Heliyon 7 (2021) e06048

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Comparative analysis of statistical tools for oil palm phytochemical research

Nur Ain Ishak [a,*], Noor Idayu Tahir [a], Syafi'ah Nadiah Mohd Sa'id [b], Kathiresan Gopal [c], Abrizah Othman [a], Umi Salamah Ramli [a]

[a] Advanced Biotechnology and Breeding Centre (ABBC), Malaysian Palm Oil Board (MPOB), No. 6, Persiaran Institusi, Bandar Baru Bangi, 43000 Kajang, Selangor, Malaysia
[b] Faculty of Chemical Engineering, Universiti Teknologi MARA (UiTM) 40450 Shah Alam, Selangor, Malaysia
[c] Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang Selangor, Malaysia

ARTICLE INFO

Keywords:
Phytochemical analysis
Metabolites
Oil palm
Multivariate data analysis
Statistical tools

ABSTRACT

Recent advances in phytochemical analysis have allowed the accumulation of data for crop researchers due to its capacity to footprint and distinguish metabolites that are present within an organisms, tissues or cells. Apart from genotypic traits, slight changes either by biotic or abiotic stimuli will have significant impact on the metabolite abundances and will eventually be observed through physicochemical characteristics. Apposite data mining to interpret the mounds of phytochemical information from such a dynamic system is thus incumbent. In this investigation, several statistical software platforms ranging from exploratory and confirmatory technique of multivariate data analysis from four different statistical tools of COVAIN, SIMCA-P+, MetaboAnalyst and RIKEN Excel Macro were appraised using an oil palm phytochemical data set. As different software tool encompasses its own advantages and limitations, the insights gained from this assessment were documented to enlighten several aspects of functions and suitability for the adaptation of the tools into the oil palm phytochemistry pipeline. This comparative analysis will certainly provide scientists with salient notes on data assessment and data mining that will later allow the depiction of the overall oil palm status in-situ and ex-situ.

1. Introduction

Phytochemicals are low molecular weight small molecules found in plant cells and tissues and are easily absorbed or damaged by spontaneous reactions, enzymatic reactions or conjoined with other molecules (Zhou et al., 2012). Extensive phytochemical analysis in relation to functional genomics and systems biology within an organism, tissue or cell under a given set of conditions at specific time is dubbed as metabolomics (Sindelar and Patti, 2020). The size and complexity of plant metabolites collections, i.e., metabolome vary by species and samples (Beisken et al., 2015). A metabolome can be examined by two distinct approaches; nontargeted and targeted analysis of endogenous and exogenous metabolites (Brown et al., 2009). The nontargeted methodologies is the most suitable technique to detect unexpected changes in phytochemical concentrations. It involves profiling of metabolites with maximum metabolome coverage over a wide range of complex phytochemical structures to provide more opportunity to identify changes without bias and *a priori* knowledge regarding the examined specimens. In the targeted approach, a relatively small number of metabolites of interest is to be predefined prior to running the experiment for detection and quantification.

Metabolome analysis is particularly employed to monitor, measure and assess the myriad diversity of various physical properties of metabolites and demands high end analytical instrumentations and software solutions, i.e., gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), capillary electrophoresis–mass spectrometry (CE-MS) and nuclear magnetic resonance (NMR) spectroscopy (Johnson and Gonzalez, 2012). Analysis of intricate biological samples are being routinely carried out using LC-MS compared to other metabolomics approaches due to a combination of sensitivity and the amount of information generated by the analysis for instance retention time (t_R) information, mass-to-charge ratios (m/z), signal intensities and ion abundance, which can be further used as an additional information for indexing metabolites (Dunn et al., 2013). The resulting metabolome profiles provide functional signatures that can be analysed using chemometrics which involves simultaneous measurement of two or more variables in an experiment to capture relationship among the variables caused by changes of metabolite abundance (Saccenti et al., 2014). Post-hoc analysis of high-throughput

* Corresponding author.
E-mail address: nurain@mpob.gov.my (N.A. Ishak).

https://doi.org/10.1016/j.heliyon.2021.e06048
Received 1 June 2020; Received in revised form 28 September 2020; Accepted 18 January 2021

Table 1. Summary of unsupervised and supervised method strategies.

Type of methods	Examples of analysis	Goal	Application	Input	Output
Unsupervised method	Principal Component Analysis (PCA)	Reduction of data dimensionality, visual inspections of data grouping and pattern recognition	Dimensional reduction recognition into an observed variation data and extraction of components that explain maximum variance(s)	Data tables without class associations: each row represents a subject and each column represents concentration/abundance of a metabolite (e.g., MS and NMR peak list or spectral bin)	Data summary in scores and loadings plots for pattern recognition
	Hierarchical Clustering		Display of subjects' connectivity in cluster formation		Grouping of data into dendrogram and heat map
Supervised method	Partial Least Squares (PLS)	Biomarker discovery and class membership prediction	Assessment of variables contributing to discrimination of subjects	Data tables with prior class membership dictation	Selection of dependent variable (metabolites) to represent class membership
	Support Vector Machine (SVM)		Construction of model that can assign new subject(s) to one category or the other		Selection of metabolites as predictors to construct prediction model

multi-dimensional data generated from metabolomics platforms may require the involvement of multidisciplinary skilled personnel, e.g., analyst, bioinformatician or scientist who are proficient in chemometric analysis. A number of available tools for metabolomics data mining are dedicated for pattern or attributes recognition and dimension reduction based on two strategies of exploration and confirmation; the unsupervised and supervised methods respectively.

Unsupervised method generates data clusters without any prior knowledge about the group structure (Ren et al., 2015). This method is commonly used in preliminary evaluation of the information contained in the data sets. Examples of the analysis of unsupervised method are the principal component analysis (PCA) and clustering. On the other hand, supervised method such as partial least squares data analysis (PLS-DA) and support vector machine (SVM) are widely used in discovering biomarkers, classification, and prediction. This method is used to confirm the results obtained from the unsupervised method. Unsupervised and supervised methods are not completely independent and each method serves different research goals (Ren et al., 2015). The different functions and features of the methods do not imply superiority of supervised over unsupervised method and vice versa. Each method has a different purpose and is used according to fittingness of the exploration. The summary for the methods mentioned above is described in Table 1.

The human population is estimated to reach 9 billion in 2050 which will manifest into the shift and rise of food demand (Béné et al., 2019). Oil palm has become one of the most versatile and important crops globally due to its status as high oil-yielding source of vegetable oil which can then be widely utilised for both food and non-food products (oleochemical industries and biofuel) (Kushairi et al., 2018). Producing the highest yield of oil per unit area compared to any other crops (Oettli et al., 2018), its harvests include crude palm oil, crude palm kernel oil, palm kernel and palm kernel cake as commodities, and biomass products as energy and non-energy feedstock (Sukiran et al., 2018). However, the oil palm industry faces challenges of sustainability and struggles to remain competitive in catering the demand for the vegetable oil which is steadily increasing. With the aim of minimum acreage with optimum yield, the research and development (R&D) towards achieving the full oil palm genetic yield potential competes with gaps caused by agro-climatic factors (Barcelos et al., 2015). The focus of this survey will be on the analysis output of oil palm metabolome data sourced from one of our field trials using several statistical platforms from both unsupervised and supervised methods. Oil palm systems biology study is still at infancy and the depth and extent of these statistical analysis has allowed us to exhaustively examine and decipher the raw data into meaningful interpretation in effort to uncover and understand the complexities and diverse physiological processes of the species. We will also discuss the advantages and limitations of these methods to assess their suitability to be adapted into our phytochemical analysis pipeline of oil palm field data.

2. Materials and methods

2.1. Sample preparation, liquid chromatography-mass spectrometry (LC-MS) analysis, data collection and pre-processing

All experiments were performed in accordance with the protocols established for rapid and wide range of metabolite extraction for oil palm leaves that has permitted the identification of phytochemicals from a single extraction (Tahir et al., 2012, 2013). The general plant metabolomics workflow of oil palm tissue sampling up to data analysis and storage steps is schematised in Figure 1. LC-MS data was obtained from the metabolomics profiling of spear leaves from clonal oil palm planted on two different planting sites of Keratong and Teluk Intan of different soils as previously described by Tahir et al. (2016).

In this particular analysis, LC-MS profile data from 1.0 to 59.5 min analysis time were pre-processed with Find Molecular Features (FMF) parameter in ProfileAnalysis™ (Version 2.0) software from Bruker Daltonics GmbH. A signal-to-noise (S/N) limit of 5 was set for the peak finder for a chromatographic peak to be eligible in peak detection. The correlation coefficient threshold was applied at 0.7 for the minimal time correlation between two related isotope traces in peak clusters to be

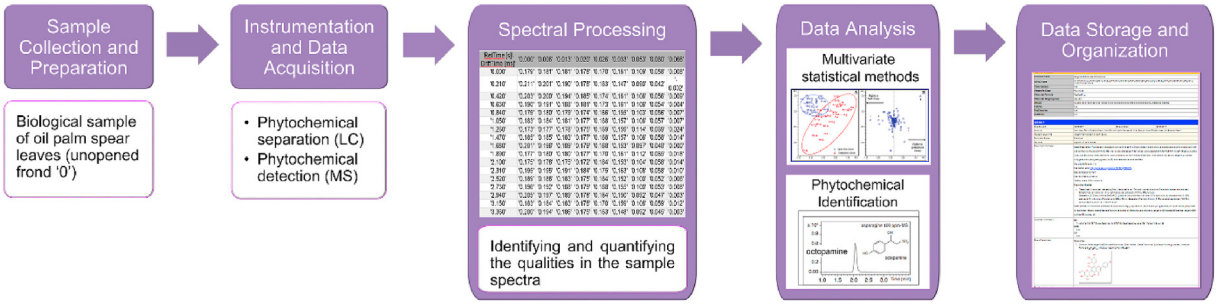

Figure 1. General metabolomics workflow for oil palm.

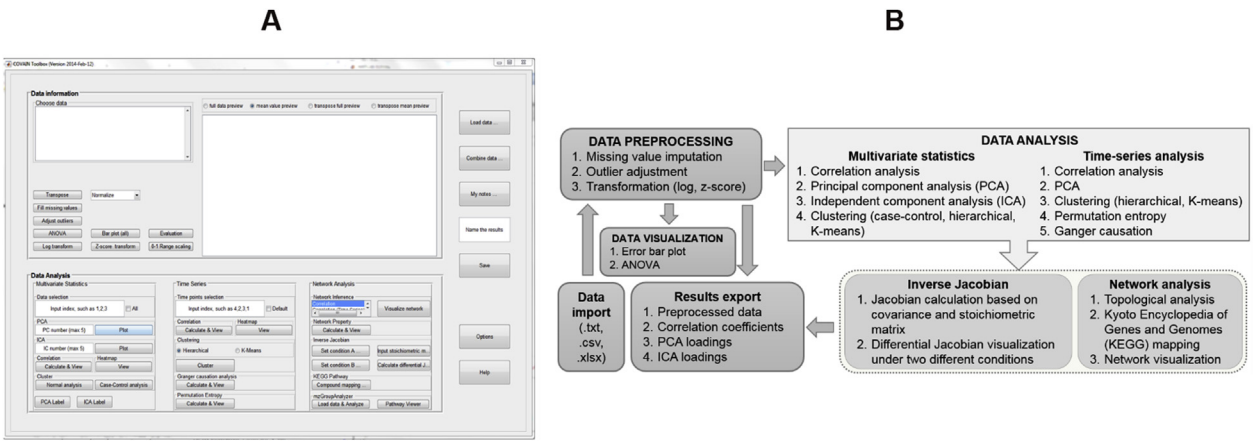

Figure 2. Interface (A) and workflow (B) of COVAIN toolbox.

combined to a charge state at a given retention time. The minimum compound length of a consecutive spectra in an MS signal was set at 10 spectra for a chromatographic peak. This value principally depends on the chromatographic peak width and the acquired spectra rate and can be determined by counting the number of data points at the full width at half maximum (FWHM) of a peak. A 'bucket' or data matrix was generated with LC-MS peaks components from *m/z* range of 50–1000. Each data file contains 'mass-to-charge (*m/z*) ratio: retention time' pairs with the corresponding peak intensities. The buckets of 60 s (s) and *m/z* 1.0 deltas were arranged into a tabular data set format with retention times (rows) against peak intensities (columns) and exported into appropriate data format for further use, e.g.,.csv or.xls. No further data treatment was applied onto the data set prior to data export.

2.2. Program implementation

To date, there are many software and web servers that offer easy analysis and interpretation tools for metabolome data. All statistical software tools in this investigation were operated with the default or recommended settings from a server equipped with 3.10 GHz Intel(R) Xeon(R) CPU E5-1607 and 16 GB of RAM. The operating system was Windows 7 Professional (64 bits). In this study, four tools were selected for data mining and the results for each utilized tools were compared.

Four software tools used in the study:

i. **Covariance Inverse 'COVAIN' toolbox version 2017-May-16** developed by Molecular Systems Biology Department, University of Vienna, Austria using Matrix Laboratory 'MATLAB©' software.
ii. **SIMCA-P+ version 12.0.1** licensed software from Umetrics (now Sartorius Stedim Biotech) and SIMCA-P+ 15.2 trial version.
iii. **MetaboAnalyst version 4.0**, a free online tool by Xia Lab at McGill University, Canada.
iv. **RIKEN Excel Macro tool**, a tool for statistical analysis on Microsoft Excel 2010 Suite, metabolomics platform by The Japanese Institute of Physical and Chemical Research (Rikagaku Kenkyūjyo, RIKEN)

2.2.1. COVAIN toolbox version 2017-May-16

COVAIN toolbox is a uni- and multivariate statistics, time series and network analysis core for metabolomics covariance data supported in MATLAB Statistics and Machine Learning Toolbox environment (Sun and Weckwerth, 2012, 2013). COVAIN toolbox can be downloaded free of charge from https://mosys.univie.ac.at/resources/software/, however the potential user(s) are required to procure MATLAB software annual license to run the toolbox. Figure 2 shows the COVAIN interface and the

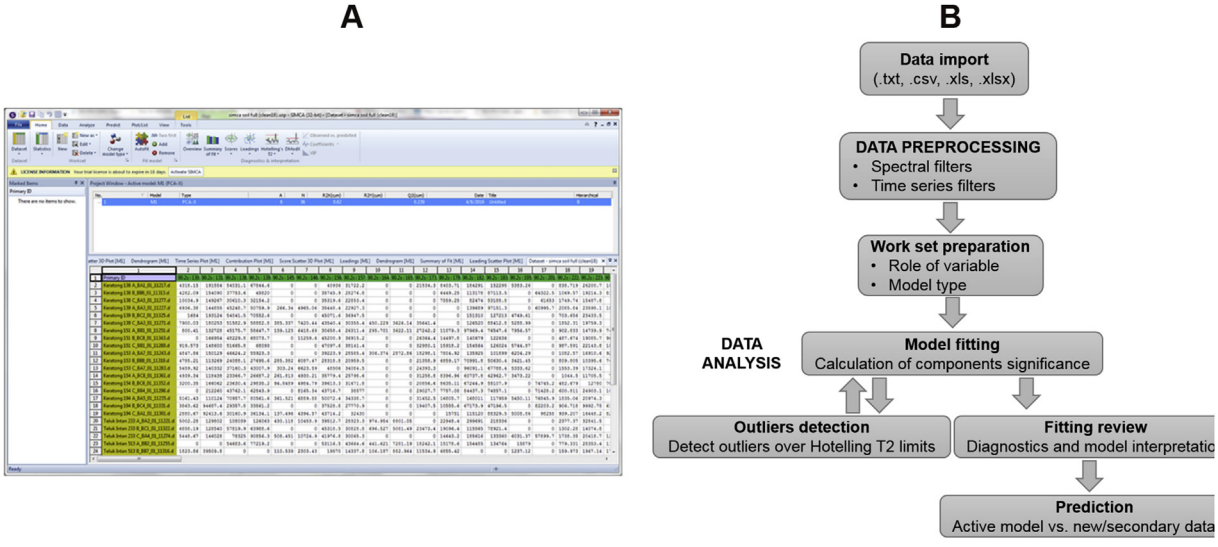

Figure 3. Interface (A) and workflow (B) of SIMCA-P+.

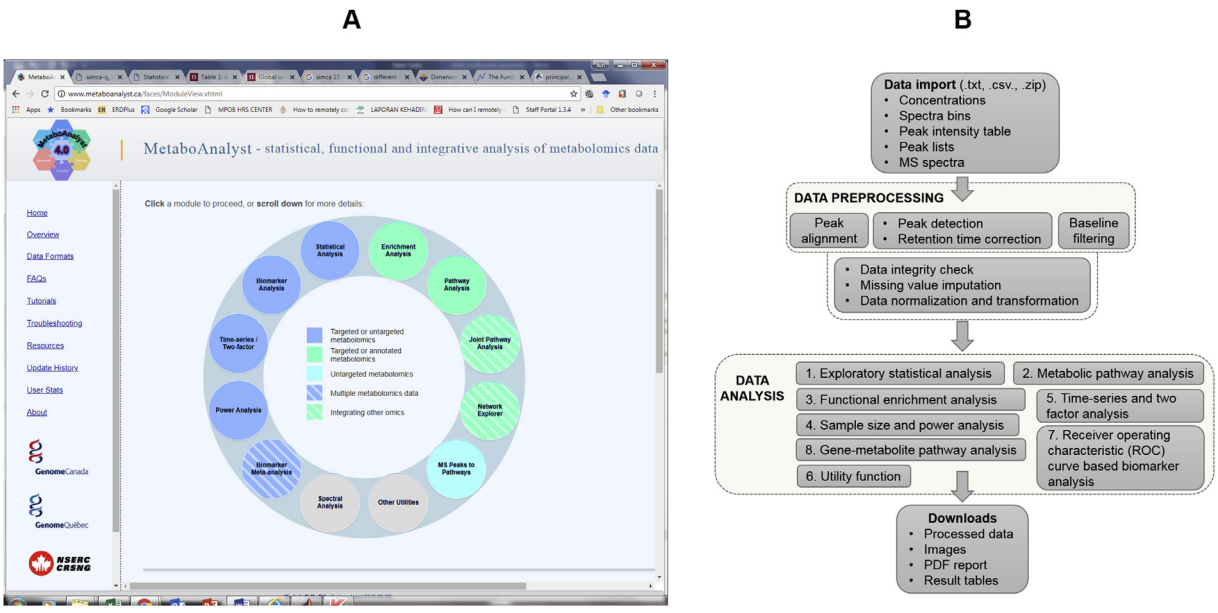

Figure 4. Interface (A) and workflow (B) of MetaboAnalyst.

toolbox workflow. The format of raw data that can be read by COVAIN are.txt (tab delimited text),.csv (comma-separated values) and.xls/.xlsx (Excel). The default missing value imputation is carried out by filling in half of the minimal value of all data.

2.2.2. *SIMCA-P+ version 12.0.1*

SIMCA-P+ software is a licensed software developed by Umetrics (now Sartorius Stedim Data Analytics AB). SIMCA-P+ is a suite of multivariate data analytics solution and is mainly used for the methods of PCA and PLS regression (PLS-R) (Wu et al., 2010). Figure 3 shows the interface of SIMCA-P+ and the data analysis workflow. SIMCA-P+ accepts raw data set in.txt,.csv and.xls or.xlsx file formats with many non-numerics or zeros. For the results to be reliable, e.g., to be used for model fitting, the data set should contain no more than 50% missing values in the observations of variables. However, the tolerance of missing value limit can be adjusted and the software will prompt a message to include or exclude the variable with zero-valued observation. A variable with zero value is given a scaling weight of 1 (Sartorius Stedim Data Analytics AB, 2017). There are several parameters that need to be determined before running the analysis including the identifiers for the variables, the roles of the variables, the data type and indication of data inclusion or exclusion. Many types of diagnostics and interpretation can be performed with SIMCA-P+ such as scores plot, loadings plot, Hotelling's T2 and Distance to the model (DModX) for a meaningful confirmatory data analysis. While using SIMCA-P+ version 12.01 with a perpetual license, the dataset was also tested on the trial version of 15.2

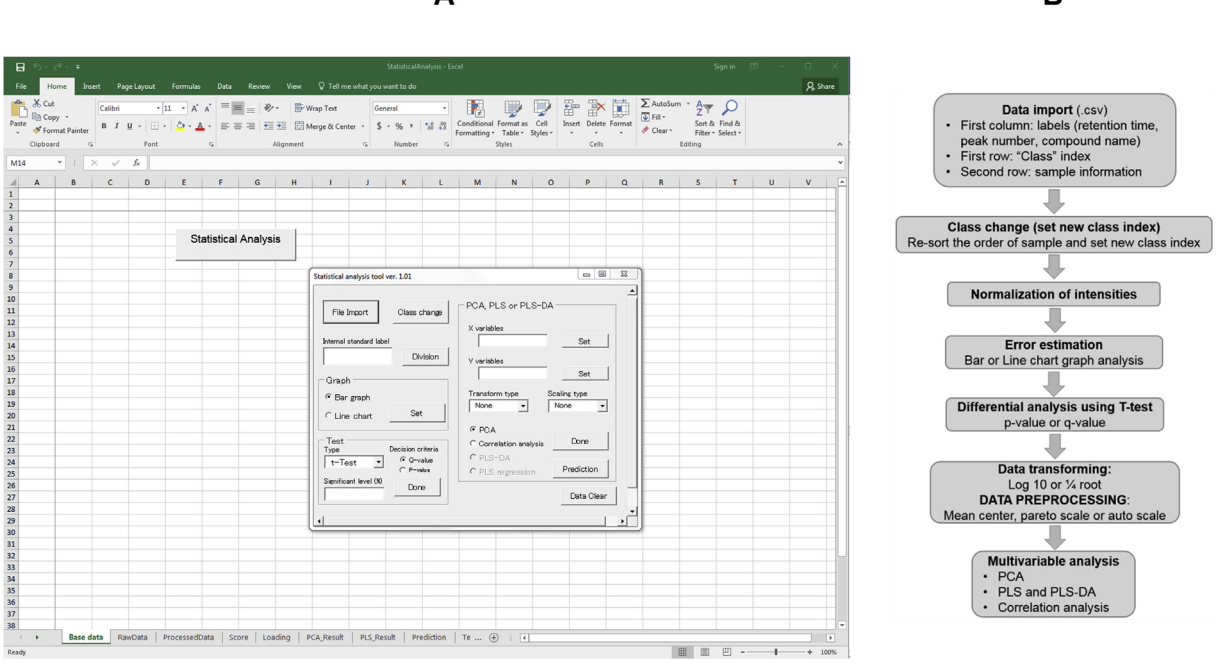

Figure 5. Interface (A) and workflow (B) of RIKEN Macro tool in Excel.

Table 2. Scaling methods.

Scaling type	Calculation	Details
Unit variance	$1/\sigma$	• Data analysis based on correlations instead of covariances • Inflation of measurement errors
Pareto	$1/\sqrt{\sigma}$	• Metabolites of large fold less dominant, balances data intensity • Data does not become dimensionless compared to unit variance scaling • Data remains nearer to original measurement
Variance	$1/\sigma^2$	• Emphasise relative responses • Inflation of measurement errors
Vast (variable stability)	$1/(\sigma/m)$	• Concentrate on metabolites with small variations • Unsuitable for data of large induced variation(s)

*m = mean, adapted from Van den Berg et al. (2006).

(free download with registration from https://landing.umetrics.com/simca-free-trial-offer) to look out for significant upgrade to the software.

2.2.3. MetaboAnalyst version 4.0

MetaboAnalyst is a publicly accessible, user-friendly, online metabolomics data analysis and interpretation tool. It is available at http://www.metaboanalyst.ca/. Since MetaboAnalyst is a web-based data analysis tool, internet connection is required. This program is regularly updated and improvement activities are usually announced via its portal or publication in refereed periodicals. As an alternative, MetaboAnalyst provides Web Application Resource (.war) file for local installation and operation at a computer or a server. However, this format requires the user to have basic computer skills to successfully employ the R programming language. MetaboAnalyst can run on Mac OS X or Linux (Redhat Fedora, Ubuntu, etc.) operating system with memory of 4 Gigabytes or more, Java software version 1.7 or more recent, R package version 3.4 or more recent with the following packages installed: "Rserve", "ellipse", "scatterplot3d", "pls", "caret", "lattice", "Cairo", "randomForest", "e1071″, "gplots", "som", "xtable", "RColorBrewer", "xcms", "impute", "pcaMethods", "siggenes", "globaltest", "GlobalAncova", "Rgraphviz", "KEGGgraph", "preprocessCore", "genefilter", "pheatmap", "igraph", "RJSONIO", "SSPA", "caTools", "ROCR", "pROC", "sva" packages and other application servers (Glassfish or Payara version 4.0 or above). To date, MetaboAnalyst has now an additional section of companion R interpreter, MetaboAnalystR for more flexible data analysis and batch processing.

MetaboAnalyst accepts data in.csv,.txt and.zip format, and also a tab-delimited text file format for MS-based proteomics and metabolomics called.mzTab. The software is able to process the data even though there are many zeros in the data set. The default missing value imputation method exchanges missing values with the half of the minimum positive values in the original data with the assumption of it to be the detection limit. Other missing values replacement methods are by column mean/median, k-nearest neighbour (KNN), Bayesian PCA (BPCA), probabilistic PCA (PPCA) and Singular Value Decomposition (SVD). Prior to data upload, the format and arrangement of the data must follow the data format guidelines for the data formats provided in the MetaboAnalyst main menu at its webpage. The statistical analysis provided by MetaboAnalyst is listed in its interface image in Figure 4. After uploading the raw data, just like COVAIN, this data analysis tool will pre-process the data by filling in missing values, filtering and normalisation. After data pre-processing, various analysis paths of uni- and multivariate analyses can be applied to explore the data.

2.2.4. RIKEN Excel Macro tool

RIKEN released a Microsoft Excel-based statistical analysis platform for statistical analyses such as principal component analysis (PCA) and projection to latent structure-based multivariate data analysis, e.g., partial least squares (PLS) regression (PLS-R), PLS discriminant analysis (PLS-DA) and correlation analysis. This tool requires Windows operating system (OS) and Microsoft Excel (32 bits and Microsoft Excel, 2007/2010 version or 64 bits and Microsoft Excel, 2010 version). Figure 5 shows the interface and the workflow of the tool. The imported data must be in the.csv file format. After the data is imported, the x variables need to be determined along with the transform type and scaling type. If PLS-R or PLS-DA is to be performed, a Y-variable must be selected. The analysis will proceed after clicking the 'Done' button. The tool can be downloaded for free from the following link; http://prime.psc.riken.jp/Metabolomics_Software/StatisticalAnalysisOnMicrosoftExcel/index.html. As the tool appears to aim for simplicity, comprehensive data processing function prior to multivariate analysis such as excluding metabolite peaks by setting a threshold at the 'peak count filter' parameter and comparison to quality control (QC) sample can be performed by missing-value interpolation in MS-DIAL software, available at the RIKEN website (Matsuo et al., 2017).

The four platforms have been chosen for this study based on their availability and usage history in Proteomics and Metabolomics (PROMET) Research Laboratory, Malaysian Palm Oil Board (MPOB). All of these platforms have been individually applied by different researchers based on their personal preferences in PROMET lab to analyse their metabolome data. In this study, the unsupervised method available from each platform, e.g., PCA and cluster analysis have been used onto the data set while for supervised methods, we focused on PLS-DA available from SIMCA-P+, MetaboAnalyst and RIKEN Excel Macro tool.

3. Results and discussion

Ideally, statistical methods for data mining require a homogenous and complete data set. However, in actuality several factors may contribute to missing values in a phytochemistry data corpus:

i. natural absence of the phytochemicals in specimen (biological replicates)
ii. error in specimen preparation (technical replicates)
iii. error from analytical platforms, e.g., injector
iv. analytical platform detection limits, e.g., ultraviolet (UV), mass spectrometry
v. error in data collection/acquisition or data export

In particular, raw data set of phytochemical analysis from chromatography paired with mass spectrometry contain many zeros as results of data acquisition cut-offs (Yang et al., 2015). These zeros or missing values are either removed in pairwise or list-wise manner (Szymańska, 2018). There are also instances of raw data from vendor data export software containing empty fields and have to be replaced with zero or other appropriate numeric imputations according to the types and randomness of the absence (Wei et al., 2018). For COVAIN tool, half of the smallest value of all data will be used to fill the missing values (Sun and Weckwerth, 2012). From our observation, heterogeneous data with too many non-numerics or zeros cannot be analysed by most of the tools

Table 3. Normalisation and scaling methods applied prior to data analysis.

Tools	Normalization	Scaling	Notes
COVAIN (Sun and Weckwerth, 2013; 2012)	Several options provided: -Normalisation by internal standard - Normalisation by sample fresh weight	N/A	Other pre-treatment process options: - data transformation (Log transformation and z-score transformation)
SIMCA-P+ (Wu et al., 2010)	- Standard Normal Variate – SNV	Pareto scaling	Other types of scaling selections: - mean centering - auto scaling
MetaboAnalyst (Chong et al., 2019)	Several options provided: - None (no normalization applied) - Sample-specific normalization (i.e., weight, volume) - Normalization by sum - Normalization by median - Normalization by reference sample (probalistic quotient normalization, PQN) - Normalization by a pooled sample from group - Normalization by reference feature - Quantile normalization	Pareto scaling	Other pre-treatment process options: - data transformation (Log transformation and cube root transformation) Other types of scaling selections: (Mean centering, auto scaling and range scaling)
RIKEN Excel Micro tool (Matsuo et al., 2017)	- Normalization method by internal standard	Pareto scaling	Other pre-treatment process options: - data transformation (Log_{10} transformation and 1/4 root transformation) Other types of scaling selections: (Mean centering and auto scaling)

*N/A: not available.

thus, the data has to be 'cleaned' prior to uploading (Sun and Weckwerth, 2013). In addition to this, the SIMCA-P+ software requires a data set with less than 50% missing value (default value) as compared to RIKEN Excel Macro tool which does not state any missing value threshold or percentage. The data set was then manually reviewed according to the modified 80% rule (Yang et al., 2015) as follows:

i. Removal of columns with zero values of peak intensity across all rows of 'mass-to-charge (m/z) ratio: retention time' pairs
ii. Removal of columns with zero values of peak intensity across more than 80% rows from both sites of Keratong and Teluk Intan. This step prepares a better data set than SIMCA-P+ requirement.

Data pre-processing or data 'cleaning' represents an important step in the data mining process to cope with values from analytical platform such as mass spectrometry that will be transformed into a sound data format. Data pre-processing is considered the crux of data interpretation as any steps applied at this stage will affect subsequent statistical analysis and poor data handling could result in or introduce unwanted variation (Engel et al., 2013). The data can be pre-treated and analysed only after a meticulous pre-processing step. This requires an effective process to address variations due to measurement deviations, experimental artefacts and complexity across the samples that can subsequently affect the performance of multivariate data analysis. For instance, data collected from mass spectrometry analyses need to be pre-processed with noise filtering, automatic peak detection or feature detection, chromatographic alignment or normalisation. Further steps of pre-treatment involve mean-centering, scaling or data transformation, although not all of these methods are necessarily used each time (Karaman, 2017).

Data for multivariate data analyses are normally brought to relativity to a factor or 'scaled' to adjust fold differences or to reduce dominance of large spectral feature, if any. This pre-treatment calculates the data dispersion or the size measurements of the data to obtain the mean value. Inevitably, different types of pre-treatment applied to input data can

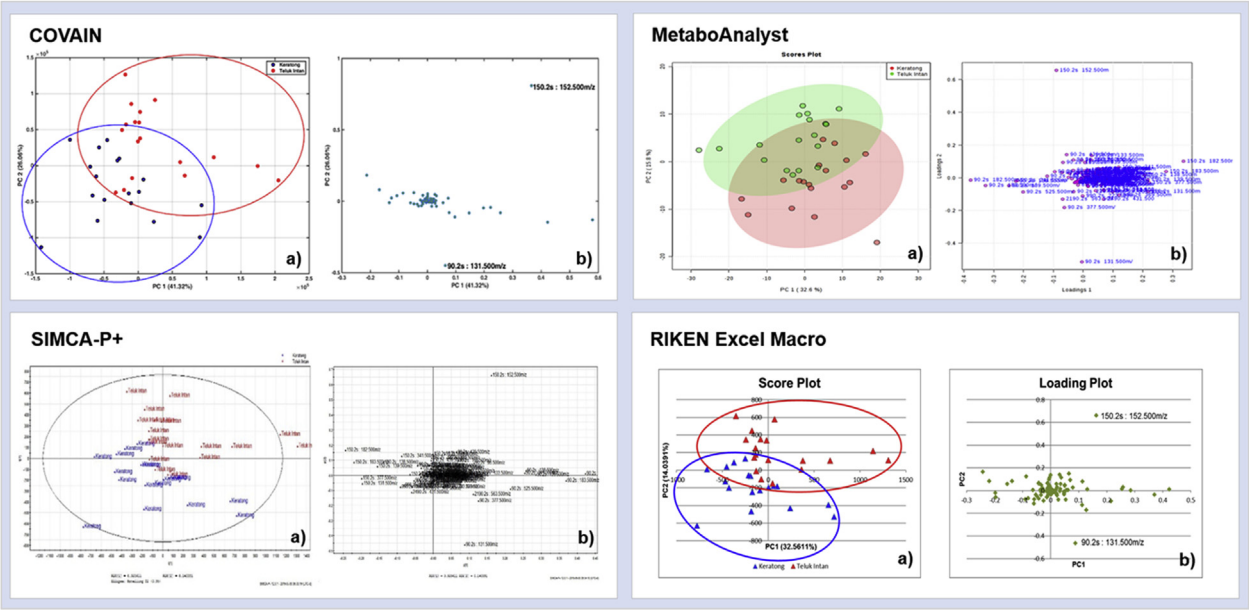

Figure 6. PCA scores (a) and loadings plots (b) of oil palm leaf metabolome of different planting sites generated by COVAIN toolbox, SIMCA-P+, MetaboAnalyst and RIKEN Excel Macro tool.

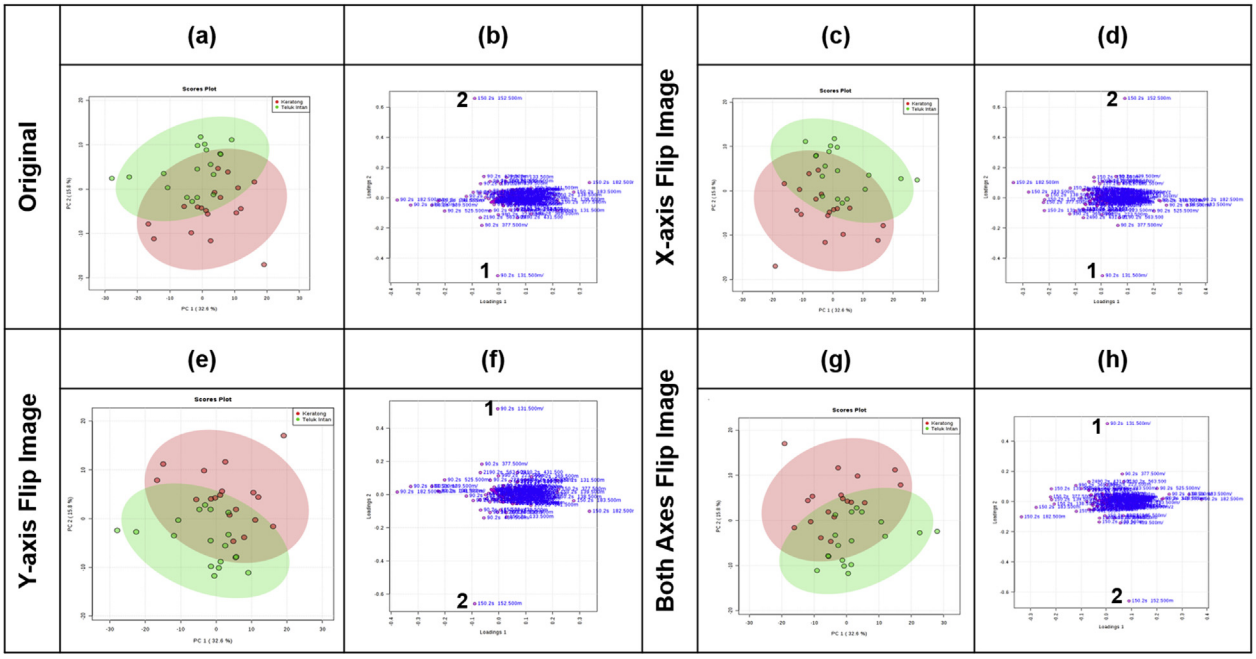

Figure 7. Scores (a) and loadings plots (b) of oil palm leaf metabolome of two different planting sites generated by MetaboAnalyst. The image orientations changed with "flip image" button function on X-axis (c and d), Y-axis (e and f) or both axes (g and h).

influence the results (Van den Berg et al., 2006). Scaling is one of the pre-treatment steps often employed prior to multivariate data statistical model construction which attempts to regulate the fold differences of metabolites by dividing each variable by a factor of dispersion to avoid total dominance of a particular variable as summarised in Table 2 below.

For most of our analysis, Pareto was preferred as the scaling method based on its balanced calculation between mean-centering (subtraction of the variable averages from the data) and unit variance scaling and its ability to decrease metabolite level variation without increasing the measurement deviation of low abundance metabolites (Yang et al., 2015). On another note, this data set was normalised and scaled according to method availability of the individual tools as explained in Table 3.

3.1. Unsupervised method

3.1.1. Principal component analysis (PCA)

PCA is the primary and most widely used unsupervised analysis technique in metabolomics which helps phytochemists to reveal outliers, groups and trends in metabolome data (Madsen et al., 2010). It is usually used for dimensionality reduction of the data by decomposing the spectral data into several principle components that are linear combinations of the original spectral data (Elmasry et al., 2012). In each software platform, the PCA generates scores and loadings plots by which the scores plot represents the original data in a new coordinate arrangement and provide an overview of the observation groups (Xia and Wishart, 2011). The loadings plot reveals the variables that exert influence on the model and are responsible in clustering the groups (Worley and Powers, 2013). Figure 6 shows the PCA scores and loadings plots generated by the four investigated platforms of COVAIN toolbox, SIMCA-P+, MetaboAnalyst and RIKEN Excel Macro tool from the metabolome data of oil palm leaf sample planted on different soil types.

To investigate general interrelation between the oil palm specimens and to observe any clustering and outliers among the samples, COVAIN generated PCA variance occupancy (%) of each principal components (PC), scores plot of all samples in two dimensional space (2D) and three dimensional space (3D), and loadings plot of pair-wise PCs that facilitated interpretation and evaluation of the data (Sun and Weckwerth, 2012). The scores plot shows the vertical plane dispersion of the samples at PC2. The loadings plot reveals the two furthest variable points from the center, conforming to the direction of scores dispersion in the scores plot. Loading point **1** (retention time (t_R); 90.2 s; m/z 131.5) is in the furthermost area corresponding to the samples from Keratong trial while the loading point **2** (t_R 150.2 s; m/z 152.5) approximates to the samples from the Teluk Intan trial in the scores plot.

In SIMCA-P+ visualization of the same data, the PCA scores plot was exhibited as variables principal properties while the loadings plot illustrates variable correlations and model contribution. The scores and loadings plots show distribution between the vertical elements of principal component 2 (PC2) that signify for different soil types of the planting trials. The variables were found to be in association with the dispersion in the scores plot and were labelled as **1** and **2**, with similar attributes to the variables found in COVAIN PCA output, with the same orientation.

For PCA in MetaboAnalyst, user can specify the PCs on the X and Y-axes for the 2D scores and loadings plots to observe the plots in different orientations whether on X-axis, Y-axis or both. The concentration distribution of each variable in the form of box plot can be visualised by clicking on the corresponding variable point in the loadings plot. The server will then generate a detailed report describing each applied method embedded with graphical and tabular outputs (Xia et al., 2009). The results for all analyses performed on the data set will be available in a downloadable.zip file for user for further use. The data will remain on the server for 72 h before being automatically erased. However, if left unattended for 30 min, the website will prompt user to keep working before the session expires in about 10 s. Similar variables were discovered responsible to separate the two planting sites, Keratong and Teluk Intan; loading point **1** (t_R 90.2 s; m/z 131.5) and loading point **2** (t_R 150.2 s; m/z 152.5). However, the orientation of the scores and loadings points are inverse (at x-axis) to that of COVAIN and SIMCA-P+. As MetaboAnalyst provide a function of image flipping at both axes, the PCA scores and loadings after flipping the image at X-axis are comparable to the other tools (Figure 7).

Table 4. Distance and linkage metrics for available tools for clustering.

	Tools	Distance	Linkage	Reference
1	COVAIN (dendrogram + heatmap)	**Euclidean distance**	**Average**	(Sun and Weckwerth, 2013; 2012)
2	SIMCA-P+ (dendrogram)	**Euclidean distance**	**Ward** Single	(Wu et al., 2010)
3	MetaboAnalyst (dendrogram + heatmap)	Euclidean Pearson Minkowski	**Ward** Average Complete Single	(Chong et al., 2019)
	MetaboAnalyst (dendrogram)	Euclidean Pearson Spearman		(Matsuo et al., 2017)

*items in **bold** are default metrics.

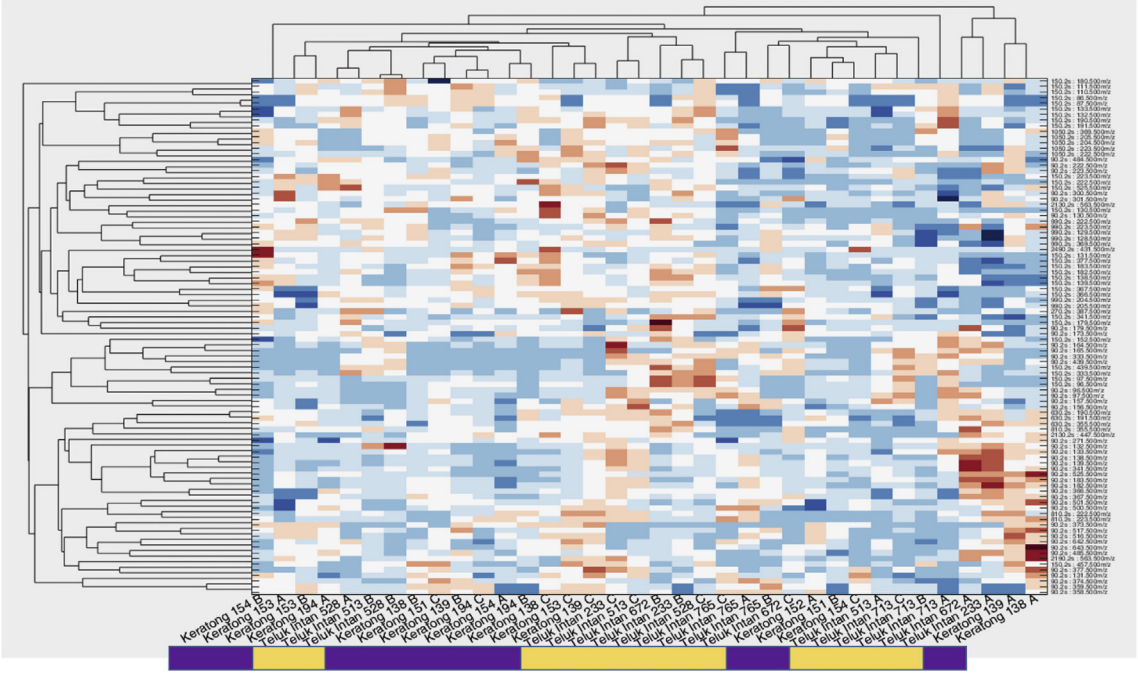

Figure 8. Dendrogram and heat map generated by COVAIN.

Microsoft Excel-based platform offers the approachability and ease of adjustment of figures (Tsugawa et al., 2015). This has led to the development of RIKEN Excel Macro tool for metabolomics data that only necessitates ample proficiency of basic Excel operation. For PCA calculated using the RIKEN Excel Macro tool, the analysis results were in tabular processed data, scores plot, loadings plot, and a PCA result sheet. The scores and loadings data were plotted based on the PCs chosen by the user, in this instance, PC1 *vs.* PC2.

In all PCA results from the tools of SIMCA-P+ platform, COVAIN toolboxes and RIKEN Excel Macro generated PCA plots of similar orientation except to that of MetaboAnalyst with inverse plots with flexible image rotation. This option eventually allowed us to obtain comparable results to other three software with groupings according to planting trials in all scores plots and similar phytochemical candidates from their loadings plots. When comparing the scores plot generated by all four tools, it is evident that grouping could be seen for the samples originating from different planting sites. The loadings plot shows corresponding variation amongst the different groups. In order to determine the variables responsible for separating the data, the loading points that were distinct from the origin in comparison to other points were chosen. All loadings plots from the tools displayed at least two loading points (labelled as points **1** and **2**) located at the furthest points from the origin or center. From here, we can conclude that the loading points **1** and **2** are accountable in the scores and loadings plots pattern. These revealed metabolites or signatory phytochemical markers (m/z ions) which are influential on the clusters as seen on the scores plot are unique for the two different planting sites. Identification of the metabolites using MS/MS (fragmentation of selected ions using collision-induced dissociation (CID) in quadrupole-time-of-flight (Q-TOF) mass spectrometer and liquid chromatography t_R comparison with commercial standards deduced the loadings point **1** as asparagine while metabolite of loadings point **2** was dopamine (Tahir et al., 2016). The identification of these phytochemicals fitted into Level 1 of identity confidence of the Metabolomics Standards Initiative (MSI) (Viant et al., 2017). In other cases of metabolites that are unidentified using available chemical standards, identification can be facilitated by referring to the public databases containing metabolomics data established by metabolome scientists over the years, e.g., the Human Metabolome Database (HMDB), MassBank, METLIN, PubChem, Lipid Metabolites and Pathways Strategy (LIPID MAPS), Kyoto Encyclopaedia of Genes and Genomes (KEGG), Chemical Entities of Biological Interest (ChEBI) and others (Klassen et al., 2017). Each database contains assorted metabolites data information, e.g., spectral data from mass spectrometry and NMR spectroscopy of samples from different species.

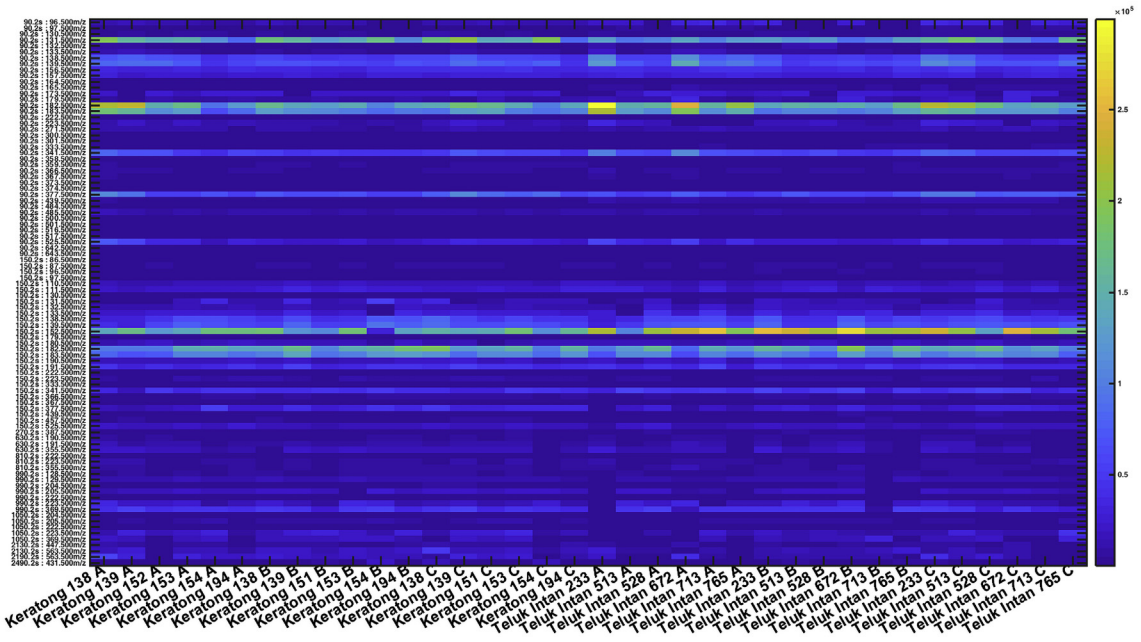

Figure 9. Heat map generated by COVAIN toolbox.

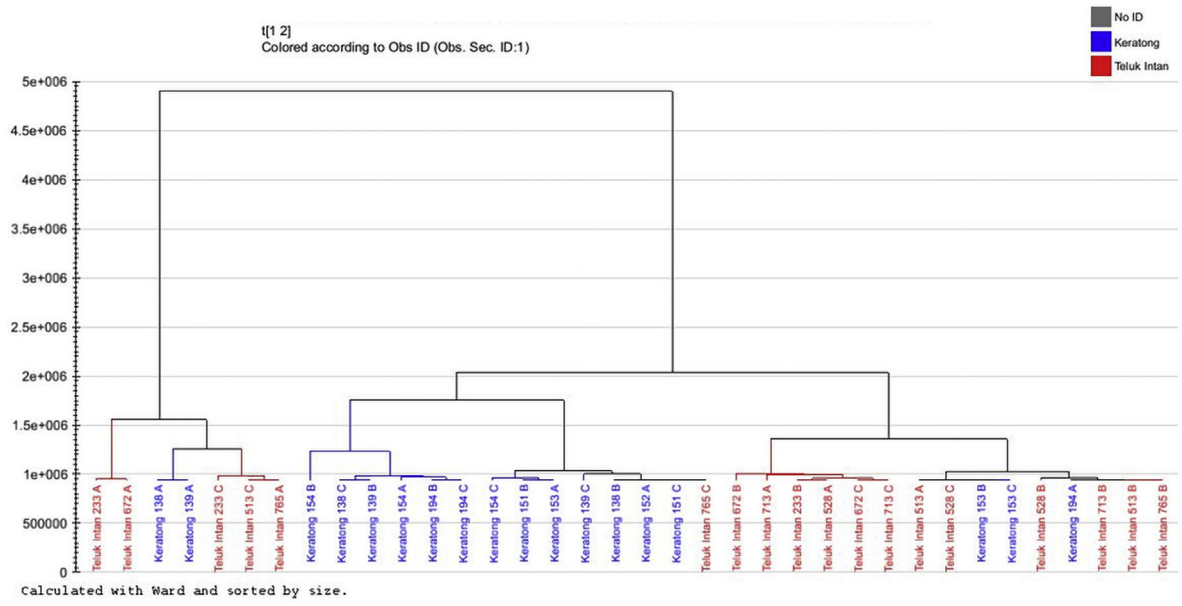

Figure 10. Dendrogram generated by SIMCA-P+.

3.1.2. Clustering

Another unsupervised method for identifying groups in the original data commonly adopted in metabolomics is clustering. Clustering techniques group the subjects in such a way that subjects in the same group are more similar to each other than to subjects in another (Ren et al., 2015). The greater the similarity within a group and the greater the difference amongst groups, the more distinct the clustering is (Tan et al., 2013). Primarily, clustering is employed to examine underlying structure of a data set; to generate hypotheses and to recognise phytochemical features and anomalies. It is also used to classify specimens into their natural forms or relationships and to organise and compress a data set for further assessment (Jain, 2010). Several methods in clustering are hierarchical cluster analysis (HCA), K-means clustering and self-organising map (SOM).

HCA is commonly preferred for small molecule data set due to its straightforward and universal approach as the number of clusters is unknown *a priori* (Boccard et al., 2010). There are two approaches in hierarchical clustering: agglomerative and divisive. The agglomerative clustering technique begins with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains. On the other hand, the divisive clustering is a method for cluster splitting and is known better as top-down clustering. All points are gathered in one big cluster, and as one moves down the hierarchy it splits into smaller size and more similar clusters until each point has its own singleton cluster. Graphically, HCA builds a hierarchy and uses a dendrogram to represent the ranked structure based on similarity levels at which groupings change. HCA is often used together with a heat map to visualize the data matrix with different colours representing different entries in the data matrix (Ren et al.,

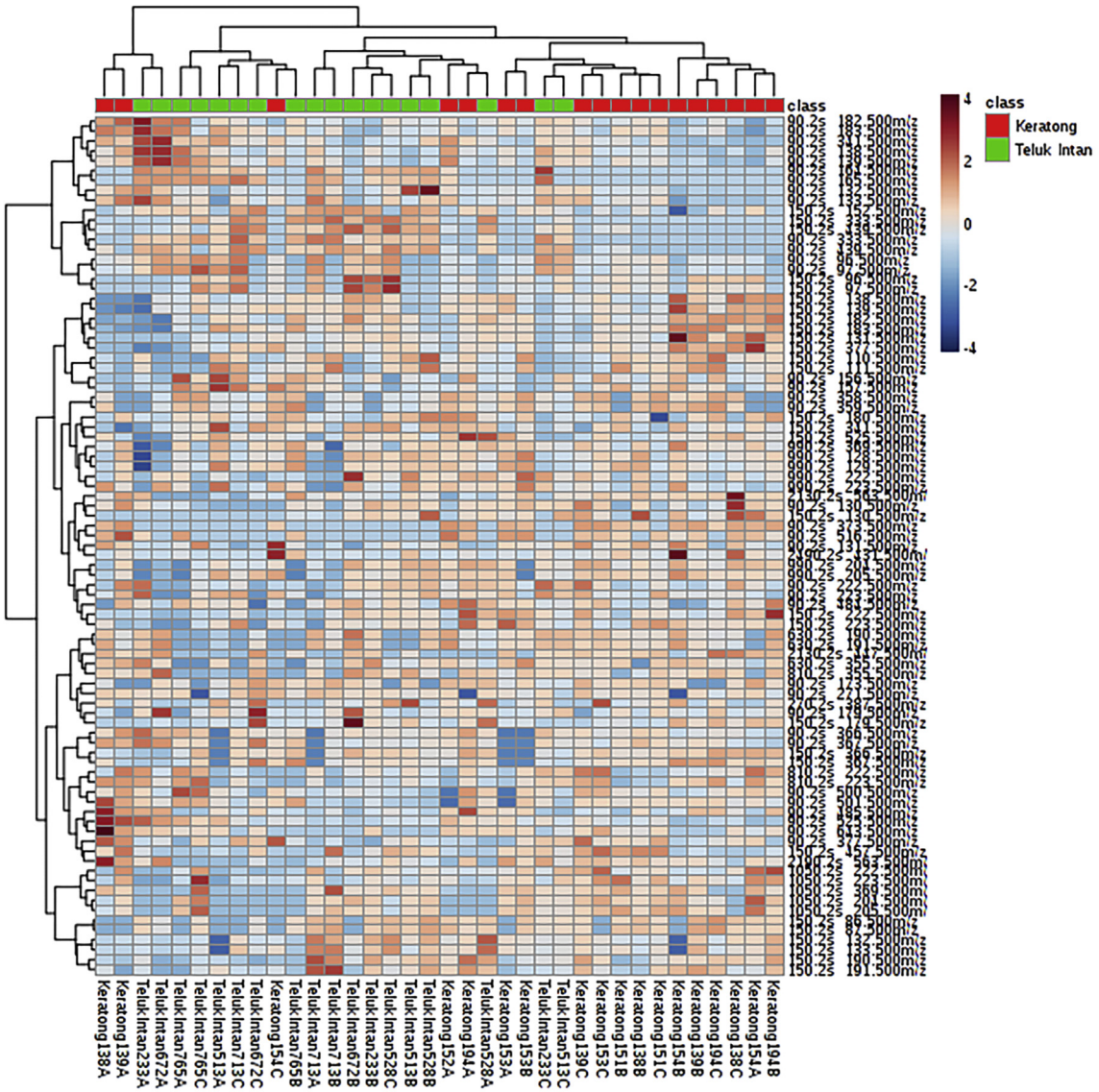

Figure 11. Dendrogram and heat map generated by MetaboAnalyst using default metrics.

2015). The (dis)similarity metrics between pairs of subjects and pairs of clusters must be chosen beforehand in order to form the hierarchical tree since this will influence the shape of the clusters. This kind of metrics are called 'distance' and 'linkage'. 'Distance' defines how far apart two data points are in general while 'linkage' describes and calculates the expanse between two clusters. From our review on the four tools, we found that RIKEN does not offer clustering analysis while MetaboAnalyst provides comprehensive clustering exploration for its user. Ward linkage is the default linkage metrics for MetaboAnalyst and SIMCA-P+. Table 4 summed up the distance and linkage metrics available by the tools for clustering.

Figure 8 shows the dendrogram and heat map generated by COVAIN based on Average linkage of Euclidean distance metric as its default distance. Euclidean is the most commonly used distance function for clustering due to its simplicity and instinctive appeal to reflect the dissimilarity between two patterns. It evaluates the vicinity of objects in two or three-dimensional space (Terziyan, 2017). Here, the samples were not clearly clustered into groups based on their sampling sites. COVAIN also provides another graphical representation of data set that enable a more focused visualisation in a form of a heat map (Figure 9). The locus of several samples from planting trials in contrasting group could be due to technical errors while preparing the samples. The samples are listed in the columns while the mass to charge (m/z) ratio of metabolites are in the rows. The yellow colour indicates higher metabolite abundance with decreasing abundance towards bluer hue. From a close observation of the heat map, higher abundance of the following metabolites in 't_R: m/z' pairs were indicated;

SIMCA-P+ provides two options of 'Ward' or 'Single' linkages for clustering and for this analysis, 'Ward' variance function was applied based on the deduction that 'Ward' method takes into account both the between-cluster and within-cluster distances (Strauss and von Maltitz, 2017). The resulting clusters from our initial experiments using this method produced better output in terms of consistency. While 'Single' linkage assigns two closest points from each cluster according to their similarity, the 'Ward' linkage measures the error increment of each merging cluster (Sartorius Stedim Data Analytics AB, 2017) and computes the increase of error after combining two clusters and minimize the error in consecutive clustering steps (Clifford et al., 2011). The dendrogram generated by SIMCA-P+ in Figure 10 showed slightly different results compared to the output of the COVAIN tool which employs a default Average linkage. Although it produced a clear clustering, more samples were admixed together beyond their sampling soil types.

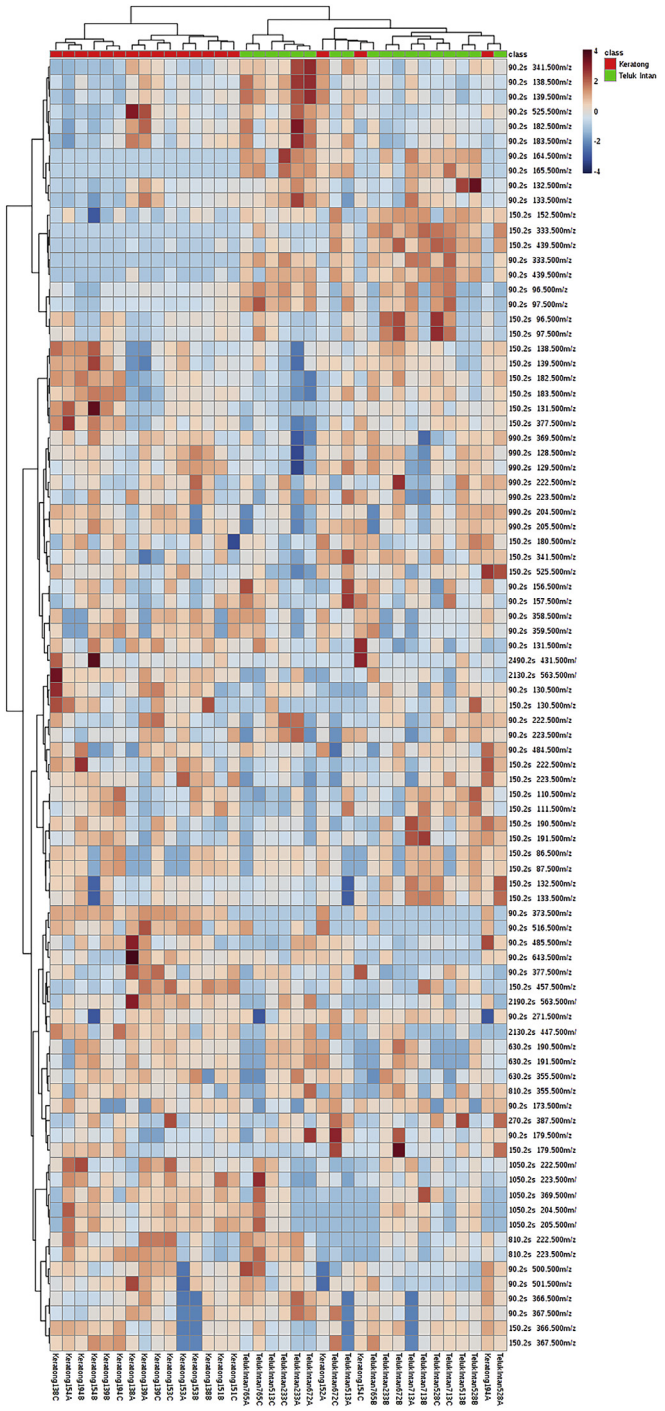

Figure 12. Dendrogram and heat map generated by MetaboAnalyst using Pearson distance metric.

In MetaboAnalyst, there are three types of distances that can be chosen for clustering (dendrogram with heatmap) from the settings namely Euclidean distance, Pearson distance and Minkowski distance while for an individual dendrogram, different options are given; Euclidean distance, Pearson distance and Spearman distance. This tool also provides several types of linkage selection which are Ward, Single, Complete or Average linkages. Figure 11 is the dendrogram and heat map produced from MetaboAnalyst version 4.0 using the default Euclidean distance and Ward linkage.

In Figure 12, the dendrogram and heat map was produced using Pearson distance and Ward linkage to see if better clustering can be obtained for the samples according the planting sites. Pearson distance metric delineates the correlation coefficient between two lists of values (Cox et al., 2005). This result produced a dendrogram with a slightly better cluster for the two planting sites. Figure 13 shows the comparison of the two constructed dendrograms employing the two distance metrics.

In MetaboAnalyst's dendrogram and heat map, the red colour indicates higher metabolite abundance in contrast with lower ones in blue colour. There are several straightforward utility for users to probe into the heatmap results. The results can be seamlessly rearranged according to distance and linkage metrics and several view selections. Based on the dendrogram and heatmap in Figure 14, the top 10 significant variables were viewable from the t-test/ANOVA options. The variable pairs of 90.2 s; *m/z* 131.5, 90.2 s; *m/z* 373.5 and 90.2 s; *m/z* 516.5 were present in most samples from Keratong while variables of 90.2 s; *m/z* 165.5, 90.2 s; *m/z* 164.5, 150.2 s; *m/z* 152.5, 150.2 s; *m/z* 333.5, 150.2 s; *m/z* 439.5, 90.2 s; *m/z* 333.5 and 90.2 s; *m/z* 439.5 were detected in more samples from Teluk Intan. This function allows users to indicate important variables and samples for further exploration and inspection.

In a separate dendrogram utility in MetaboAnalyst, another comparison using default Euclidean and Spearman distance metrics was reviewed as shown in Figure 15. The Spearman metric is a nonparametric (distribution-free, not adhering to assumptions) distance measure and is less influenced by outliers, such as the presence of lower or higher intensity (Cox et al., 2005). Better clusters were recorded with Spearman distance and Ward linkage by which the samples were generally assembled into groups based on their sampling sites compared to its Euclidean default.

Altogether, different outputs of clustering have been produced from MetaboAnalyst, COVAIN toolbox and SIMCA-P+. From the four statistical tools used in this analysis, MetaboAnalyst and COVAIN are found to be equipped with both dendrogram and heat map function while SIMCA-P+ provides dendrogram construction tool. Although the clustering in SIMCA-P+ and MetaboAnalyst began by using default method of similar Euclidean distance and Ward linkage metrics, the result of the dendrograms from these tools are not alike. This could probably be due to different background or underlying algorithm behind their "distance" and "linkage" metrics in addition to the various pre-processing steps involved prior to clustering. It was also found that several samples deviated from their respective technical replicates in the clusters as listed in Table 5 below:

As observed from the dendrograms generated by these three tools, differences within the groups of samples were managed to be detected by the clustering methods of dendrogram and heatmap as compared to the output of PCA although all of these approaches are of unsupervised nature. This is due to the fact that the clustering algorithm consecutively pairs entities together for the highest degree of similarity while PCA reduces dimensions of the data for total variability or divergence (Robledo et al., 2015). More biological and technical replicates for a solid data set in addition to scrutiny into the pre-processing parameters could

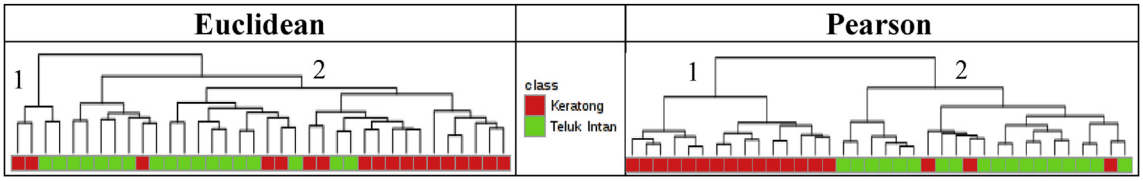

Figure 13. Dendrograms from heat map generated by MetaboAnalyst using Euclidean and Pearson distance metrics.

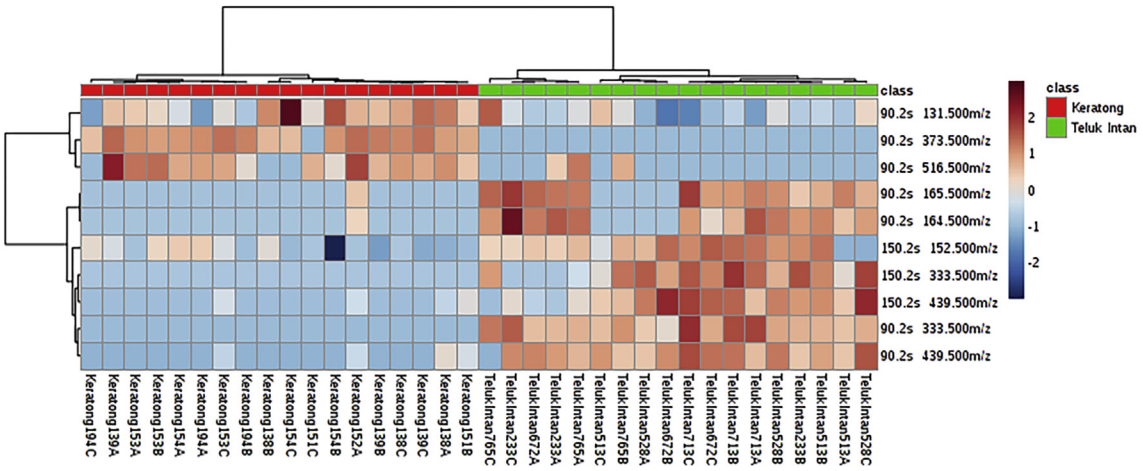

Figure 14. Top 10 variables based on t-test/ANOVA analysis using Pearson distance and Ward clustering algorithm presented in dendrograms and heat map generated by MetaboAnalyst.

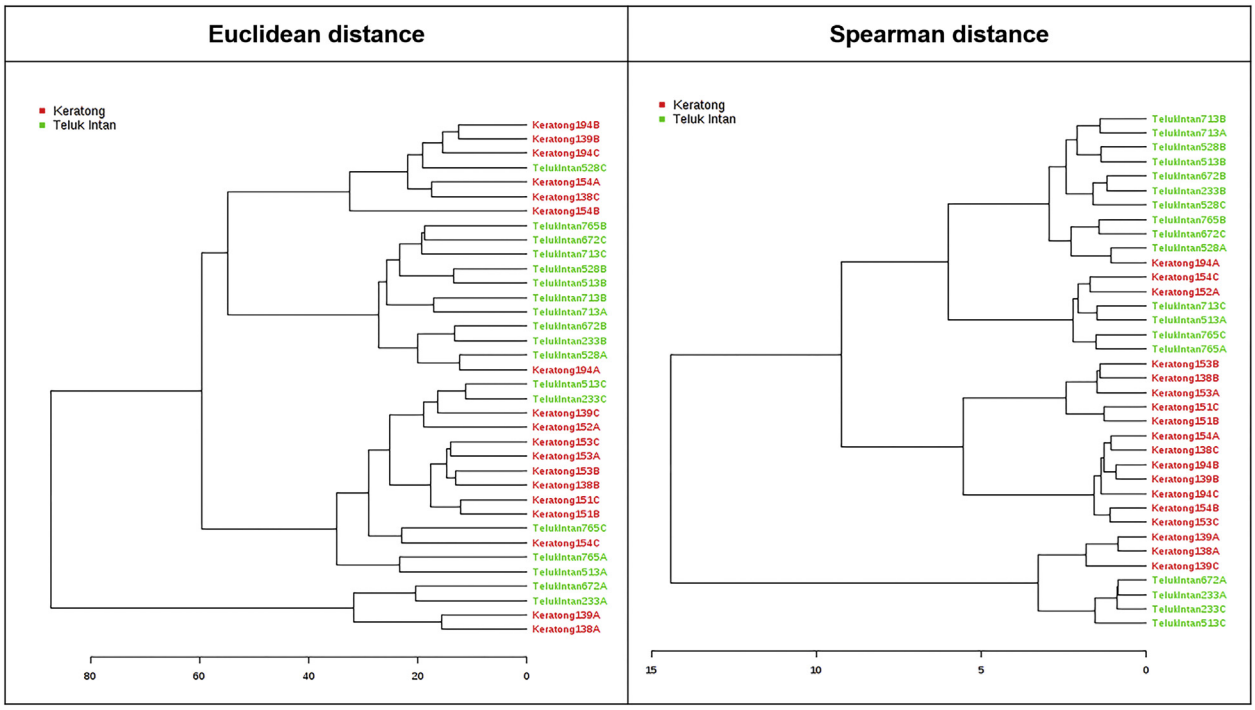

Figure 15. Dendrograms generated by MetaboAnalyst with Euclidean and Spearman distances.

illuminate the factors contributing to these different findings. While pre-processing steps for multivariate data analysis aided in the scaling and normalisation of the data set throughout PCA, the technical variation found in several samples were quite profound that it was eventually observed in the cluster analysis. It is no doubt that the linearity, precision, repeatability, stability and accuracy of an analytical method are the critical criteria for a sound phytochemical analysis (Sun et al., 2017). Human errors that contributed to the sample mishmash should be traced and investigated and more technical replicates should be prepared to investigate the magnitude of the error. Multistep extraction procedures increases the chance of error introduction and should be minimised (Emwas et al., 2016) and attention should be given to the protocols and technical approaches for maximum reliability, reproducibility, and sensitivity of analysis due to the fact that a metabolome is a very dynamic unit and is highly responsive to different stimuli, e.g., sample handling and sensitive to surrounding for instance thermal degradation (Fang et al., 2015).

3.2. Supervised method

3.2.1. Partial least squares discriminant analysis (PLS-DA)

PLS-DA is often applied in the encounter of many possible correlated predictor variables in a matrix of responses. It maximises and exploits the covariance between the variables (X) and the informative response (y), when response (y) is available in a challenge to understand which variables carry the class separating information. PLS-DA produces scores vectors and loadings vectors, similar as PCA output. The strategy differs from PCA in that it includes the additional input (vector y) by adjusting the model to capture the Y-related variation in X and also to enhance the poor clustering obtained with the PCA model. PLS-DA is frequently employed in cases with two classifications of participants such as treated versus untreated control groups or infected versus healthy control. Several members of PLS-DA family are PLSR-DA (PLS Regression-DA), Orthogonal-PLS (OPLS), Orthogonal PLS-DA (OPLS-DA), Power-PLS

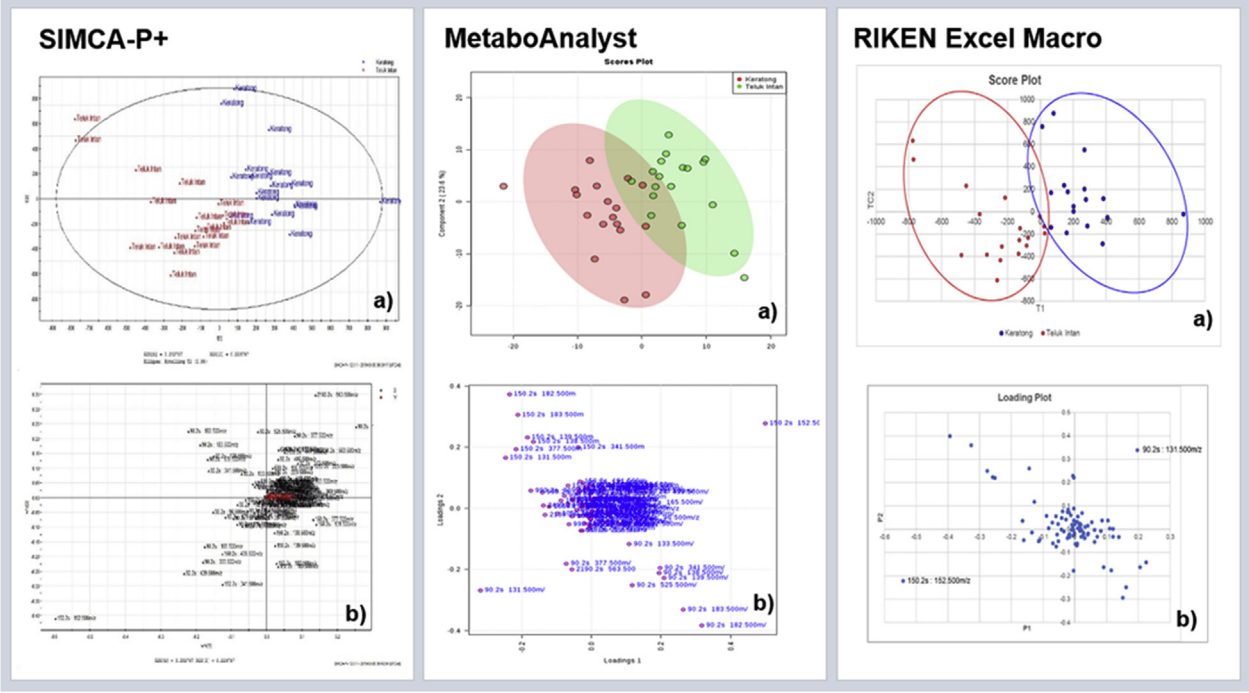

Figure 16. PLS-DA scores (a) and loadings plots (b) of oil palm leaf metabolome of different planting sites generated by SIMCA-P+, MetaboAnalyst and RIKEN Excel Macro tool.

(PPLS) and its adaptation; PPLS-DA and Canonical Powered PLS (CPPLS) (Liland, 2011).

The PLS-DA plots in SIMCA-P+ displayed the relation between the X-variables and the Y-variables with values of R2X (cum) of 0.500, R2Y (cum) of 0.904929 and Q2 (cum) of 0.63923. The PLS-DA scores plot in Figure 16 shows a complete and significant separation of these two sample groups of different planting trials. While in the loadings plot, similar variables as in PCA were discovered responsible to separate the two planting sites, Keratong and Teluk Intan; loadings point **1** (t_R 90.2 s; m/z 131.5) and loadings point **2** (t_R 150.2 s; m/z 152.5). The loadings line plot in Figure 17 shows the X-variables with red circled peak pinnacles that are responsible in separating the data into two groups. The Variable Importance for the Projection (VIP) plot (Figure 18) summarises the importance of the variables for interpreting X and correlating to Y. VIP values larger than 1.0 indicates "important" X-variables and values lower than 0.5 indicates "unimportant" X-variables. These "important" X-variables are the variables responsible for distinguishing the data into two group of clusters. The interval between 1.0 and 0.5 is a grey area, where the importance level depends on the size of the data set. However, only variables with VIP values > 2.0 were selected for further data analysis. Such a strict criterion was set because of the large number of variables in the plot. The following variables of 't_R: m/z' pairs should be chosen for further analysis, e.g., identification and pathway mapping:

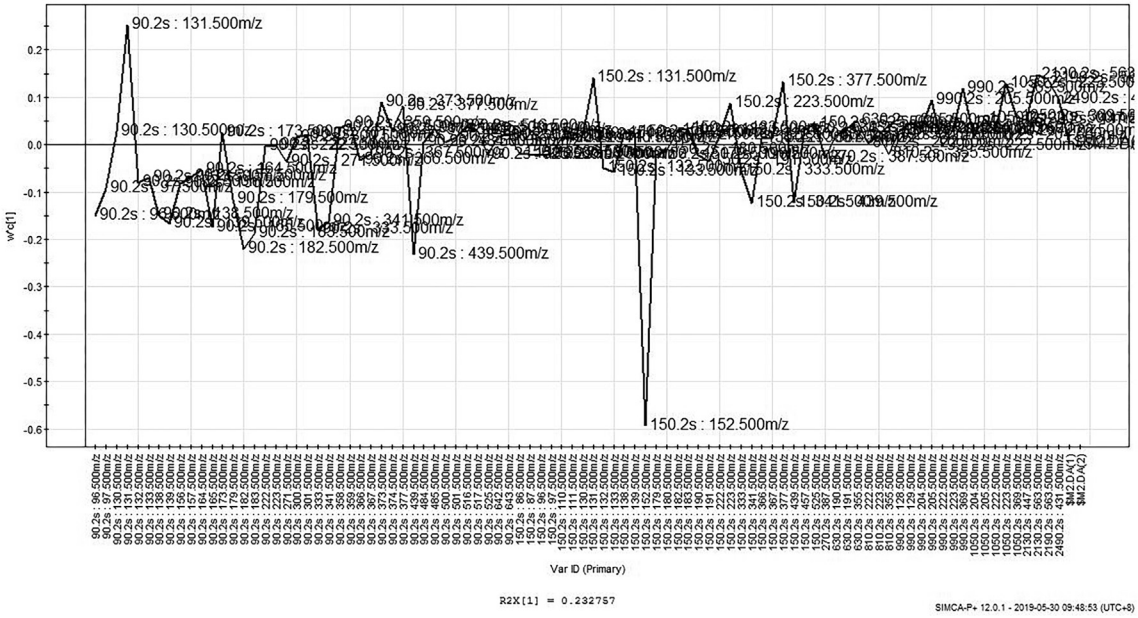

Figure 17. PLS-DA loadings line plot generated by SIMCA-P+.

Table 5. Examples of deviating samples discovered from clustering analysis.

Tools	Teluk Intan Cluster	Keratong Cluster
COVAIN (dendrogram + heatmap)	Keratong 154	Teluk Intan 528
	Keratong 152	Teluk Intan 513
	Keratong 151	
	Keratong 139	
	Keratong 138	
SIMCA (dendrogram)	Keratong 194	Teluk Intan 765
	Keratong 153	Teluk Intan 672
		Teluk Intan 513
		Teluk Intan 233
MetaboAnalyst (dendrogram + heatmap using Pearson distance)	Keratong 194	
	Keratong 154	
	Keratong 152	
MetaboAnalyst (dendrogram using Spearman distance)	Keratong 194	Teluk Intan 672
	Keratong 154	Teluk Intan 513
	Keratong 152	Teluk Intan 233

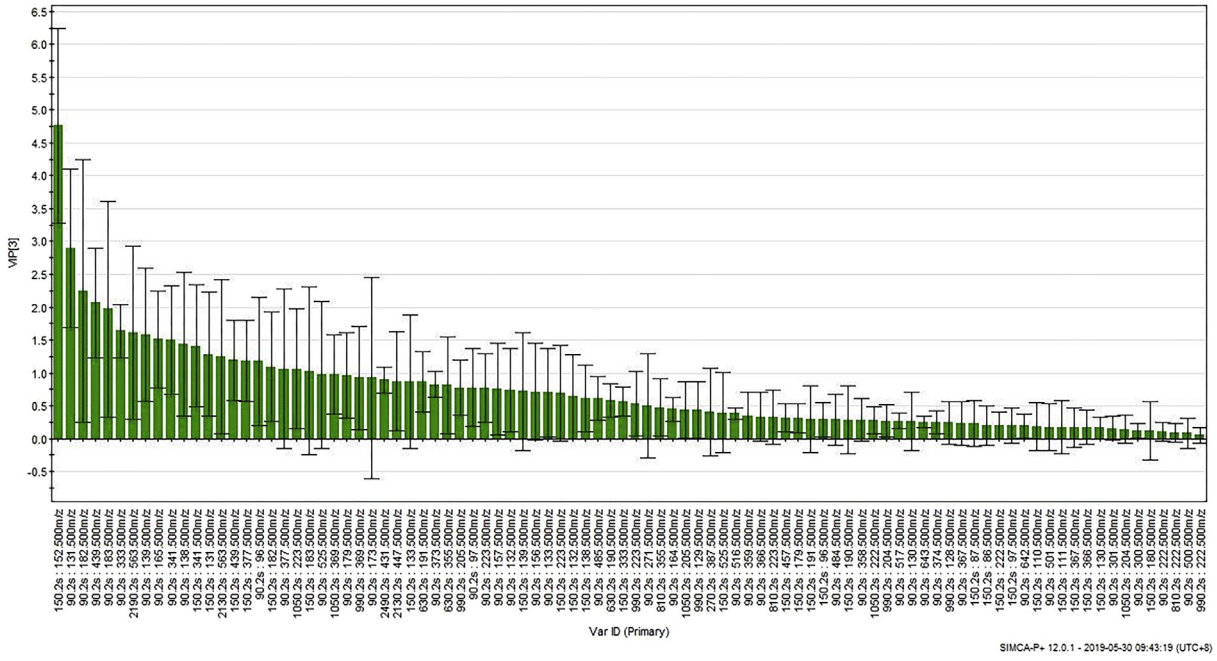

Figure 18. PLS-DA VIP plot generated by SIMCA-P+.

The VIP scores generated in MetaboAnalyst version 4.0 shown in Figure 19 demonstrates the top 15 important features identified by PLS-DA in Figure 16. These identified variables were responsible for discriminating phytochemical profiles of the two planting sites in the PLS-DA scores and loadings plots. The coloured boxes on the right indicate the relative abundances of the corresponding metabolites in the two planting sites; red colour indicates high abundance and green colour indicates low abundances.

From the VIP scores of the top 15 variables, *m/z* 152.5 at 150.2 s and *m/z* 131.5 at 90.2 s were identified as the metabolites that significantly contributed to the class separation of oil palm leaf metabolome profiles from the different planting sites. Metabolite *m/z* 152.5 is of higher abundance in samples from Teluk Intan and metabolite *m/z* 131.5 is higher in samples from Keratong. The MetaboAnalyst loadings plot of PLS-DA (Figure 16 (b)) shows that the variables similar to the VIP findings are plotted further away from the origin which are *m/z* 131.5 at 90.2 s and *m/z* 152.5 at 150.2 s. These variables are responsible in separating the data into two groups of clusters as shown in the scores plot. To validate the statistical modelling of the two different planting sites in the PLS-DA model, a cross validation test was conducted in MetaboAnalyst with the result of R2 = 0.95995 and Q2 = 0.58703. R2 reading represents the model's degree of fitness and Y variables quality while Q2 reading explains the predictive quality of the PLS model with the best model having value closer to 1.0 (Khoo et al., 2015). The loadings plot generated by RIKEN Excel Macro Tool in Figure 16 (b) shows *m/z* 131.5 at 90.2 s and *m/z* 152.5 at 150.2 s as the most influential metabolites in the separation of the two different sample groups. The loadings plot produced by the RIKEN Excel Macro tool was comparable to the loadings plots from the MetaboAnalyst platform in terms of orientations and dispersal of elements in its scores and loadings plots. A

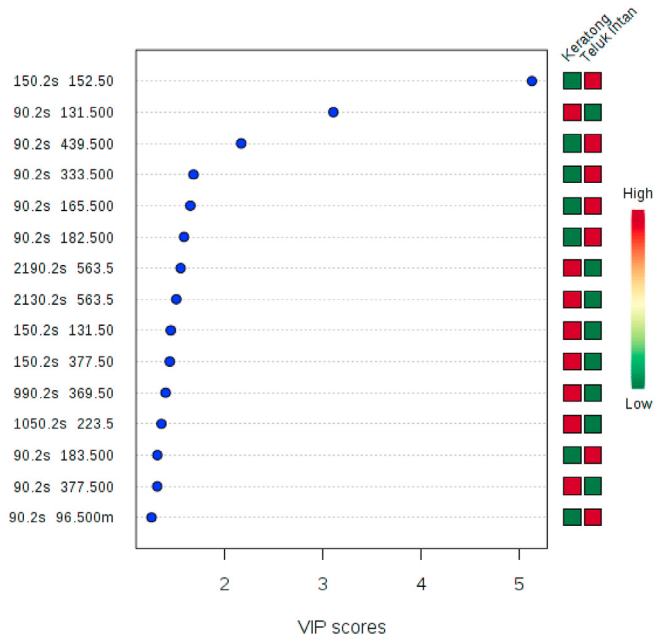

Figure 19. PLS-DA variable importance in projection (VIP) by MetaboAnalyst.

cross-validation test validated the model with a result of Q2 = 0.63924. More than 15 metabolites with VIP scores greater than 1.0 were observed across different samples from both Keratong and Teluk Intan trials. However, due to the large number of variables in the plot, only variables with VIP values >2.0 were selected for further analysis (Figure 20). They are:

From the three software platforms providing supervised methods, the recorded important variables from SIMCA-P+ output are similar to those of RIKEN Excel Macro tool. MetaboAnalyst on the other hand, listed 90.2 s: *m/z* 333.5 in its lists of VIP values >2.0. The influential loadings of *m/z* 131.5 at 90.2 s and *m/z* 152.5 at 150.2 s for the PLS-DA from all three tools were consistent. When SIMCA-P+ version 12 was compared to one of the more recent edition, we noticed its improved performance in term of processing speed and its user-friendliness, i.e., pop-up windows. Nevertheless, our existing SIMCA-P+ version still works reasonably on par with the other tested platforms. Based on the testing of each tool, every platform has its own function and features in analysing the data. The runtimes of all installed tools were comparably fast with only MetaboAnalyst running online and every statistical plot was generated in less than 2.0 s Table 6 summarises the platforms that were used for the oil palm metabolome data analysis.

4. Conclusion

Subtle changes in the environment affect crop productivity and performance whether on the instant or in the long term. As part of our effort to streamline the ecometabolomics workflow for sustainable oil palm, the utilization of a rapid, straightforward and cost-effective method of field data interpretation is ideal. In this investigation, the raw LC-MS data set from an earlier published work (Tahir et al., 2016) was used for comparison of COVAIN toolbox, MetaboAnalyst, SIMCA-P+, and RIKEN Excel Macro statistical analysis tools. The real data set represents the individual biological palms of similar genetic background to put the genetic factor as a constant and to monitor the plants components such as the metabolome that responds to environmental stimuli in a metabolic homeostasis (Nagler et al., 2018). The four software that were assessed have helped us to reveal the different methods, purposes and interfaces of each approach in interpreting the metabolome data. This comparison allows identification of relationships and the capabilities of respective statistical analysis tool for oil palm metabolomics research. There is no single statistical tool that possess an entirely desirable features as every tool has its own advantages and limitations. The statistical tools are comprised of different multivariate data analysis for complex metabolomics data as such used in the literatures (Bartel et al., 2013; Worley and Powers, 2013). There are many methods that can be applied in analysing raw metabolome data, and both unsupervised and supervised techniques were used in this study. Depending on functions and objectives of the exploration, the unsupervised and supervised approaches have no superiority over one another. Nevertheless, unsupervised methods are routinely used as an initial step to limit and consequently delineate the most significant remaining variables, often to attain a global overview of the data set and to allow generation of hypotheses that

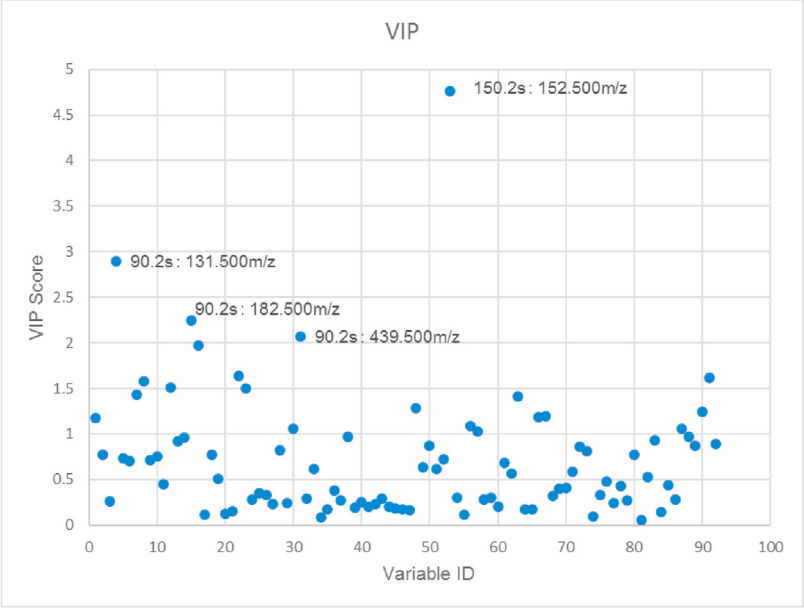

Figure 20. PLS-DA VIP scores by RIKEN Excel Macro tool.

Table 6. Comparison of investigated metabolomics statistical tools.

Tools	COVAIN toolbox Version 2017-May-16	SIMCA-P+ version 15.2	MetaboAnalyst version 4.0	RIKEN Excel Macro Tool
Type of platform	Toolbox with window or pane with quick access to common operation functions in the program	Licensed software	Web server	Tool
Statistical analysis offered	PCA ICA Clustering (Dendrogram and heat map utility) Correlation analysis	PCA PLS, OPLS PLS-DA, OPLS-DA Clustering (Dendrogram utility, PLS-tree, etc.)	PCA PLS-DA OPLS-DA Clustering (Dendrogram and heat map utility, K-means, self-organizing map (SOM), etc.)	PCA PLS-R PLS-DA Correlation analysis
Cost	COVAIN itself is an open source software but annual renewal of MATLAB license costs at least USD29 (student license) on top of one-time software and tools package purchase	Perpetual software license for one time purchase	Free online and offline open source local installation using Web Application Resource (.war) file	Microsoft Excel which cost at least USD140 (for Home & Student version)
Runtime (Analysis time)	1.79 s	1.91 s	Runtime of the online tool depends on user's internet connection unless locally installed.	1.29 s
Limitations	• Requires MATLAB licensed software to run the toolbox • Supports.xls or.xlsx and.txt file format • No supervised methods function	• The software product need to be purchased at a cost • Support.xls or.xlsx,.csv and.txt file format • No heat map utility	• Requires internet connections • Support.csv,.txt and mass spectrometry file formats • The information for pathway and metabolite identification for plant and microbial metabolism is still improving	• Requires: -Windows operating system -Microsoft Excel Version 2010 • Supports.csv file format • No dendrogram and heat map utility
Advantages	• Freely accessible for download • Easy to use	• Excellent graphic capabilities • Comprehensive analysis options for three multivariate data analysis methods	• Freely accessible • Easy to use • Offers many statistical analysis methods • Has its own metabolite and pathway identification tools	• Freely accessible • Easy to use • Easy adjustment of figures
Experience of use	Generated figures from statistical analysis are adjustable. However, data pre-processing parameters and supervised methods are limited. Apart from statistical analysis, COVAIN tool consist of Granger time-series analysis, pathway mapping, correlation network topology analysis and visualization.	A well-known data mining software for more than 30 years. However, users need to clearly define variables including identifiers for the variables, the roles of the variables, the data type which can be quite confusing for the newbies. This could be due to usage of SIMCA-P+ for multiple fields other than metabolomics.	A complete pipeline for high-throughput metabolomics starting from data pre-processing, multivariate data analysis and data annotation. It provides interesting functions-biomarker analysis, various pathway analysis, etc. The software is constantly updated for public use.	A user friendly tool with comprehensive manual. Figures generated are easy to adjust due to familiarity of Microsoft Excel.

can then be verified statistically using supervised approach (Noto et al., 2016).

In common circumstances, the methods and tools that are chosen by users for their data analysis are based on the capability and availability of these platforms. Cost for obtaining and renewing software licenses may influence the decision to choose or switch over to publicly available tools. At present, there are many methods and tools that have been established to aid phytochemists and these methods need to be explored in order to find and tailor the most efficient and optimum data interpretation platform into a metabolomics workflow. Several initiatives have been carried out to fill the gaps to tackle the data rich, information poor quandary (Tachibana, 2014) and the tools that go as far as distinguishing and mapping the metabolites into pathways and integrates the findings with other "omics" platforms are the crème de la crème (Klupczyńska et al., 2015). Overall, with regards to cost and versatility, this investigation gave a better understanding and a foundation for the application of the best statistical tools into the oil palm research pipeline which at present was found to be MetaboAnalyst.

While applying the tools onto the raw data set, it was found that apart from discovering differences and similarities between the sample groups relevant to the hypothesis, multivariate data analysis is useful for critical assessment of data integrity. While biological contamination could cause a sample fail to cluster within the group, technicalities issues such as repeated measurements of biological replicates would allow the evaluation of the clustering robustness. In addition to this, precision and accuracy of the analysis could be improved by looking at the intra- and inter-day relative standard deviation (RSD) for example analysing at least five replicates of quality control (QC) samples at three different concentrations on the same day and for at least three consecutive days. The concentration of each QC sample is then calculated using a standard curve of the day. The precision can be determined as the RSD while the accuracy is established as the relative error (RE) (Peng et al., 2014). Robust analytical design for LC-MS is crucial as the method is less reproducible than other phytochemistry methods in terms of its hardware/parts (e.g., column, septa) and other parameters (mobile phase composition and preparation, analysis temperature throughout the setup) (Defernez and Le Gall, 2013).

Declarations

Author contribution statement

Nur Ain Ishak, Noor Idayu Tahir: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Syafi'ah Nadiah Mohd Sa'id: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Kathiresan Gopal: Analyzed and interpreted the data; Wrote the paper.

Abrizah Othman, Umi Salamah Ramli: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Malaysian Palm Oil Board (Board Approved Program) (R009911000).

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors thank the Director-General of MPOB for the permission to publish this article. We are also grateful to the Proteomics and Metabolomics (PROMET) research team especially Madam Hasliza Hassan and Nurul Liyana Rozali for their valuable technical assistance.

References

Barcelos, E., Rios, S.A., Cunha, R.N.V., Lopes, R., Motoike, S.Y., Babiychuk, E., Skirycz, A., Kushnir, S., 2015. Oil palm natural diversity and the potential for yield improvement. Front. Plant Sci. 6, 190.

Bartel, J., Krumsiek, J., Theis, F.J., 2013. Statistical methods for the analysis of high-throughput metabolomics data. Comput. Struct. Biotechnol. J. 4, e201301009.

Beisken, S., Eiden, M., Salek, R.M., 2015. Getting the right answers: understanding metabolomics challenges. Expert Rev. Mol. Diagn. 15 (1), 97–109.

Béné, C., Oosterveer, P., Lamotte, L., Brouwer, I.D., de Haan, S., Prager, S.D., et al., 2019. When food systems meet sustainability -Current narratives and implications for actions. World Dev. 113, 116–130.

Boccard, J., Veuthey, J.-L., Rudaz, S., 2010. Knowledge discovery in metabolomics: an overview of MS data handling. J. Separ. Sci. 33 (3), 290–304.

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.

Chong, J., Wishart, D.S., Xia, J., 2019. Using metaboanalyst 4.0 for comprehensive and integrative metabolomics data analysis. Curr. Protoc. Bioinformatics 68–86.

Clifford, H., Wessely, F., Pendurthi, S., Emes, R.D., 2011. Comparison of clustering methods for investigation of genome-wide methylation array data. Front. Genet. 2 (88).

Cox, B., Kislinger, T., Emili, A., 2005. Integrating gene and protein expression data: pattern analysis and profile mining. Methods 35 (3), 303–314.

Defernez, M., Le Gall, G., 2013. Strategies for data handling and statistical analysis in metabolomics studies. In: Rolin, D. (Ed.), Advances in Botanical Research, 67. Academic Press, pp. 493–555. Chapter 11.

Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., et al., 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. J. Separ. Sci. 9 (1), 44–66.

Elmasry, G., Kamruzzaman, M., Sun, D.-W., Allen, P., 2012. Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review. Crit. Rev. Food Sci. Nutr. 52 (11), 999–1023.

Emwas, A.-H., Roy, R., McKay, R.T., Ryan, D., Brennan, L., Tenori, L., et al., 2016. Recommendations and standardization of biomarker quantification using NMR-based metabolomics with particular focus on urinary analysis. J. Proteome Res. 15 (2), 360–373.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J.J., Downey, G., Blanchet, L., et al., 2013. Breaking with trends in pre-processing? Trac. Trends Anal. Chem. 50, 96–106.

Fang, M., Ivanisevic, J., Benton, H.P., Johnson, C.H., Patti, G.J., Hoang, L.T., et al., 2015. Thermal degradation of small molecules: a global metabolomic investigation. Anal. Chem. 87 (21), 10935–10941.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31 (8), 651–666.

Johnson, C.H., Gonzalez, F.J., 2012. Challenges and opportunities of metabolomics. J. Cell. Physiol. 227 (8), 2975–2981.

Karaman, I., 2017. Preprocessing and Pretreatment of Metabolomics data for statistical analysis. In: Sussulini, A. (Ed.), Metabolomics: from Fundamentals to Clinical Applications. Springer International Publishing, Cham, Switzerland, pp. 145–161.

Khoo, L.W., Mediani, A., Zolkeflee, N.K.Z., Leong, S.W., Ismail, I.S., Khatib, A., et al., 2015. Phytochemical diversity of *Clinacanthus nutans* extracts and their bioactivity correlations elucidated by NMR based metabolomics. Phytochem. Lett. 14, 123–133.

Klassen, A., Faccio, A.T., Canuto, G.A.B., da Cruz, P.L.R., Ribeiro, H.C., Tavares, M.F.M., et al., 2017. Metabolomics: definitions and significance in systems biology. In: Sussulini, A. (Ed.), Metabolomics: from Fundamentals to Clinical Applications. Springer International Publishing, Cham, Switzerland, pp. 3–17.

Klupczyńska, A., Dereziński, P., Kokot, Z.J., 2015. Metabolomics in medical sciences - trends, challenges and perspectives. Acta Pol. Pharm. 7 (24), 629–641.

Kushairi, A., Loh, S.K., Azman, I., Hishamuddin, E., Ong-Abdullah, M., Izuddin, Z.B., Shamala, S., Parveez, G.K.A., 2018. Oil palm economic performance in Malaysia and R&D progress in 2017. J. Oil Palm Res. 30 (2), 163–195.

Liland, K.H., 2011. Multivariate methods in metabolomics–from pre-processing to dimension reduction and statistical analysis. Trac. Trends Anal. Chem. 30 (6), 827–841.

Madsen, R., Lundstedt, T., Trygg, J., 2010. Chemometrics in metabolomics-a review in human disease diagnosis. Anal. Chim. Acta 659 (1), 23–33.

Matsuo, T., Tsugawa, H., Miyagawa, H., Fukusaki, E., 2017. Integrated strategy for unknown EI-MS identification using quality control calibration curve, multivariate analysis, EI-MS spectral database, and retention index prediction. Anal. Chem. 89, 6766–6773.

Nagler, M., Nägele, T., Gilli, C., Fragner, L., Korte, A., Platzer, A., Farlow, A., Nordborg, M., Weckwerth, W., 2018. Eco-metabolomics and metabolic modeling: making the leap from model systems in the lab to native populations in the field. Front. Plant Sci. 9, 1556.

Noto, A., Pomero, G., Mussap, M., Barberini, L., Fattuoni, C., Palmas, F., et al., 2016. Urinary gas chromatography mass spectrometry metabolomics in asphyxiated newborns undergoing hypothermia: from the birth to the first month of life. Ann. Transl. Med. 4 (21), 417.

Oettli, P., Behera, S.K., Yamagata, T., 2018. Climate based predictability of oil palm tree yield in Malaysia. Sci. Rep. 8 (1), 2271.

Peng, J.-B., Luo, C.-H., Wang, Y.-C., Huang, W.-H., Chen, Y., Zhou, H.-H., et al., 2014. Validation of a liquid chromatography-electrospray ionization-tandem mass spectrometry method for determination of all-trans retinoic acid in human plasma and its application to a bioequivalence study. Molecules 19 (1).

Ren, S., Hinzman, A.A., Kang, E.L., Szczesniak, R.D., Lu, L.J., 2015. Computational and statistical analysis of metabolomics data. Metabolomics 11 (6), 1492–1513.

Robledo, J.I., Saánchez, H.J., Leani, J.J., Pérez, C.A., 2015. Exploratory methodology for retrieving oxidation state information from X-ray resonant Raman scattering spectrometry. Anal. Chem. 87, 3639–3645.

Saccenti, E., Hoefsloot, H.C.J., Smilde, A.K., Westerhuis, J.A., Hendriks, M.M.W.B., 2014. Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics 10 (3), 361–374.

Sartorius Stedim Data Analytics, A.B., 2017. SIMCA® 15 Multivariate Data Analysis Solution User Guide. ID #2084. Guide Edition Date: November 17, 2017. Sartorius Stedim Biotech, Umeå, Sweden.

Sindelar, M., Patti, G.J., 2020. Chemical discovery in the era of metabolomics. *J.* Am. Chem. Soc. 142 (20), 9097–9105.

Strauss, T., von Maltitz, M.J., 2017. Generalising ward's method for use with manhattan distances. PloS One 12 (1), e0168288.

Sukiran, M.A., Loh, S.K., Bakar, N.A., 2018. Conversion of pre-treated oil palm empty fruit bunches into bio-oil and bio-char via fast pyrolysis. J. Oil Palm Res. 30 (1), 121–129.

Sun, X., Weckwerth, W., 2012. COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. Metabolomics 8 (1), 81–93.

Sun, X., Weckwerth, W., 2013. Using COVAIN to analyze metabolomics data. In: The Handbook of Plant Metabolomics. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 305–320.

Sun, Y., Li, B., Lin, X., Xue, J., Wang, Z., Zhang, H., et al., 2017. Simultaneous determination of four triterpenoid saponins in aralia elata leaves by HPLC-ELSD combined with hierarchical clustering analysis. Phytochem. Anal. 28 (3), 202–209.

Szymańska, E., 2018. Modern data science for analytical chemical data -A comprehensive review. Anal. Chim. Acta 1028, 1–10.

Tachibana, C., 2014. What's next in 'omics: the metabolome. Science 345 (6203), 1519–1521.

Tahir, N.I., Shaari, K., Abas, F., Ishak, Z., Tarmizi, A.H., Amiruddin, M.D., Parveez, G.K.A., Ramli, U.S., 2016. Metabolome analysis of oil palm clone P325 of different planting trials. J. Oil Palm Res. 28 (4), 431–441.

Tahir, N.I., Shaari, K., Abas, F., Parveez, G.K.A., Tarmizi, A.H., Ramli, U.S., 2013. Identification of oil palm (*Elaeis guineensis*) spear leaf metabolites using mass spectrometry and neutral loss analysis. J. Oil Palm Res. 25 (1), 72–83.

Tahir, N.I., Shaari, K., Abas, F., Parveez, G.K.A., Ishak, Z., Ramli, U.S., 2012. Characterization of apigenin and luteolin derivatives from oil palm (*Elaeis guineensis* Jacq.) leaf using LC-ESI-MS/MS. J. Agric. Food Chem. 60 (45), 11201–11210.

Tan, P.N., Steinbach, M., Kumar, V., 2013. Data mining cluster analysis: basic concepts and algorithms, 2005. In: Introduction to Data Mining, 1. Pearson, London, UK, pp. 487–568. Chapter 8.

Terziyan, V., 2017. Social Distance metric: from coordinates to neighborhoods. Int. J. Geogr. Inf. Sci. 31 (12), 2401–2426.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., et al., 2015. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat. Methods 12 (6), 523–526.

Van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genom. 7 (1), 142.

Viant, M.R., Kurland, I.J., Jones, M.R., Dunn, W.B., 2017. How close are we to complete annotation of metabolomes? Curr. Opin. Chem. Biol. 36, 64–69.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., et al., 2018. Missing value imputation approach for mass spectrometry-based metabolomics data. Sci. Rep. 8 (1), 663.

Worley, B., Powers, R., 2013. Multivariate analysis in metabolomics. Curr. Metabolomics 1 (1), 92–107.
Wu, Z., Li, D., Meng, J., Wang, H., 2010. Introduction to SIMCA-P+ and its application. In: V., Esposito Vinzi, W.W., Chin, J., Henseler, H., Wang (Eds.), Handbook of Partial Least Squares, Springer Handbooks of Computational Statistics. Springer-Verlag, Berlin Heidelberg, pp. 757–774. Chapter 32.
Xia, J., Psychogios, N., Young, N., Wishart, D.S., 2009. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 37 (Suppl.2), W652–W660.
Xia, J., Wishart, D.S., 2011. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. Curr. Protoc. Bioinformatics. 34 (1), 14.
Yang, J., Zhao, X., Lu, X., Lin, X., Xu, G., 2015. A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. Front. Mol. Biosci. 2, 4.
Zhou, B., Xiao, J.F., Tuli, L., Ressom, H.W., 2012. LC-MS-based metabolomics. Mol. Biosyst. 8 (2), 470–481.