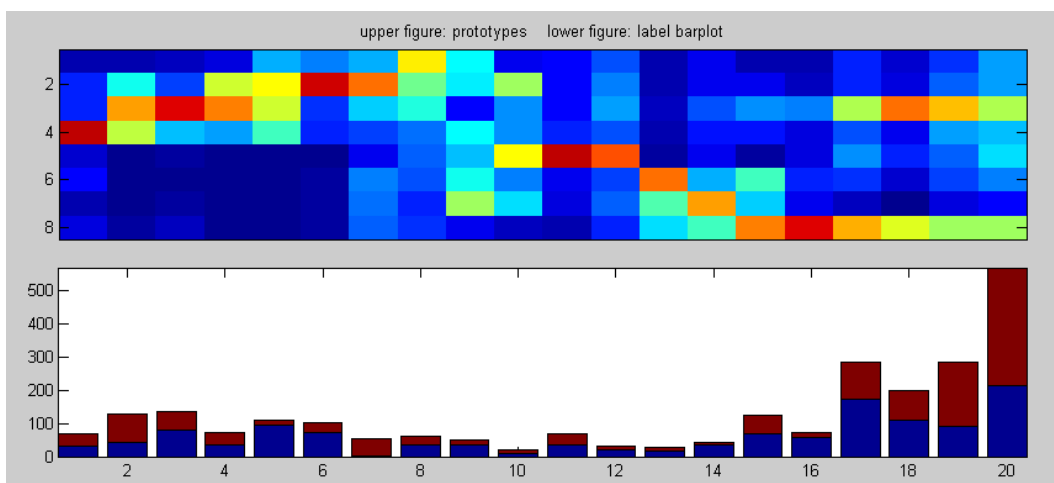


Figure 4.13: Data set label barplot for the data sets `wound_neg_raw.csv` and `wound_pos_raw.csv` after filtering, combination, and clustering: For each cluster, the dark blue proportions of the bars correspond to the number of candidates from the data set obtained in negative ionization mode and the dark red proportions correspond to the candidates from the positive ionization mode.

4.4.3 General visualization properties

Via the `View` menu, general visualization properties of MarVis-Cluster can be changed. By clicking the `Colormap editor` menu item, the MATLAB[®] colormap editor is opened. The colormap editor displays the current colormap in the upper region of the dialog. MarVis plots the intensity values according to this map. Low intensity values are mapped to the colors on the left-hand side of the color spectrum (by default blue). High intensity values are mapped to the colors on the right-hand side of the spectrum (by default red). Intensities between the minimum and the maximum are mapped based on a linear projection onto the colormap. Color values for the original, log-transformed, and normalized intensity profiles are calculated independently according to their global minimum and maximum values. The reference marks below the spectrum can be moved to the left or to the right to change the color contrast. By clicking below the spectrum, further marks can be added for a finer contrast adjustment. Via the `Standard colormaps` menu item in the `Tools` menu, one of the MATLAB[®] standard colormaps can be selected. By default, MarVis uses the MATLAB[®] Jet map. If the checkbox `Immediate apply` is activated, the changes are immediately applied. By clicking `OK`, the dialog is closed and the changes are applied. Use the `Cancel` button to close the dialog without applying any changes (`Immediate apply` must be deselected). Clicking the `Apply` button applies the colormap to all intensity profile plots in the main window. View the online MATLAB[®] colormap editor or colormap help pages for further information.

Note: Several MATLAB[®] functions of the colormap editor are not integrated in MarVis and the application may produce an error message.

Select the `Cursor color editor` entry in the `View` menu to change the color for the cursor representation. This can be useful to improve distinguishability between the cursor (usually a white rectangle) and the intensity profiles, e.g. if a new colormap was selected.

If the `Logarithmic intensities` checkbox within the `View` menu is activated, MarVis calculates the color mapping for original intensity profiles logarithmically. This can be useful if the data set contains very large intensities (see figure 4.14). The color mapping for normalized intensity profiles is not affected. By default, this checkbox is deactivated.

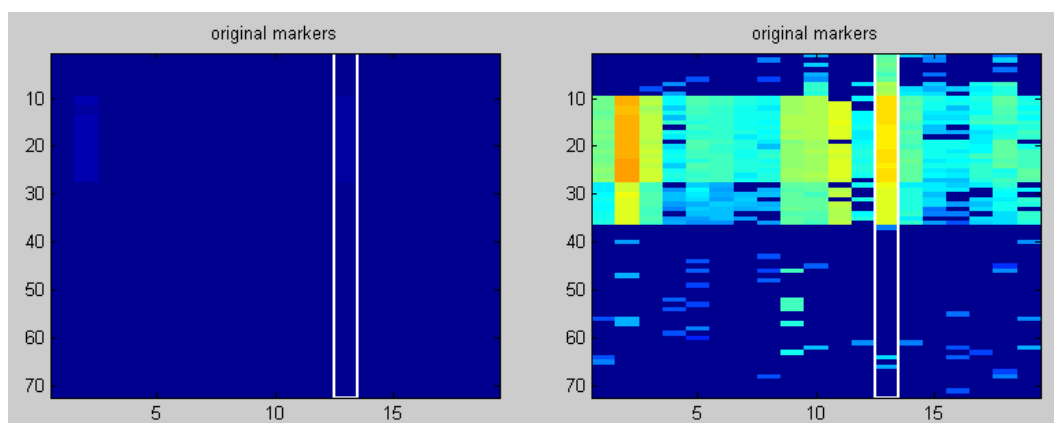


Figure 4.14: Original intensities of marker candidates from `dataset1.csv` in a cluster without logarithmic scaling (left-hand side) and with logarithmic scaling (right-hand side).

4.5 Data exchange with MarVis-Pathway

The selected marker candidates can be analyzed in MarVis-Pathway (see chapter 5) using the `Goto MarVis-Pathway` entry in the MarVis-Suite menu. In MarVis-Pathway, the marker candidate profiles are visualized according to the projection onto the 1D-SOM used in MarVis-Cluster or a new one (user's choice). If marker candidates were labeled (see section 4.4.2), a distinct color can be selected for each specified label. After database query, entries (e.g. metabolites or genes) exclusively mapped to candidates associated with a particular label are colored accordingly. Entries mapped to data set features associated with different labels or labeled and unlabeled features are by default colored in gray (see section 5.3.1).

4.6 Example data sets

In the following, two example data sets are introduced, which can be used to explore the capabilities of MarVis-Cluster in the context of a metabolomic and a gene expression study.

4.6.1 The wound response of *Arabidopsis thaliana*: A metabolomic case study

Data set one (file `examples/dataset1.csv`) contains 837 metabolite marker candidates for the wound response of the thale cress *Arabidopsis thaliana* under 8 conditions,

which correspond to a time course after wounding for wild type (wt) plants (conditions 1-4) and the jasmonate-deficient *dde2-2* mutant plants (conditions 5-8) [1]. Each condition contains 9 replicate samples. The data file starts with 4 comment rows and 2 comment columns, which contain explanatory information about the file format. The header information is contained in several columns within the fifth row starting at column three. The third column contains unique integers representing marker candidate IDs. Columns four (x-values) and five (y-values) contain the retention times (rt) and the mass-to-charge-ratios (m/z) from liquid chromatography/mass spectrometry (LC/MS) analysis, respectively. These columns are followed by the 72 intensity data columns (number 6 to 77, sorted by conditions and replicate samples). As an additional column (number 78, KWpValue), the p-values of a Kruskal-Wallis test were added. These values were used as a measure of marker-specific quality in [1]. In this data set, the columns are separated by comma. In the previous chapters, this data set was already used to demonstrate particular features of MarVis.

Figure 4.15 shows the main window of MarVis-Cluster after clustering of this data set using 50 prototypes and minimal smoothing (final clustering state). Intensity profiles were aggregated and scaled using the mean value and the Euclidean norm (2-norm), respectively. The prototype plot reveals

- a block of marker candidates that show high intensities in the conditions representing wt plants only (prototype 1 to 18, condition 1 to 4),
- an intermediate block of different profiles representing high intensities across wt and *dde2-2* mutant plants (prototype 19 to 24),
- a block of prototypes that show high intensities in mutant plants only (prototype 27 to 36, condition 5 to 8),
- and a block of marker candidates which particularly represent high concentrations in the third and eighth condition (prototype 40 to 50).

The corresponding bar diagram shows a number of clusters that contain just a few or no marker candidates at all (e.g. cluster 30). These "sparse" clusters usually indicate the use of too many prototypes. Despite the large number of prototypes, candidates with similar intensity profiles can be found in neighboring clusters.

For better interpretability, all clusters except for clusters 19 to 24, which show quite

indistinct prototype profiles, were selected (item `Select all` in the `Selection` menu and pressing `c` after activating clusters 19 to 24). The selected marker candidates are clustered once again via the `Recluster markers` entry in the `Selection` menu. This time, a lower number of prototypes (30) is used. The two different clustering results can be compared if the `New window checkbox` in the `Clustering` dialog is activated. The file `dataset1Results.csv` in the `examples` directory contains the exported clustering results in CSV format.

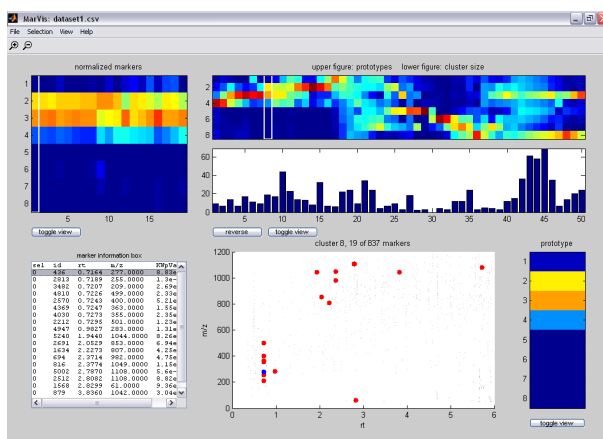


Figure 4.15: MarVis-Cluster main window for dataset1.csv using 50 prototypes.

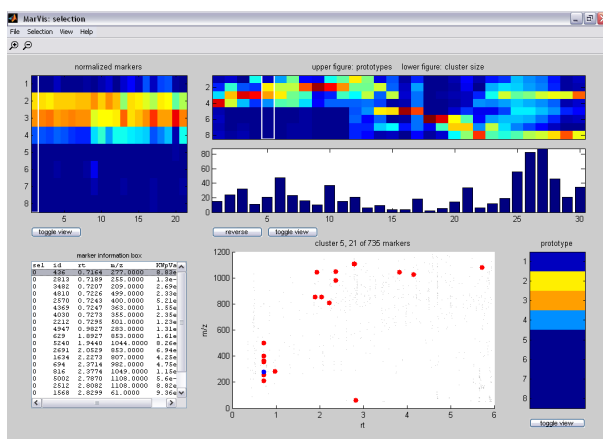


Figure 4.16: MarVis-Cluster main window for the selection of dataset1.csv using 30 prototypes.

4.6.2 The yeast cell cycle: A gene expression experiment

Data set two (file `examples/dataset2.csv`) contains the (normalized) expression levels of 384 selected yeast genes over two cell cycles represented by 17 time points (which are used as conditions). The 384 genes were selected from the original data set from [23] based on expression level peaks at different time points representing the five phases of cell cycle [24]². To obtain suitable x and y-values, the loadings of the first two principal components (PC1 and PC2) were calculated. As IDs the ORF (Open Reading Frame) identifiers from the original data set were used. The information about the dominant group/phase of each gene was retained as an additional data column (last column). Additionally, four comment rows and two comment columns (start of the header in row 5 and column 3) were added. All columns are separated by a comma as delimiter character.

The data set contains no replicate measurements for the conditions (insert 1 as number of samples per condition), therefore the selected aggregation method has no effect. The Euclidean norm (2-norm) was used for normalization and the data set was clustered using minimal smoothing and 20 prototypes. Figure 4.17 shows the main window containing the clustering results. The prototype plot reveals roughly two parallel diagonal lines of high expression levels, which represent two cell cycles. Due to the ordering, a number of adjacent prototypes with very similar expression profiles can be identified (e.g. number 1 and 2 or 13 and 14). The prototype number 9 seems to be alien to the adjacent prototypes (see figure 4.17, cursor in the prototype plot). The associated cluster contains only eight marker candidates, which show quite different expression profiles. This and the similarity of adjacent prototypes indicate a much lower adequate number of clusters. Despite the overestimated number of clusters, the order of prototypes reflects the five phases of the cell cycle (see additional column `phase`). Cluster 1 to 5 contain mainly marker candidates associated with phase 2, cluster 6 to 8 represent mainly candidates of phase 3, cluster 9 to 12 are generally associated with phase 4, cluster 13 to 15 represent mainly phase 5, and cluster 16 to 20 contain primarily candidates of phase 1. When the data set is clustered using a lower number of prototypes (e.g. 10 or 5, see figure 4.18), this effect becomes even stronger.

²The original data set of selected genes can be downloaded from <http://faculty.washington.edu/kayee/model/>

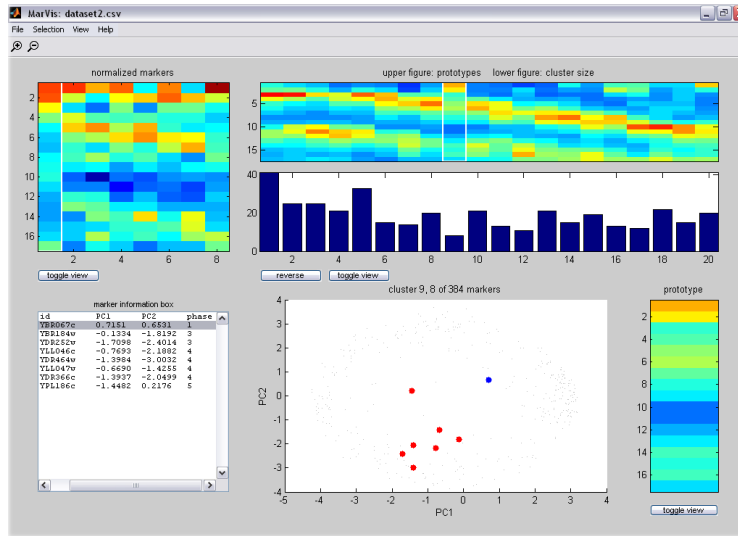


Figure 4.17: MarVis-Cluster main window for dataset2.csv using 20 prototypes. The “alien” prototype number 9 is highlighted by the white cursor in the prototype plot.

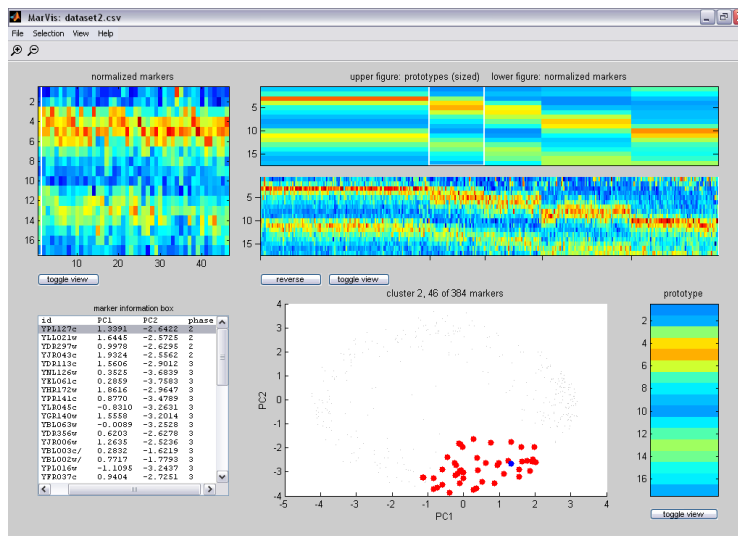


Figure 4.18: Main window for dataset2.csv using 5 prototypes (prototypes are scaled according to cluster size and the lower box of the prototype plot displays normalized marker candidate profiles).

MarVis-Pathway

5.1 Data import

For data import, MarVis-Pathway utilizes the import capabilities of MarVis-Filter and MarVis-Cluster (see section 3.1 and 4.1). Filtered and selected data can be analyzed by means of MarVis-Pathway via the `Goto MarVis-Pathway` function in the respective MarVis-Suite menu (see section 3.6 and 4.5). Additionally, previously saved MarVis-Pathway projects may be loaded via the `Load project` function in the File menu (see section 6.2).

5.2 Database selection and query

In the current version, MarVis-Pathway provides pathway databases from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the BioCyc collection [3, 4]. The databases are either integrated in the MarVis-Suite program directory (subfolder `local`) or can be downloaded separately from the project home page. In the latter case, the folder where the database files have been extracted has to be specified when running MarVis-Pathway for the first time. The local database folder can be changed by means of the `Set local database dir` function in the Analysis menu.

The included KEGG collection¹ contains one reference (Reference pathways) and about 3000 organism-specific databases. The included BioCyc collection² provides about the same number of organism-specific databases and one reference database (MetaCyc). Each KEGG reference pathway is associated with a number of compound, EC (Enzyme Commission), and KO (KEGG ORTHOLOGY) IDs (e.g C00084, K00001, or 3.2.1.86) and names. Each MetaCyc reference pathway variant is associated with a

¹KEGG FTP Release Dec 9, 2013, <http://www.kegg.jp>

²biocyc-17.5, <http://biocyc.org/>

number of compound and EC IDs/names. The databases in both collections also contain the monoisotopic masses for all compounds (which may be used for annotation of MS data sets). In case of the organism-specific databases, the pathways are associated with compound IDs/names/masses and gene IDs/names (e.g. AT5G42650 or AOS for *A. thaliana*) instead of the EC and KO numbers. Each pathway can be considered as a set of associated entries (see also section 5.4).

Additionally, customized databases may be loaded from CSV files. The first row in such a file is reserved for header information (e.g. column labels). Each following row contains information about the association of an entry (e.g. compound or gene) with a set (e.g. pathway). The first column is reserved for entry IDs, the second and third column for retention time and compound mass values (insert zeros for genes or enzymes), and the fourth column for the corresponding entry names. The fifth column is reserved for the IDs of associated sets/pathways, followed by the set/pathway name in the sixth column. The last column is reserved for molecular formulas of the respective compounds (insert '-' if no formula is available, e.g. for genes).

Note that entry IDs should not contain the characters '/' or ':', which are reserved for entry type definitions (see also section 5.4.1).

5.2.1 Database selection

After switching to MarVis-Pathway or using the Database lookup function in the Analysis menu, the databases which should be queried have to be selected in the Database selection dialog (see figure 5.1). The user can select one or both of the collections from the listbox `Top level order` on the left-hand side of the dialog. The available databases in the selected collection(s) are presented in the middle listbox with the title `Databases`. For both collections, the first entry represents the respective reference database (`MetaCyc` for the `BioCyc` collection and `Reference pathways` for `KEGG`) followed by the organism-specific databases sorted in alphabetic order of their names. Single entries can be marked by clicking into the listbox or using the arrow keys, multiple entries may be marked by holding the `Control` or `Shift` key. Marked databases can be selected by clicking the `Select` button. The corresponding entries are then listed in the `Selection` listbox on the right-hand side. The databases (in the selected collections) may be searched for a particular organism by entering the corresponding name, abbreviation, or substring in the textfield below the listbox and clicking the `Find` button.

The Database listbox shows then only entries which contain the entered text in their names. All entries can be restored by clicking onto the collection entries (left-hand side) once again. Marked entries in the Selection listbox can be deselected by clicking on the Remove button. By clicking the OK button, the dialog is closed and all selected databases will be used for query. In the following file dialogs, additional databases may be loaded from CSV files (see start of section). By clicking the Cancel button, the selection of additional CSV files is finished.

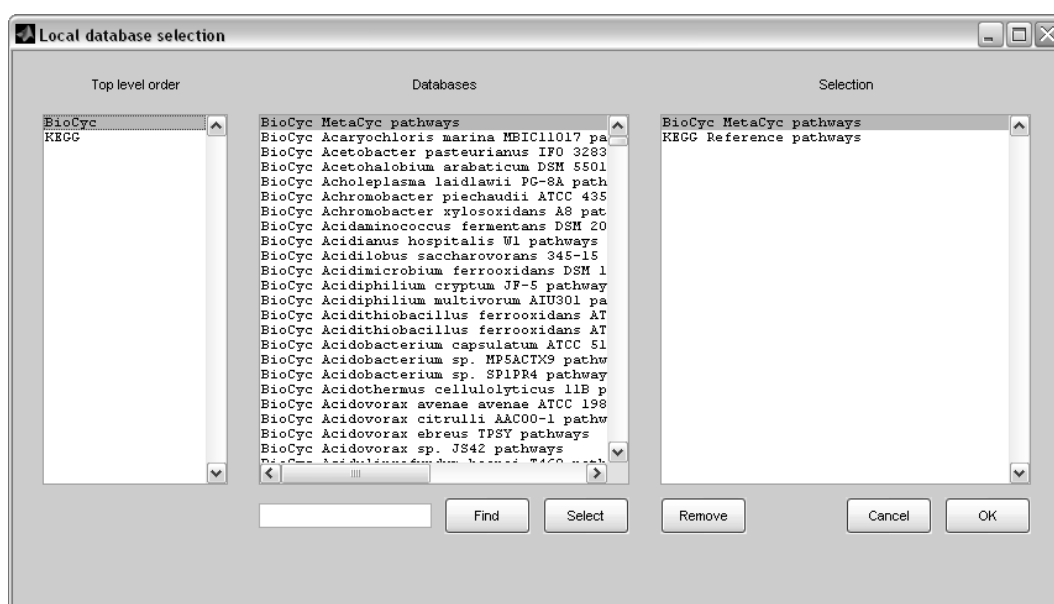


Figure 5.1: MarVis-Pathway database selection dialog

5.2.2 Entry mapping

After database selection, the mapping method for each data set has to be specified. The marker candidates in a particular data set may be mapped to the database entries based on their IDs and the entry ID/names or based on their corrected monoisotopic masses (y-values, see section 3.1 and 3.5) and the entry masses. In the Match entry IDs for data sets dialog, the user can specify the data sets which should be mapped based on marker candidate ID (as imported with the data set) and entry ID (as stored in the databases, e.g. KO/EC numbers or gene IDs). In the Match entry names for data sets dialog, the data sets which should be mapped based on marker candidate ID and entry name

(e.g. common gene name) can be selected. Multiple data sets can be selected by holding the `Control` or `Shift` key. The ID and name matching is case-insensitive. Note that in contrast to the ID matching, the name matching is error-prone (especially for compound names). By clicking the `Cancel` button, no data sets are selected for ID or name matching. Data sets which have not been selected are used for mass-based mapping. In this case, the user has to specify a mass tolerance (e.g. 0.005 Da) and a single or several correction factors in the `Mass matching` dialog. A marker candidate is mapped to a database entry if the absolute mass difference (marker candidate y -value minus entry mass) is below the specified tolerance. The correction factors are added to each marker candidate mass (y -value) before calculating the difference. By this means, multiple corrected masses per feature can be used for matching. For data sets corrected for adducts and isotopes (see section 3.5), the correction term should be set to 0. For uncorrected MS data these factors may represent different ionization rules (e.g. 1.0078 or -1.0078 for negative/positive ionization mode and $[m - H]^-$ or $[m + H]^+$). Different correction factors (e.g. corresponding to adducts) may be specified separated by space characters.

5.2.3 Scoring of pathways

After specifying the method(s) for marker candidate mapping, different options for the scoring and ranking of pathways which contain matched entries can be selected in the `Ranking options` dialog (see figure 5.2). For each pathway p , a score

$$S_p = \sum_m \sum_e c_{m,e,p} a_{m,e,p} W_{m,e} \quad (5.1)$$

is calculated. $W_{m,e}$ defines a positive weight for each assignment of a marker candidate m to an entry e . $a_{m,e,p}$ is a boolean assignment variable: $a_{m,e,p} = 1$ if marker candidate m was mapped to entry e and e is associated with pathway p , $a_{m,e,p} = 0$ otherwise. $c_{m,e,p}$ is a scaling factor for normalization.

If the radio button `count` in the `Marker score` box is activated, $W_{m,e} = 1$ for all assignments. If the radio button `rank` is selected, the marker candidate ranking from MarVis-Filter (see section 3.3) is utilized. In this case, the weights are calculated as rank score $W_{m,e} = 1 - \frac{R_m}{N_d}$, where R_m denotes the marker candidate rank in the (first) MarVis-Filter ranking and N_d the number of candidates in the corresponding original data set. This option is useful when annotating unfiltered data sets.

The following options can be used in order to normalize the pathway scores according to the number of entries a particular marker candidate matches ($M_{m,p}$, option `Marker hit normalization`) and/or the number of candidates a particular database entry is mapped to ($E_{e,p}$, option `Entry hit normalization`). The global normalization factor $c_{m,e,p}$ (see equation 5.1) is defined as a function of $M_{m,p}$ and $E_{e,p}$. If the `max` radio button in the `Overall normalization box` is activated,

$$c_{m,e,p} = \frac{1}{\max(M_{m,p}, E_{e,p})}. \quad (5.2)$$

If the `product` radio button is selected,

$$c_{m,e,p} = \frac{1}{M_{m,p} E_{e,p}}. \quad (5.3)$$

If the `mean` radio button is selected,

$$c_{m,e,p} = \frac{1}{\text{mean}(M_{m,p}, E_{e,p})}. \quad (5.4)$$

If the option `local` in the `Marker hit normalization box` is activated, $M_{m,p}$ is calculated as the number of assignments of marker candidate m to different entries in pathway p . In case `global` is selected, $M_{m,p}$ is the number of assignments of marker candidate m to different entry-pathway pairs (in this case, $M_{m,p}$ is independent of p). If the option `local` in the `Entry hit normalization box` is activated, $E_{e,p}$ is defined as the number of assignments of marker candidates to entry e in pathway p . In case `global` is selected, $E_{e,p}$ is the number of assignments of marker candidates to entry e over all pathways. If the option `none` is selected, $M_{m,p} = 1$ and $E_{e,p} = 1$, respectively.

Additionally, the pathway scores may be normalized according to the number of entries associated with each pathway (checkbox `normalize by set size` selected). This option in combination with the entry-based normalization is useful when searching for pathways which show a high coverage by experimental evidence.

If the `count` option is activated and no normalization method is selected (`none` for `Marker hit normalization` and `Entry hit normalization`), the pathway scores correspond to the number of assignments per pathway. If the `Marker hit normalization` is set to `local`, the pathway scores equal the number of unique marker candidate hits per pathway. In case the `Marker hit normalization` and

Entry hit normalization are set to local and the max option is activated in the Overall normalization box (default options), the pathway scores are normalized for marker candidates which match multiple entries in one pathway (e.g. because the entries have the same mass) and for entries which are associated with multiple candidates (e.g. due to different ionizations in MS analysis). The global option and product function are useful when searching for pathways which are exclusively associated with a set of entries which are exclusively matched by a set of marker candidates.

After specifying the ranking options and clicking the OK button, MarVis-Pathway performs the database query and scoring of pathways. The results are presented in the MarVis-Pathway main window (see figure 5.3).

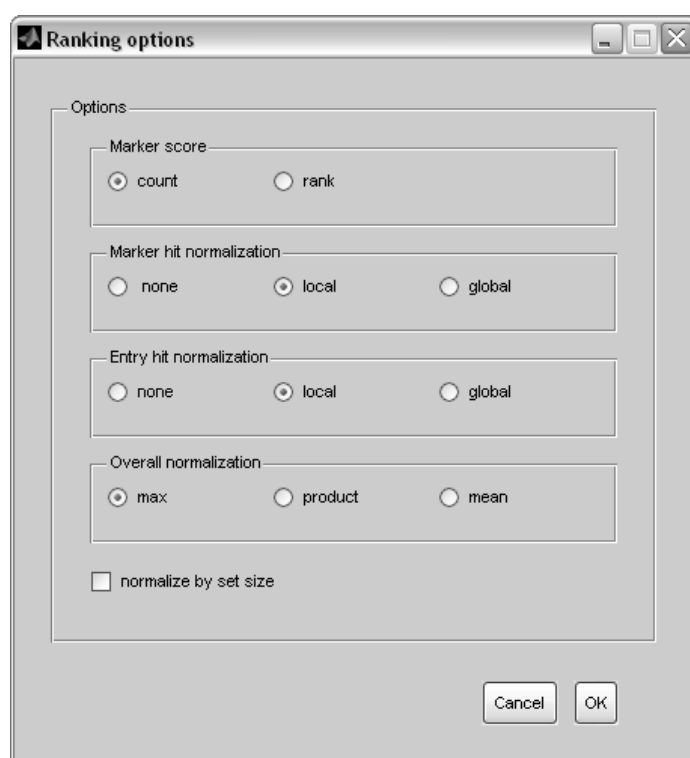


Figure 5.2: MarVis-Pathway ranking options dialog (default options)

The pathway/set information box

The pathway/set information box (region 2) contains additional information regarding the currently selected pathway. For the KEGG and BioCyc collection, this includes information about the flat files used for database construction and copyright details.

The marker profile map

The marker profile map (region 3) shows the heatmap of marker candidate profiles (columns) which could be mapped to the current pathway. Each row represents one of the experimental conditions in the current order. The profiles are sorted according to the projection onto the 1D-SOM used for clustering. By default, the average intensities per condition are represented by colors from the MATLAB[®] Jet colormap (e.g. red colors represent high intensities and blue colors represent low intensities). Single marker candidates may be selected by clicking into the plot or using the left and right arrow keys. A white rectangular cursor marks the current candidate (profile).

The entry assignment list box

The entry assignment list box (region 4) shows the assignments of marker candidates to entries in the selected pathway sorted according to entry name. For each assignment, the first column contains the color which is used to represent the corresponding entry in pathway maps (e.g. pathway maps on the KEGG web site) followed by the corresponding marker candidate ID, the associated entry name and ID, the marker candidate x and y-value, the correction factor used for mass matching, and additional marker-specific information (e.g. annotations from MarVis-Filter). A particular assignment can be selected by clicking into the list box or using the up and down arrow keys. After selecting an assignment, the corresponding marker candidate profile is highlighted in the marker profile map (region 3). When selecting a profile in the marker profile map, the first assignment of the corresponding marker candidate is highlighted in the assignment list box.

By holding the `Control` key and pressing one of the color keys (`r`: red, `b`: blue, `g`: green, `y`: yellow, `o`: orange, `m`: magenta, `a`: gray), the color for the entry associated with the currently selected assignment can be changed. This function and additionally the selection of customized RGB colors are also available via the `Set entry color` item in the `Analysis` menu. The dialog for customizing the color is shown after canceling the first dialog. Via the `Set colors` item in the `Analysis` menu, the color can be changed for all entries.

After clicking the `Map` button below the list box, MarVis-Pathway opens the corresponding

online pathway map in a new browser window or tab. In case of pathways from the KEGG collection, the entries in the pathway map are marked in the chosen colors (by default red). For the KEGG reference pathways, the user may specify an KEGG organism code (e.g. 'ath' for *Arabidopsis thaliana*). In this case, the organism-specific pathway with all matching entries is shown. Leave the corresponding text field empty or insert 'map' in order to inspect the reference pathway.

After clicking the `Entry` button, the online resources for the corresponding entry are presented in a new browser window or tab.

The marker profile plot

The marker profile plot (region 5) shows the raw intensity profile of the marker candidate associated with the currently selected assignment as bar plot. Intensity values of replicate samples for the same experimental condition are marked in the same color. The intensities are sorted according to conditions and the current sample order.

The related pathways/sets list box

The related pathways/sets list box (region 6) shows all pathways that contain entries mapped to the marker candidate which is currently selected in the marker profile map (region 3) (or which is associated with the currently selected assignment). The list items contain the same information as presented in the pathway/set list box (region 1) and are in the same order. If the current marker candidate is mapped to more than one entry in the same pathway, the corresponding row is repeated. By clicking onto one of the rows, the corresponding pathway, assignment, and marker candidate is selected and highlighted. When activating the `Hold` checkbox below the list box, the current contents are retained when a new pathway/candidate/assignment is selected and the user can browse the list. This is especially useful when checking different assignments of a single marker candidate.

5.3.2 Pathway and entry search

By means of the `Search set list` and `Search entry list` in the `Analysis` menu, the user can search the results of database query for pathway/entry names and IDs. In the first case, the rows of the pathway/set list box are searched for a given substring (case-insensitive) starting with the next pathway in the list. In the second case, the entry assignments are searched starting with the next pathway in the pathway/set list box. Additionally, the user can search for particular marker candidates via the `Find markers . . .` function in the `Analysis` menu (see section 6.3).

5.4 Set Enrichment Analysis

For statistical analysis of pathway sets containing matched entries, MarVis-Pathway provides an extensive framework for (Gene/Metabolite) Set Enrichment Analysis (SEA) [5, 6] (see function `Set Enrichment Analysis` in the `Analysis` menu). SEA intends to identify sets of database entries (e.g. metabolic pathways) which are enriched in high-ranked (e.g. highly differential) data set features mapped to the corresponding entries.

The SEA framework in MarVis-Pathway offers three different types of enrichment analysis: The entry-based, marker/feature-based, and sample-based analysis. In the first case, the number of entries in a pathway matched by the filtered/selected marker candidates (e.g. in MarVis-Filter or MarVis-Cluster) in comparison to the number of entries which could be matched over all pathways is evaluated based on a hypergeometric distribution. This method is useful when dealing with small or strictly filtered data sets in comparison to large organism-specific databases.

The marker/feature-based enrichment analysis uses the ranks of marker candidates (as calculated in MarVis-Filter) which match entries in a particular pathway. For statistical evaluation, a static or iterative hypergeometric test, a rank-sum, or a Kolmogorov-Smirnov test is applied. This type of analysis relies on data sets containing a large number of marker candidates (but not necessarily large databases). The method is able to incorporate information from adduct and isotope correction performed in MarVis-Filter (see section 3.5).

The sample-based enrichment analysis uses the ranks of marker candidates which are recalculated for a large number of random permutations of sample condition labels (assignments of samples to experimental conditions). For (re-)ranking, the signal-to-noise ratio (see section 3.3.2) is used. This method does not depend on the assumption of independent marker candidates or independent database entries but requires a sufficiently high number of independent replicate samples and considerably more computing time in comparison the first two methods. In the following sections, the three types of analysis and the combination of results from multiple data sets in a meta-analysis are described in detail.

5.4.1 Entry-based enrichment analysis

The entry-based enrichment analysis utilizes the cumulative hypergeometric distribution [25, 26] based on the probability mass function

$$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad (5.5)$$

which represents the probability of observing exactly k successes when drawing n times from a population of size N containing M successes (without replacement). In the context of SEA, the population size N is calculated as the number of different entries in the selected database(s). M is the number of matched entries (successes). For each pathway, n is set to the number of associated entries and k to the number of matched entries in this pathway. The p-value for each pathway is calculated based on the cumulative probability

$$P(X \geq k) = \sum_{i=k}^n P(X = i). \quad (5.6)$$

Instead of calculating the numbers k , n , M , and N by counting all (matching) entries in the database/pathways, the user can select a particular type of entries (cpd/: compounds, ko/: KEGG ORTHOLOGY entries, ec/: enzymes, gene/: genes, all: all entries in the selected databases). This function is useful when a data set contains only marker candidates which are matched to a specific type of entries (e.g. compounds in case of MS-based Metabolomics experiments). In case mass-based mapping of marker candidates was performed (see section 5.2.2), entries with very similar masses are merged in order to reduce this systematic dependence. For this purpose, the entry masses are clustered (hierarchical clustering with complete linkage) before the statistical analysis is performed. The clusters are obtained by cutting the dendrogram at a maximum distance corresponding to the doubled mass tolerance (as specified, see section 5.2.2).

The pathway-specific p-values may be adjusted for multiple testing based on the Bonferroni (option `Direct p-value calculation, FWER control (Bonferroni)`) or Holm-Bonferroni method (`Direct p-value calculation, FWER control (Holm-Bonferroni)`) [14], which control the familywise error rate (FWER), or the Benjamini-Hochberg method (`Direct p-value calculation, FDR control (Benjamini-Hochberg)`) [15], which controls the false discovery rate (FDR). Additionally, the error rates can be estimated based on the random permutation of entry hits (options `Random permutation of ranks, FWER control` and `Random permutation of ranks, FDR control`). In this case, the p-values are directly calculated (see above) and then compared to the p-values recalculated for a high number of random permutations. In each permutation step, the entries which are matched by marker candidates and the entries which could not be matched are permuted randomly (with the number of matched entries constant). This approach takes the dependence of different pathways (e.g. pathways that contain a common subset of entries) into account when

calculating the error rates.

After calculation, the pathways in the pathway/set list box are sorted according to the corresponding p-values. An additional column contains the adjusted p-values/error rates. In case of direct p-value calculation and adjustment, the corresponding values may be greater than one (e.g. due to Bonferroni adjustment).

5.4.2 Marker/feature-based enrichment analysis

The marker/feature-based enrichment analysis provides four different types of tests: A test based on the cumulative hypergeometric distribution (option `Hypergeometric`, see previous section), an iterative hypergeometric test (`Hypergeometric (iterative)`), a rank-sum (Mann–Whitney U) test (`Rank-sum`), and a Kolmogorov–Smirnov test (`Kolmogorov-Smirnov`).

In the first case, the p-values are calculated based on the cumulative probability from equation 5.6. In contrast to the entry-based analysis, N is the number of marker candidates in the unfiltered data set, M the number of selected candidates (in MarVis-Filter or MarVis-Cluster), n the number of candidates in the unfiltered data set that match entries in a particular pathway, and k the number of selected and matching candidates (in a particular pathway). In case of the mass-based mapping of marker candidates (see section 5.2.2), the overall number of hits in a pathway (n) is calculated by applying all available adduct rules and isotope numbers used in the mass correction procedure (see section 3.5) to all marker candidates in the unfiltered data set and mapping the resulting masses to the database entries. In case mass correction factors were specified (see section 5.2.2), the factors are applied to all candidates and the resulting masses are mapped to the selected database(s).

The iterative hypergeometric method evaluates the different filtering possibilities in MarVis-Filter within one test. Starting with the first selected candidate in the MarVis-Filter ranking, followed by the first two, three, and up to all candidates which were selected (e.g. in MarVis-Filter or MarVis-Cluster), the cumulative probabilities are calculated. For each simulated filtering, the numbers M and k in equation 5.5 include only the selected candidates below/above the currently simulated filter threshold. For the final pathway-specific p-value, the minimum is taken over all simulated selections and multiplied by the number of matching selected candidates (applying all adduct/isotope rules in case of mass-based

mapping). If the data set was ranked and filtered several times, MarVis-Pathway utilizes the first ranking after data import in MarVis-Filter. This type of test is useful when applying the hypergeometric test to unfiltered data sets or when there is no suitable threshold for filtering.

The rank-sum (Mann–Whitney U) and Kolmogorov-Smirnov tests are more powerful but less robust alternatives to the iterative hypergeometric test. In case of the rank-sum test, the p-value for each pathway is calculated based on the ranks of all matching marker candidates in the unfiltered data set (considering all adduct rules, isotope numbers, and correction factors in case of mass-based mapping) in comparison to the ranks of all candidates which do not match entries in the current pathway. In case of the Kolmogorov-Smirnov test, the p-value is calculated by comparing the ranks of all matching candidates to the distribution of ranks in the unfiltered data set. Note that the resulting p-values reflect the rank distribution of all matching marker candidates (not only the selected candidates which are displayed in MarVis-Pathway).

For all four types of tests, the p-values may be calculated and adjusted directly or based on random permutations of marker candidate ranks/hits (see previous section). After calculation, the pathways in the pathway/set list box are sorted according to p-values and an additional column contains the adjusted values.

5.4.3 Sample-based enrichment analysis

The sample-based enrichment analysis uses the signal-to-noise ranking of marker candidates (see section 3.3.2) in combination with the rank-sum or Kolmogorov-Smirnov test (see previous section) and random permutations of sample condition labels (assignments of samples to experimental conditions). In contrast to the first two approaches, this method does not assume independent marker candidates or independent database entries for the final error rate estimation. It should be applied to unfiltered data sets only. The user has to specify the number of random permutations and dependence labels of samples (e.g. 1 2 3 for three independent samples, see section 3.3.2).

In the first step of this procedure, the available marker profiles (as selected in MarVis-Filter or -Cluster) are ranked using the signal-to-noise ratio. For each pathway, a p-score is calculated based on the rank-sum or Kolmogorov-Smirnov test (user's choice) of associated marker candidate ranks. Then, for a high number of random permutations of sample labels, the ranking of the permuted intensity profiles and the corresponding pathway-specific

p-scores are recalculated. The observed pathway-specific p-scores are then compared to the p-scores calculated for all pathways and random permutations. Finally, the error rate is estimated for each pathway (option `FWER control` or `FDR control`, see also section 3.3.2). After calculation, the pathways in the pathway/set list box are sorted according to the error rates and an additional column contains the adjusted values.

Note that the re-ranking of marker candidate profiles for all random permutations takes considerably more computing time compared to the marker/feature or entry-based SEA.

5.4.4 Meta-analysis of multiple data sets

MarVis-Pathway provides a framework for the joint (entry, marker/feature, or sample-based) Set Enrichment Analysis of combined data sets, e.g. resulting from positive and negative ionization mode in MS analysis, from different omics platforms, or replicate experiments (see section 3.4.1). For this purpose, the pathway-specific p-values/scores, which are calculated for each data set and pathway separately (see previous sections), are merged in a meta-analysis [7, 27] using Fisher's [28] or Stouffer's method [29] for independent data sets.

Additionally, groups of dependent data sets, e.g. obtained from the same biological samples, may be specified. For each pathway and group of data sets, the corresponding p-values are then combined by taking the minimum value multiplied by the number of group members (`Minimum p-value and Fisher's method`). The adjusted group-specific p-values are then used to calculate the meta-p-value for each pathway based on Fisher's method. In case no groups of dependent data sets were specified, all data sets are assumed to be independent. An exception applies to the sample-based analysis (see following paragraph).

In case a sample-based enrichment analysis is performed, representations of the same biological sample in different data sets may be linked and the condition labels are permuted together, e.g. a particular sample is always assigned the same condition label in all linked data sets. The linking option may also be combined with technical replicates belonging to independent biological samples (see section 3.3.2). For example: In case of linking two data sets with dependent sample labels 1 1 2 2 3 3 4 4 (four independent biological samples and two dependent technical replicates, respectively) and 1 2 3 4 (the same four independent biological samples, no technical replicates), the four biological samples

are always assigned the same condition label for both data sets and the technical replicates are assigned the label of the corresponding sample.

Finally, the error rates (FWER or FDR) are calculated based on the meta-p-values. In case a random permutation test is performed, the observed meta-p-value for a pathway is compared to the meta-values obtained for all pathways and all random permutations.

If no marker candidates in a data set match entries in a particular pathway, this data set is left out in the meta-analysis of the corresponding pathway. The user may select a subset of data sets which should be used for SEA and meta-analysis. The error rates of pathways which contain no hits for the selected data sets are set to infinity.

5.5 Data export

Results from functional annotation and statistical analysis in MarVis-Pathway may be exported via the `Export sets`, `Export annotated markers`, or `Export markers in set(s)` entries in the `File` menu. In the first case, the annotated set/pathways (see figure 5.3 region 1) are exported as CSV file. In the second case, all marker candidates are exported in MarVis CSV format (see section 3.1) with an additional entry column. This column contains the IDs and names of all matched database entries and the corresponding set/pathway IDs and names. In the third case, only marker candidates which could be mapped to selected pathways are exported. The pathways have to be selected by specifying their current rank indices (e.g. `1 : 3 6` for indices 1 to 3 and 6, see figure 5.3 region 1).

5.6 Data exchange with MarVis-Cluster

The annotated marker candidates can be analyzed in MarVis-Cluster (see chapter 4) using the `Goto MarVis-Cluster` entry in the `MarVis-Suite` menu. In this case, the candidates are annotated with an additional entry column. This column contains the IDs and names of all matched database entries and the corresponding pathway IDs/names. A method for aggregation of replicate intensities and for scaling of aggregated intensity profiles has to be specified (see section 3.6). The aggregated and scaled condition-specific profiles are then used for clustering and visualization in MarVis-Cluster (see section 4.3). After clustering, the marker candidates within a cluster may be sorted according to the projection onto the 1D-SOM (and thereby sorted according to similarity of their intensity profiles).

General functions in the MarVis-Suite

6.1 Toolbar

The toolbar in the MarVis-Suite main windows (below the menu bar) contains two toggle buttons for zooming. After activating the `Zoom in` button (on the left-hand side) with a mouse click, the MATLAB[®] zoom mode can be used to zoom into a particular plot by clicking into the plot (for details see the online MATLAB[®] zoom help page). Within a plot, a rectangular area which should be zoomed in can be specified by holding the left mouse button pressed and moving the cursor along the plot. After releasing the mouse button, the selected area will be zoomed in. By activating the `Zoom out` button (right-hand side), the plots can be zoomed out, respectively. A click with the right mouse button within a plot shows a small context menu. Via the `Reset to Original View` menu item, the original view can be restored. The zoom mode can be deactivated by pressing the respective button with another mouse click.

For plots in additional windows, MarVis utilizes the MATLAB[®] toolbar. Via the floppy disk symbol, the figure content can be saved in various image formats. By means of the symbols representing a magnifying glass or hand, the zoom mode can be activated/deactivated and the user can slide the plot (click on the hand symbol, click into the plot and hold the mouse button, move the mouse for sliding). By clicking on the datacursor symbol (third from right-hand side), the datacursor mode (see MATLAB[®] help page) can be activated and used for inspection of associated data. A colorbar representing the current color mapping can be added by clicking on the respective symbol (second from right-hand side). For changing the colormap, see section 4.4.3 about the MATLAB[®] colormap editor.

6.2 Save and load projects

Via the entries `Save project` and `Load project` in the `File` menu of all MarVis-Suite main windows, the current data set and all user-specific settings (e.g. visualization properties, dialog entries, etc.), can be saved or restored. MarVis displays a file dialog where the input/output file has to be selected. The project is stored in the MATLAB[®] `mat`-format. Note that projects saved in MarVis-Filter, MarVis-Cluster, or MarVis-Pathway can only be loaded in the respective tool. By default, MarVis saves all user-specific settings at the end of each session in the files `MFilterUserSettings.mat`, `MClusterUserSettings.mat`, and `MPathwayUserSettings.mat` in the MarVis-Suite program directory. At the start of each session, MarVis automatically loads these files and the corresponding settings.

Via the menu entry `Reset default settings` in the `File` menu, the default settings, e.g. the default colormap, can be restored.

6.3 Search for marker candidates

Marker candidates with particular IDs, x , or y -values (e.g. retention time or mass) can be searched for and selected/deselected using the `Find markers` dialog via the `Selection` or `Analysis` menu in all MarVis-Suite main windows (see figure 6.1). The (range of) values to be searched for can be specified in the `ID/x/y Value` text fields (left column). If a value should not be limited, the corresponding field has to be left empty. The ID must be specified as a regular expression¹ or substring (e.g. “`^123`” for marker IDs starting with “123”), while the x and y -value must be numerical. The ID search is case-insensitive and leading/trailing whitespaces are removed. In the `Deviation` fields the maximum absolute x and y -deviation of the search results can be specified. Note that these values must be positive. After clicking the `Find` button, MarVis searches for marker candidates

- whose IDs match the regular expression or substring (if specified)
- AND whose x -values are in the range [given x -value - x -deviation, given x -value + x -deviation] (if specified)

¹For information about regular expressions, we refer to the corresponding MATLAB[®] help page or the wikipedia web site.

- AND whose y-values are in the range [given y-value - y-deviation, given y-value + y-deviation] (if specified).

The search results are presented in the Marker information box at the bottom of the dialog (see also section 4.4.1). Here, the selection status (1: marker candidate selected, 0: marker not selected), the associated cluster number, the ID, x and y-values, and additional values of matching candidates can be inspected. By clicking on a single row or by using the up and down cursor keys, a particular marker candidate can be activated. Clicking Select/Deselect or pressing m on the keyboard adds the current candidate to the selection list (only in MarVis-Cluster). By clicking the Close button, the Find markers dialog is closed. After closing the dialog, MarVis jumps to the activated marker candidate in the corresponding main window.

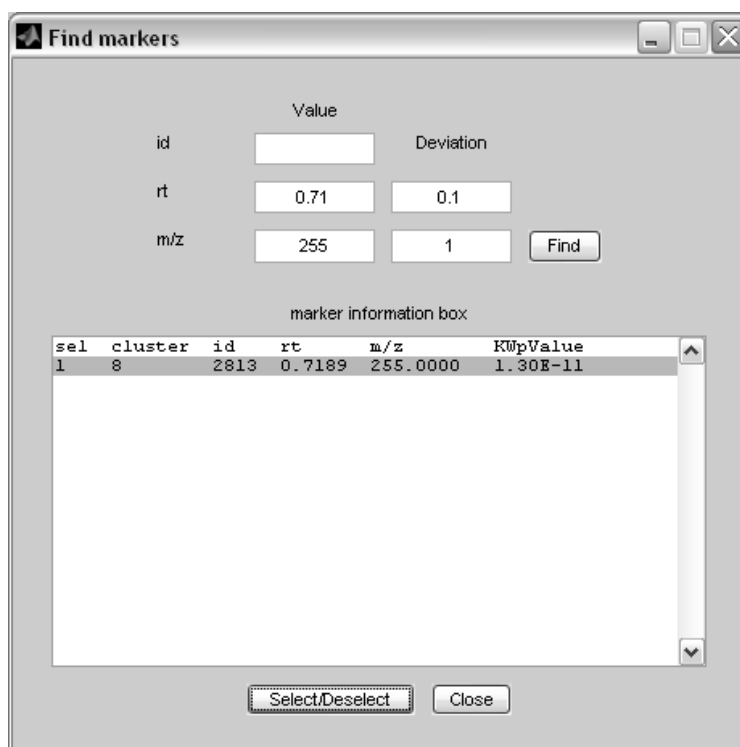


Figure 6.1: Find markers dialog for example dataset1.csv

6.4 General visualization functions

6.4.1 Standard deviation, box, and scatter plots

The raw intensity profile of the current marker candidate can be visualized as mean/standard deviation plot or boxplot using the `Errorplot` and `Boxplot` entries in the `MarVis-Suite` menu. In the first case, `MarVis` plots the mean intensities for all conditions and the sample standard deviation within all conditions in a new window. In the second case, a condition-specific boxplot is displayed (see `MATLAB`[®] `boxplot` function for details).

Additionally, in `MarVis-Filter` and `MarVis-Pathway`, the `x` and `y`-values of marker candidates (e.g. retention times and masses) can be visualized as scatter plot (entry `Plot x/y values`). The `x` and `y`-values of all marker candidates in the unfiltered data set are plotted as small gray dots. The values of all marker candidates in the current filtered data set are marked as big red dots, the values of currently selected candidates (`MarVis-Filter`) or of candidates in the currently highlighted pathway/set (`MarVis-Pathway`) as big black dots.

6.4.2 Export of graphics

Via the `Export graphics` item in the `MarVis-Suite` menu, all plots of the current main window can be copied into separate windows, inspected, and exported. By using the `MATLAB`[®] toolbar (beneath the menu bar, see section 6.1), the plots can be adjusted (see `pan` function), zoomed in (see `zoom` function), inspected with a data cursor (see `datacursor` function), annotated with a colormap bar (see `colorbar` function), and exported in various image formats (e.g. in TIF or EPS format, see `saveas` function). For general export options, such as the resolution in dpi, the `MATLAB`[®] `Export Setup` function in the `File` menu may be used (see `Export Setup` function).

Via the `Copy figure` entry in the `Edit` menu, the whole plot can be copied into the system clipboard. Using the `Colormap editor` entry (see section 4.4.3 for details), the colormap can be modified. Via the `Figure properties` entry, general properties of the current plot can be adjusted for export (e.g. the title, the `x` and `y`-labels, and the axes ticks and labels). All described functions can also be applied to standard deviation and boxplots generated by `MarVis` (see section 6.4.1).

Via the `Hide cursors` and `Show cursors` entries in the `MarVis-Suite` menu of each main window, the cursors within each plot can be hidden or restored. These functions can be useful when exporting graphics without highlighting.

Note that for the standard export of plots (not in case of the `Export Setup` function), the corresponding window size determines the resolution of the resulting images.

6.5 MarVis-Suite log function

All MarVis-Suite tools log the workflow of data analysis and the applied functions. The user can inspect the log via the `View log` function in the `MarVis-Suite` menu. The log text is shown in a new window and can be exported via the `Save file` button or copied into the system clipboard using the `Control+C` keys. Additionally, the user can edit the log and confirm the changes by clicking the `OK` button. When closing the window or clicking the `Cancel` button, changes are not applied to the log.

6.6 Molecular formula calculation

In MS data analysis, putative molecular formulas can be calculated for the selected (corrected) marker candidate using the `Mass2Formula` entry in the `MarVis-Suite` menu. In the following dialog (see figure 6.2), the chemical elements used for calculation (`elements`), the input mass (`measured mass`), and the mass tolerance (`tolerance`) can be specified. By default, `measured mass` is set to the (corrected) `y`-value of the current marker candidate (see section 3.5 for details about mass correction).

When clicking on the `Std` button, MarVis fills the `elements` field with a set of default elements (`C H N O P S`). If the checkbox `chemical rules` is activated, MarVis filters out formulas which do not satisfy the heuristic Seven Golden Rules [30]. If the checkbox `RDBE` is activated, MarVis also performs a Rings-plus-Double-Bonds-Equivalent check for the obtained formulas (see [30]). Via the checkbox `number of carbons estimated` and the text fields `number of carbon atoms` and `tolerance`, the obtained formulas can be filtered further. The (estimated) number of carbon atoms may be a floating-point number, the tolerance must be a number between 0 and 1 (e.g. 0.25 for a 25 percent tolerance). MarVis automatically fills the first text field with the estimated number of carbon atoms from isotope correction (if the estimation was possible, see section 3.5).

The calculation can be started via the `Calculate` button. Note that the calculation of molecular formulas for large masses (> 500 Da) without applying chemical rules may take some time. After calculation and filtering, the remaining formulas and their deviation from the measured mass are displayed in the text field below the button, sorted according to absolute deviation. Single lines of the output can be copied into the system clipboard using the `Control+C` keys.

The molecular formula calculator can also be applied to the difference of y -values of two selected marker candidates (function `DifferenceMass2Formula`, only in MarVis-Cluster). This option is useful for the identification of adducts in MS data analysis.

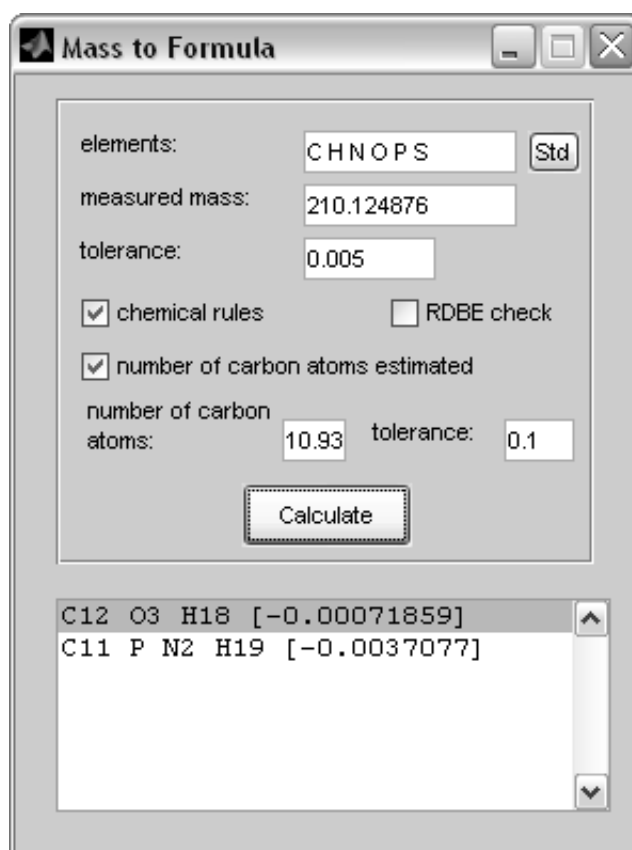


Figure 6.2: Mass-to-formula dialog

References

- [1] P. Meinicke, T. Lingner, A. Kaefer, K. Feussner, C. Göbel, I. Feussner, P. Karlovsky, and B. Morgenstern. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology*, 3:9, 2008.
- [2] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.
- [3] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.
- [4] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1):D742–D753, 2012.
- [5] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.
- [6] J. Xia and D. S. Wishart. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(suppl 2):W71–W77, 2010.
- [7] A. Kaefer, M. Landesfeind, K. Feussner, B. Morgenstern, I. Feussner, and P. Meinicke. Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets. *PLoS ONE*, 9(2):e89297, 2014.

- [8] A. Kaefer, T. Lingner, K. Feussner, C. Göbel, I. Feussner, and P. Meinicke. MarVis: a Tool for Clustering and Visualization of Metabolic Biomarkers. *BMC Bioinformatics*, 10:92, 2009.
- [9] A. Kaefer, M. Landesfeind, M. Possienke, K. Feussner, I. Feussner, and P. Meinicke. MarVis-Filter: Ranking, Filtering, Adduct and Isotope Correction of Mass Spectrometry Data. *Journal of Biomedicine and Biotechnology*, 2012, 2012. doi: 10.1155/2012/263910.
- [10] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.
- [11] J. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. CRC Press, 2003.
- [12] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, 2(4):775–795, 2012.
- [13] S. Wright. Adjusted p-values for simultaneous inference. *Biometrics*, 48(4):1005–1013, 1992.
- [14] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [15] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- [16] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121, 2001.
- [17] Z. He and J. Zhou. Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. *Applied and Environmental Microbiology*, 74(10):2957–2966, 2008.
- [18] L. V. Hedges. Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2):107–128, 1981.
- [19] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):3, 2004.

- [20] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210, 2003.
- [21] V.-T. Tran, S. A. Braus-Stromeier, H. Kusch, M. Reusche, A. Kaefer, A. Kühn, O. Valerius, M. Landesfeind, K. Aßhauer, M. Tech, K. Hoff, T. Pena-Centeno, M. Stanke, V. Lipka, and G. H. Braus. Verticillium transcription activator of adhesion vta2 suppresses microsclerotia formation and is required for systemic infection of plant roots. *New Phytologist*, 202(2):565–581, 2014.
- [22] R. Tautenhahn, C. Böttcher, and S. Neumann. Annotation of LC/ESI-MS mass signals. *Bioinformatics Research and Development*, 4414:371–380, 2007.
- [23] R. J. Cho, M. J. Campbell, E. A. Winzler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, 2(1):65–73, 1998.
- [24] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.
- [25] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.
- [26] D. A. Hosack, G. Dennis Jr, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):R70, 2003.
- [27] K. Shen and G. C. Tseng. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323, 2010.
- [28] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 1925.
- [29] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. *The American soldier: adjustment during army life*. Princeton University Press, Princeton, 1949.
- [30] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8(1):105, 2007.

Curriculum Vitae

Name	Alexander Kaefer
Date of birth	February 21 st , 1985 (Hannover)
Nationality	German
Email	alex@gobics.de

Education

Since 2011	Göttingen Graduate School for Neurosciences, Biophysics, and Molecular Biosciences (GGNB), thesis project "Development of a statistical framework for mass spectrometry data analysis in untargeted Metabolomics studies"
2008-2010	Georg-August-University Göttingen, Master Applied Computer Science with focus on Bioinformatics, thesis "Entwicklung und Evaluation von Methoden zur Erkennung von Metabolit-Markern in Massenspektrometrie-Daten"
2005-2008	Georg-August-University Göttingen, Bachelor Applied Computer Science with focus on Bioinformatics
1997-2004	Secondary school, Gymnasium Burgdorf (Hannover)

Employment history

Since 2010	Georg-August-University Göttingen (Department of Bioinformatics), scientific assistant in the BMBF project "The plant-pathogenic fungus <i>Verticillium longisporum</i> and the interaction with its host <i>Brassica napus</i> "
2007-2010	Georg-August-University Göttingen (Department of Bioinformatics), student programmer: Software and method development for data mining, teaching and supervision of computer tutorials
2004-2005	Georg-August-University Göttingen (University Medical Center), civilian service

Conferences

02/2014	BioFung Seminar: Recent Advances in <i>Verticillium</i> Research, Göttingen, Germany (organization team)
09/2013	GCB (German Conference on Bioinformatics), Göttingen, Germany (organization team)
07/2012	ISMB (Intelligent Systems for Molecular Biology) and student council symposium, Long Beach, California, USA