Accepted Manuscript

Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of LC-MS data

Anastasiia Shevchuk, Lalith Jayasinghe, Nikolai Kuhnert

PII: S0963-9969(18)30243-6
DOI: doi:10.1016/j.foodres.2018.03.059
Reference: FRIN 7496

To appear in: *Food Research International*

Received date: 28 November 2017
Revised date: 19 March 2018
Accepted date: 21 March 2018

Please cite this article as: Anastasiia Shevchuk, Lalith Jayasinghe, Nikolai Kuhnert , Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of LC-MS data. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Frin(2018), doi:10.1016/j.foodres.2018.03.059

Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of LC-MS data

Anastasiia Shevchuk[1], Lalith Jayasinghe[2] and Nikolai Kuhnert[1]

[1]Jacobs University Bremen, Germany; [2]Institute of Fundamental Studies, Sri Lanka

Corresponding Author:

Nikolai Kuhnert

n.kuhnert@jacobs-university.de

Tel: 0421-200-3120

Abstract

A data set of sixty samples of diverse black tea were collected and analysed using high-performance liquid chromatography-mass spectrometry (HPLC-MS) methods. Chemical variations of black tea infusions depending on origin, botanical variety, and processing were investigated employing various multivariate statistical techniques including principal component analysis (PCA), hierarchical cluster analysis (HCA), partial least squares discriminant analysis (PLS-DA) and analysis of variance (ANOVA).

In particular, PLS-DA allowed identification of a variety of marker compounds responsible for differences among black teas of different origin, plant variety and processing methods used.

Among most variable compounds are catechins, derivatives of quercetin, apigenin, quinic acid, and kaempferol. Rutin, epigallocatechin gallate (EGCG), quinic acid and theaflavin (TF) were contributing to most variances. Products of black tea fermentation (theaflavin, theasinensin, and theacitrin derivatives) contributed to PLS-DA associated to the processing of black tea.

Keywords: Multivariate analysis; Black tea; Polyphenols; Thearubigins; Discrimination; Origin; Fermentation; HPLC-MS

1 Introduction

Tea is second to water the most consumed beverage globally with daily consumption of 500 ml per capita and total crop production of 5,95 Mtons in 2016 (http://faostat.fao.org). 78% of tea is produced as black tea, 20% as green tea and 2% as oolong tea (Yassin, Koek, & Kuhnert, 2015). Black tea is produced by a process termed "fermentation" from the fresh green leaves of either *Camellia sinensis* var. assamica or *Camellia sinensis* var. *sinensis* plant. In tea fermentation, mechanical rupture of the green tea leaves brings tea secondary metabolites, mostly catechins, into contact with oxidase enzymes mainly tea polyphenol oxidase (TPO). Green tea catechins are hereby enzymatically converted into dimeric flavonoids such as theaflavins (TFs), theasinensins, theanaphthoquinines or theacitrins and a heterogeneous fraction referred to by Roberts (Roberts, 1957) as thearubigins (TRs). The chemistry of black tea fermentation has recently been reviewed by Drynan et al. (Drynan, Clifford, Obuchowicz, & Kuhnert, 2010).

Next to the black tea polyphenols, black tea contains proteins, free amino acids, lipids, carbohydrates, fibre, minerals and purine alkaloids, mainly caffeine.

Both frequent consumption and economic importance form the motivation for investigating variations of chemical composition in black tea. Variations in chemical composition are directly related to black tea sensory properties, health benefits, quality, and price (Laddi, Prakash, & Kumar, 2014). These differences may be caused by a variety of factors and have never been addressed comprehensively.

Firstly, the green tea leaf raw material defines the chemical composition of black tea. Green tea leaves may be provided by two distinct plant varieties, with C. *sinensis* var. *assamica* containing higher levels of catechins if compared with *C. sinensis* var. *sinensis*. Variations in green tea leaf composition may be defined by different geographical origin, growth altitude or

climatic conditions. He et al. showed that grade levels and geographical origins of black tea can be distinguished using an electronic tongue (He et al., 2009). Similarly, trace metal composition varies significantly with geographical origin of black teas (Fernández-Cáceres, Martín, Pablos, & González, 2001; Moreda-Piñeiro, Fisher, & Hill, 2003). Chen et al. used Near Infrared spectroscopy (NIR) to demonstrate variations in the caffeine content of black tea depending on the geographical origin (Q. Chen, Zhao, Zhang, & Wang, 2006). Chen et al. reported that EGCG content in the green tea leaves is strongly correlated with growth altitude and season of harvest (Y. Chen et al., 2010).

A second source of variation is defined by agricultural and harvesting practices of green tea raw material, for example, machine harvesting versus manual “two leaves and a bud” harvesting. These differences have to our knowledge never been addressed.

The last origin of variations arises from black tea fermentation itself. Here two procedures are established referred to as CTC (crush, tear and curl) and the orthodox process. Chemical changes in orthodox tea manufacturing have been investigated by Brajesh (Panda & Datta, 2016) and CTC by Amit (Laddi et al., 2014). Sensory variations depending on fermentation times of black tea have been investigated by Ivarsson using cyclic voltammetry based electronic tongue (Ivarsson, Holmin, Höjer, Krantz-Rülcker, & Winquist, 2001). Phenolic content correlated with fermentation conditions, grading and tea prices have been investigated by McDowell et al. (McDowell, Feakes, & Gay, 1991).

In the available literature typically a small selection of analytes present in black tea including catechins, theaflavins, theanine, and caffeine using liquid chromatography coupled with UV/VIS detection has been studied. Volatiles in black tea varieties were determined using gas chromatography techniques (Togari, Kobayashi, & Aishima, 1995). Mineral content was investigated using inductively coupled plasma optical emission spectrometry (ICP-OES)

techniques. Most work uses sum parameters such as total polyphenol content (TPC), measured by the Folin assay or related anti-oxidant techniques (Serpen et al., 2012), electrochemical techniques (Bhattacharyya et al., 2012) or vibrational spectroscopy. The differences between green and black tea have been studied using more information-rich analytical techniques such as nuclear magnetic resonance (NMR) spectroscopy (Fujiwara, Ando, & Arifuku, 2006; Van Dorsten, Daykin, Mulder, & Van Duynhoven, 2006) or liquid chromatography-mass spectrometry (Fujiwara, Ando, & Arifuku, 2006; Van Dorsten, Daykin, Mulder, & Van Duynhoven, 2006). Chemometrics was applied to HPLC-MS data in attempts to discriminate origin, cultivar, and processing (Zheng et al., 2018); geographical location, plantation elevation and leaf grade (Zhang, Suen, Yang, & Quek, 2018); and grades of black tea (Guo, Long, Meng, Ho, & Zhang, 2018).

From a chemical and analytical point of view, black tea constitutes a special challenge. Unlike many other food materials, however, similar to many processed food materials, (Kuhnert, Dairpoosh, Yassin, Golon, & Jaiswal, 2013) black tea can be classified as an unresolved complex mixture (UCM). Kuhnert showed that within the thearubigin fraction of black tea around 6 000 different analytes can be detected using high-resolution mass spectrometry. Resolution of isomers using liquid chromatography coupled to mass spectrometry (MS) shows an average of six isomers increasing the number of compounds present in a black tea infusion to a minimum of 30 000 (Kuhnert, 2010). These compounds are formed during black tea fermentation through a process named “oxidative cascade reaction”, in which catechin derivatives are oxidized at their B-rings to form ortho-quinones, followed by addition of water to form polyhydroxylated flavonoid derivatives (Kuhnert, Clifford, & Mueller, 2010) or by nucleophilic addition to form oligomeric flavonoids (Yassin, Koek, Jayaraman, & Kuhnert, 2014; Yassin, Koek, & Kuhnert, 2014).

This complexity of black tea requires analytical methods offering an exceptionally high resolution with LC-high-resolution mass spectrometry as the most suitable option. For differentiation of chemical compositions, multivariate statistical techniques using LC-MS data have emerged as popular options, including Principal Component Analysis (PCA), Hierarchical Clustering (HC) or linear regression. In view of the complexity of black tea samples with several thousand analytes detected in each sample, the use of such bioinformatics techniques will pose a challenge. Part of this manuscript addresses this challenge; especially, what type of information can be gathered from multivariate techniques using such complex samples and how should data be treated to allow meaningful conclusions from such complex data sets.

The aim of this contribution is to establish whether black tea infusions from different origin, botanical variety or fermentation method, can be differentiated based on their chemical composition; and if yes, which chemical markers would allow such a differentiation. We decided to carry out the statistical analysis using two datasets. The aim of this division to distinguish variances originating from the secondary metabolites of the green tea leaf itself or from compounds formed during tea fermentation through the action of oxidative processes.

2 Materials and Methods

2.1 Black teas

Sixty black tea samples were analysed by HPLC-MS. Samples were purchased from the countries of origin (Sri Lanka, Kenya, Portugal) and the local supermarket chains (Bremen, Germany). A figure containing all sample details is provided in the supplementary material (**Figure S1, Appendix A**).

These 60 samples came from various origins, including India (Darjeeling, Assam and Southern India), Nepal, Africa (Kenya and, Nigeria), Sri Lanka, China, Myanmar and the Azores Islands

(Portugal, the most northern tea cultivation area globally). Samples were made from different varieties of *C. sinensis* var. *sinensis* plant and undergone different processing (CTC and orthodox). Finally, the sample set contained five commercial tea blends that were included in the analysis to assess how these samples were grouped with respect to the origin, plant variety, and processing.

2.2 Reagents

Ultrapure water was obtained from Mili-Q water purification system to provide a resistivity of 18.2 MΩ cm (Millipore, Bedford, USA). Acetonitrile used for HPLC-measurements was LC-MS grade (Acetonitrile for UHPLC Supergradient, AppliChem). Formic acid, used during experiments as a mobile phase modifier and as calibration buffer, was mass-spectrometry grade (Sigma-Aldrich, Germany). Phloretin was purchased from PhytoLab. HPLC syringe filters were purchased from Macherey-Nagel (Germany). Sodium carbonate, Folin-Ciocalteu's phenol reagent, and gallic acid were obtained from Sigma-Aldrich. Hydrochloric acid was obtained from Merck (Germany).

2.3 Infusion preparation

Black tea infusions were prepared simulating home-made cup of tea (Diniz, Barbosa, De Melo Milanez, Pistonesi, & De Araújo, 2016). To 2.0 g of tea powder or leaves, 200 ml of deionized water at 100 °C was added. The infusion was stirred at 95 °C for 5 minutes, left to cool for 5 minutes and 2 ml filtered through a nylon syringe HPLC filter. From this solution, 940 µL were taken, and 60 µL of phloretin internal standard of 0.1 mg/mL was added. The solution was used for LC/MS analysis.

2.4 Preparation of thearubigins

Freshly ground black tea leaves (8 g) were added to 150 mL freshly boiled water and kept for 10 minutes in a Thermos flask, which was inverted every 30 seconds. The flask contents were

filtered through a Whatman no. 4 filter paper to remove the leaves, and the remaining brew allowed to cool to room temperature. Caffeine sufficient to achieve 20 mM was added to the brew, stirred to ensure dissolution, and allowed to stand at 4 °C for two hours, and centrifuged at 23,300 × *g* for 20 min. The resulting precipitate was recovered and suspended in boiling water and partitioned against aliquots of ethyl acetate (40 mL) until no further colour was extracted (usually ×5).

The ethyl acetate-supernatant was removed and evaporated to dryness under nitrogen below 35 °C, and the residue (TF fraction) recovered in 10 ml distilled water. The aqueous phase was partitioned at 80 °C against two volumes of chloroform, the decaffeinated liquid stored overnight at –80 °C, and freeze-dried. The freeze-dried material (TR fraction) was stored at –20 °C until required and reconstituted as required for the analysis. TRs were obtained as orange to light brown fluffy powders. 1 mg of TR was dissolved in 1 ml 70:30 MeOH-Water for LC-MS analysis.

2.5 Folin-Ciocalteau total phenolics assay

Total phenolics assay was performed by modified Folin-Ciocalteu assay procedure (D'Souza et al., 2017)

Black tea infusions as prepared above were 5-fold diluted with deionized water. Into Eppendorf tubes containing 100 µL of deionized water, 30 µL of diluted sample was pipetted (standards and samples). Commercial Folin-Ciocalteu reagent (100 µL) was added to the black tea infusion and the resulting mixture vortexed for 5 s. The solution was allowed to stand for two minutes after which 800 µL of 5% sodium carbonate solution was added, vortexed and incubated for 20 minutes at 40°C. Samples were cooled to RT, and 200 µL was transferred to a

96-well plate. Absorbance was measured at 725 nm on a 96-well Biochrom EZ Read 2000 microplate reader (Cambridge, UK). Total polyphenol values were related to gallic acid equivalents using a gallic acid calibration curve.

2.1 Chromatographic conditions, HPLC-ESI-TOF-MS analysis, and HPLC-ESI-ion trap-MS analysis

Chromatographic separation was performed with an Agilent 1200 HPLC system, consisting of a binary pump, well-plate autosampler with 100 µL loop, column oven and UV-VIS detector (Agilent, Waldbronn, Germany) coupled to the ESI-TOF mass spectrometer (Bruker Daltonics micrOTOF Focus, Bremen, Germany). The mobile phase consisted of two solvents; A was water and B was Acetonitrile. Mobile phase A was modified with 0.005% formic acid. For chromatographic elucidation, a reversed phase packing was used (Varian Amide-C18, 250x3.0 mm, particle size 5.0 µm) at 25 °C column oven temperature. Flow rate was 0.5 mL/min. Elution was achieved by the following gradient: in 50 min 8% B to 31% B, then 31%B for 10 min, followed by washing step at 85% B for 10 min and re-equilibration for 5 min. The volume of injection was 3 µL.

Samples were measured in a negative mode in a scan range 50–1000 *m/z*. Prior to each chromatographic run 0.01 M sodium formate solution was injected through a six-port valve for internal calibration. Calibration was carried out using the enhanced quadratic mode.

HPLC-ESI-ion trap (MS^n) measurements were conducted using HCT Ultra Ion Trap mass spectrometer (Bruker Daltonics, Bremen, Germany) coupled to an Agilent 1100 HPLC system. The HPLC consisted of a binary pump, well-plate autosampler with 100 µL loop, column oven and DAD detector (Agilent, Waldbronn, Germany). For chromatographic separation same water-acetonitrile gradient as before was used. Acquisition of MS^2 and MS^3 mass spectra were performed on the HTC Ion-Trap mass spectrometer in negative ion mode using the AutoMS^n

mode. The column eluent was first directed to the UV detector and then to the ESI interface operating with a capillary voltage of 1 V, and fragmentation amplitude was set starting at 30% and ending at 200%. The capillary temperature was also set at 300 °C. Nitrogen gas was used here as nebulizing and drying gas at a flow rate of 10 L/min and a pressure of 10 psi, respectively.

2.2 Data Analysis

Fifty-two samples were included in the statistical analysis. Raw HPLC-MS data was analysed using Bruker Data analysis 4.0. Tandem-MS data was used for identification of statistically relevant compounds by comparison with fragment spectra from authentic standards, fragment spectra reported in the literature or in silico fragments predicted in online databases (Metlin, ChemSpider). For multivariate statistical analysis using the compound finder routine of Data Analysis 4.0, the most intense 1 200 compounds were exported for further analysis. The total number of compounds identified in an average chromatographic run using a TOF-MS instrument is between 2 000 and 2 600.

Statistical evaluation of the data was performed using ProfileAnalysis (Bruker Daltonics) for Hierarchical cluster analysis (HCA) and online tool Metaboanalyst for other statistical assessment. Unsupervised clustering was calculated without scaling and utilizing Spearman distance method, Minkowski exponent equals 1.5, and complete linkage method (Nunes, Alvarenga, de Souza Sant'Ana, Santos, & Granato, 2015).

The LC-MS data was set up for HCA using a bucketing approach of the raw line data. The LC-MS data was integrated in the range 2.5–60 min and 100.5–1000.5 *m/z* in time- and *m/z*-buckets of 5 min and 1 *m/z*. HCA was performed in the *m/z* range 100-1000 and 500-1000.

MetaboAnalyst (MetaboAnalyst: a web server for metabolomic data analysis and interpretation) is an online tool designed for comprehensive metabolomic data analysis (Xia &

Wishart, 2016). Samples were aligned, deconvoluted and normalized in MzMine software (Pluskal, Castillo, Villar-Briones, & Oresic, 2010).

ANOVA and t-test are parametrical statistical tools that are used for dimensionality reduction of complex data such as LC-MS chromatograms. Black tea LC-MS data is highly complex in terms of variables; therefore, ANOVA was applied to reduce the number of variables. For black tea data sets representing different origin, plant variety, and processing one-way ANOVA or t-test was performed. ANOVA and t-test cutoff p-value was 0.05. The post-hoc method used for ANOVA calculation was Fisher's LSD.

Partial least squares regression method combines features of principal components analysis and multiple regression.

PLS-DA is a targeted metabolomic technique, which allows identification of marker compounds displaying variations in defined groups of samples. Score and loading plots are describing relationship of samples as well as variables that are defining groups separation. Cross-validation and permutation test were performed using 10-fold CV cross-validation method and 2000 permutations using separation distance (B/W) statistics.

Most important features (**MIF**) diagram represent useful outputs of PLS-DA which classifies identified marker compounds according to two categories, VIP (variable importance in projection) and a weighted sum of absolute regression coefficients. VIP is computed for each variable, where the score is a measure of variable's importance in the PLS-DA model. In our study, we used the mean of absolute regression coefficient to access contribution of compounds to all PLS-DA components.

MIF diagrams indicate relative concentrations of the corresponding metabolite in each group under study in the form of a green-red colour coded heatmap, whereas green for low, yellow for medium, and red for high intensity and coefficients of each variable, which indicate

contribution of the compound to PLS-DA variance. Coefficient for each variable is a measure of compound (variable) importance, thus compounds with higher MIF coefficient is contributing more to the variance and discrimination of samples.

3 Results and Discussion

For a comprehensive characterization of black tea, aqueous infusions were prepared under standardized conditions, previously described. Next to the standardized black tea infusion, we prepared for a smaller sample subset a purified SII thearubigin fraction using a caffeine precipitation protocol developed by Roberts (Roberts, Cartwright, & Oldschool, 1957).

In terms of samples, we collected 60 different black tea samples and analysed these by HPLC-ESI-MS in the negative ion mode. Compound assignment was based on high-resolution MS measurements revealing molecular formulae and subsequent tandem MS measurements for further structure elucidation. Compound assignment was based on authentic reference standards, literature tandem MS data or tentative assignment based on mass spectral fragmentation rules (Kuhnert, Drynan, Obuchowicz, Clifford, & Witt, 2010; Kuhnert et al., 2010).

The resulting LC-MS data sets of the black tea infusions and TR fractions were subsequently subjected to multivariate statistical analysis. This type of approach has become very popular recently due to its ability to reduce data dimensionality and hence identify similarities and differences in large data sets comprising a large sample number combined with information-rich liquid chromatography coupled to mass spectrometry (LC-MS) data. Beside LC-MS data we collected a popular sum parameter the TPC value using the Folin–Ciocalteu assay.

After visual inspection of the chromatograms, eight of these sixty samples were excluded from further multivariate analysis. These samples are included in the complete sample table in **Figure S1 (Appendix A)**. Three samples showed a series of chromatographic peaks that are untypical for black tea. For two of these samples, these peaks suggested the presence of herbal adulterations of the black tea products due to the presence of high level of quercetin and luteolin glycosides along with signals corresponding to monoterpenes. One sample showed a very intense signal corresponding to the herbicide glyphosate and was excluded for this reason. Two Chinese black teas showed a very low degree of fermentation characterised by high EGCG levels and the absence of theaflavin monogallates. These teas resembled in their chromatographic profile typical oolong teas. Finally, two black tea samples from Myanmar showed a very interesting chromatographic profile. These two teas showed an extreme degree of fermentation. Indeed, only a typical thearubigin hump could be observed, with no defined peaks being observable in negative ion mode in the chromatograms. This observation suggests that all dimeric and trimeric flavonoids, typical for a black tea chromatogram, were converted into higher oligomers by further oxidation. Only the presence of theanine and caffeine in the positive ion mode identified these samples as black tea. A typical Myanmar black tea chromatogram is shown in **Figure S2, Appendix A**.

Following the acquisition of analytical data, we decided to carry out the statistical analysis using two datasets. This division aims to distinguish variances originating from the secondary metabolites of the green tea leaf itself and compounds formed during tea fermentation through the action of oxidative processes.

Multivariate analysis was thus carried out using data from a *m/z* range of 100-1000 including all components detected in the black tea infusion, including unreacted green tea leaf metabolites. A second analysis was carried out using a mass range of *m/z* 500-1000. This mass

range contains predominantly phenolic compounds produced during fermentation, in particular dimers and oligomers of flavan-3-ols formed in the oxidative cascade reactions and flavonol glucosides.

3.1 Folin-Ciocalteu total phenolics content assay

To account for variations in chemical composition based on the variation of extraction kinetics, defined by particle sizes of the black tea, the total polyphenol content determined by Folin-Ciocalteu assay was used as a guide (Astill, Birch, Dacombe, Humphrey, & Martin, 2001). For 33 black tea samples from Assam, Sri Lanka, Kenya, and Portugal. TPC assay was performed to explore difference among different groups of black tea.

The total polyphenol content of samples of *sinensis* as well as of *assamica* showed similar TPC values. No difference in TPC values was observed in the case of CTC versus Orthodox processed samples. (**Figure S4, Appendix A**). Minor differences in TPC were observed for samples of different origins, where Portugal samples showed lower TPC values if compared to samples from India, Sri Lanka or Kenya. All samples analysed show a TPC value between 0.89 and 4.05 g-GAE/kg.

ANOVA and t-test calculations revealed a non-significant difference between groups, with p-value >0.05. Boxplots and p-values are shown in **Appendix A, Figure S5**.

Consequently, total polyphenol content cannot be used as reliable parameter for discrimination of black teas, due to black tea phenolic diversity encountered among samples combined with the Folin-Ciocalteu reagent.

3.2 Hierarchical cluster analysis

Firstly, we employed Hierarchical cluster analysis (HCA), a multivariate analysis method suitable for identification of similarities between samples. The output of a HCA analysis is a

dendrogram with samples grouped in branches according to their similarity, in this case similarity of LC-MS parameters (retention time (RT), *m/z* value and intensity).

Black tea samples were colour-coded in the HCA plot according to their origin, plant variety (*sinensis* and *assamica*) and processing (CTC vs. Orthodox). Pareto scaling of raw data prior to HCA did not result in a significantly different dendrogram, if compared to unscaled data. Therefore, furtherer calculations were performed without scaling. For *m/z* range 100-1000 samples formed fewer clusters if compared to the *m/z* 500-1000 range, indicating dependence of sample similarities on fermentation degree. Hierarchical cluster analysis, performed on the *m/z* range 500-1000 without scaling and dendrograms are shown in the **Figure 1 A, B**.

Nepal, Darjeeling and Portuguese black teas are merged in a joined cluster 101. The dendrogram shows that samples cluster based on their origin, one main branch containing black teas from Portugal, Nepal, and Darjeeling; the second branch containing black teas originating from Sri Lanka, Assam and Kenya. Darjeeling and Nepalese teas are produced from the same tea cultivar of *sinensis*, which was imported to Nepal from Indian tea-growing region Darjeeling (Ahmed & Stepp, 2013). Thus, similarities of Nepal and Darjeeling tea can be expected, considering a shared history, close geographical location and the same processing method, predominantly Orthodox. Portuguese teas grown on the Azores islands are processed by the Orthodox method, however, using *assamica*, more suitable for their climatic conditions. Nepalese and Darjeeling black teas are cultivated at high altitude and processed resulting in a lower degree of fermentation, possibly due to a soil deficient in copper, essential for TPO activity (Graham, 1992). Degree of fermentation is based on the ratio of flavan-3-ols to dimeric flavan-3-ols.

We suggest that analytical similarities are based on fermentation degree with Kenya, Sri Lanka, and Indian Assam black teas showing a higher degree of fermentation if compared to Nepal, Portugal and Darjeeling samples.

HCA with colour-coded CTC and Orthodox black teas shows clusters for samples processed by the Orthodox method (101, 70, 73, 77) and a second for CTC processing (76, 66, 78, 55, 91). Clusters 74 and 96 indicate that Assam and Sri Lankan black teas show similarities, even though they are produced by different methods. It can be concluded that for these teas plant variety plays a more crucial role defining similarities than processing.

In **figure 1, B** clustering with colour-coding based on plant variety is given. Assam, Sri Lanka and Kenya black teas group together in cluster 98 as they are made from assamica variety of, whereas assamica tea from Portugal, is grouped with *sinensis* Nepal and Darjeeling black teas. Thus, geographical or climatic similarities appear to be more significant than varietal similarity or common processing method. Results of the HCA based on the processing are shown in the **Figure S9 (Appendix A)**.

3.3 PCA of black tea infusions

Secondly, we employed principal component analysis (PCA), an unsupervised technique that identifies variances in a dataset, hence allowing an overview of sample diversity. The output of a PCA is a combination of two plots, a loading plot and a score plot. The score plot shows relationship of samples with distance between samples giving a quantitative value for a variance. The loading plot identifies key variables (*m/z* and Rt pair in the case of LC-MS that allow compound assignment) responsible for variances.

PCA was performed on the data with a *m/z* range from 100 to 1000 Da and for the corresponding data with a *m/z* range 500-1000, covering fermentation products. Explained variances in PC1 and PC2 were low, however, acceptable with around 40-50% explained

variance, in view of the extremely high number of around 1 200 variables selected for the analysis.

Score plots (**Figure 2. A, B**) show good separation of *C. sinensis* var. *assamica* plant from *C. sinensis* var. *sinensis*, whereas samples grouped by origin overlap in case of geographical proximity. Similarly, as in the case of HCA, Nepal and Darjeeling samples overlap. However, Kenya samples overlap with Sri Lanka samples. Samples from Assam and Portugal, made of *C. sinensis* var. *assamica* intersect with two further groups on the score plot.

As expected from the HCA dendrograms, tea samples grouped according to their processing show no clear separation on the score plot (**Figure 2. C**). 3D PCA loading plots using *m/z* range 100-1000 are shown in **Figure S6** and using *m/z* range 500-1000 in **Figure S7** (**Appendix A),** with PC3 not significantly contributing to sample separation**.**

PCA for the fermentation products (*m/z* 500-1000) showed similar separation of samples on the PCA score plots (**Figure S8, Appendix A**). Compound structures, identified from the loading plots from both *m/z* ranges, were assigned based on their MS^n fragment spectra (**Table S1, Appendix A).**

For the most significant compounds identified in the loadings plots, p values of ANOVA and t-test were calculated. Amongst most variable compounds in the *m/z* 500-1000 range are theaflavin-3-gallate, quercetin-3-O-rutinoside, quercetin-3-O-glucoside, gallocatechin, kaempferol 3-O-rutinoside, theaflavin, theasinensin C, Theaflavin as well as quercetin, apigenin and kaempferol derivatives. These compounds constitute key markers of overall sample variation. In case of *m/z* range 100-1000, catechins (EC, ECG, GC, EGC), quinic acid and hydroxycinnamoyl derivatives, quercetin glucoside, and rhamnoside were identified. P values for these compounds are shown in **Table 2**, box plots and structures of the most significant features are shown in **Figure 3**. Quercetin rutinoside and quercetin glucoside were increased

in *assamica* if compared to *sinensis*. Theaflavin 3-O-gallate shows significantly higher concentrations in *assamica* as well as in Ceylon, Assam and Kenya black tea. It is almost absent in the case of *sinensis*, Azores, Darjeeling, and Nepal black teas. Gallocatechin, on the contrary, shows the opposite behavior to theaflavin 3-O-gallate with increased levels found in *sinensis*.

Quinic acid was found to be increased in Assam, Ceylon and Kenya tea, and equally in the *assamica* plant variety.

3.4 PCA analysis of purified SIIa thearubigin extracts

The results described so far used for multivariate statistical analysis LC-MS data from crude black tea infusions. Within tea fermentation, the majority of phenolic biomass is converted by fermentation to a material referred to as thearubigins accounting for 30-60 % of the dry mass of a black tea infusion. To compare our multivariate analysis of crude extracts with actual thearubigins we carried out a PCA analysis of 20 purified TR extracts from two locations, Sri Lanka and Kenya, obtained through our established caffeine precipitation method, characterised by LC-ESI-MS in the negative ion mode. Scores and loading plots are given in **Figure S10, Appendix A**.

In the score plot, a grouping of samples was observed according to the origin of tea. More interestingly the loading plot revealed a contribution of an excess of 300 TR constituents to the explanation of variance. Hence no useful biomarkers could be identified. This is not unexpected since the complexity of TRs is very high with up to 30 000 fermentation products detectable. Secondly, our oxidative cascade hypothesis rationalising the chemistry of tea fermentation assumes an air driven diversification of fermentation products that should be independent of processing parameters or botanical variety.

3.5 ANOVA and t-test

ANOVA calculations were carried out on the full data set including the 1 200 most intense LC-MS signals. These did not result in significant data reduction as hoped for. More than 80% of the variables (and more than 1000 in terms of quantity) were considered significant with a p-value of 0.05 or smaller (**Figure S11, Appendix A**). Thus, ANOVA and t-test have limitations for complex matrices, obscuring marker identification yielding a confusing high number of significant features. These findings agree with observations made in PCA analysis of purified TR fractions. Consequently, PLS-DA supervised technique was tried for identification of key marker compounds allowing discrimination of black teas with pairwise varying attributes. For all compounds that were identified as significant in the PLS-DA plots, ANOVA or t-test p-values were calculated. Cross-validation and permutation test results are represented in **Figure S3 (Appendix A)**.

3.6 PLS-DA of black tea origin

In the case of Origin-related PLS-DA for the *m/z* range 100-1000, separation on the score plot was low with 36% of the explained variance for the Component1/Component2 plot (**Figure 4**). The 3D graph for PLS-DA compounds 1-3 as well as score plots for compound1/compound3 (39%), and Compound1/compound4 (42%) are additionally shown in the **Figure S12 (Appendix A).**

Loadings plots identify compounds responsible for variability between two groups, similar to PCA. However, we decided to present the content of the loading plot in a Most Important Feature Diagram (MIF) with additional colour codes representing increase or decrease of concentration of key marker compounds. Most important compounds (that have more contribution to PLS-DA) were identified and ANOVA was calculated. Identification and

accurate mass, as well as p-values, are represented in **Table 2,** whereas fragmentation details are shown in the **table S1 (Appendix A).**

The MIF diagram for the *m/z* range 100-1000 is shown in **Figure 4**. The most contributing features is EGCG, which is increased in Darjeeling tea and Nepal black teas if compared to Kenya and Assam teas. Azores and Ceylon teas showed in-between values. Epicatechin gallate has a similar pattern, which can be explained by the natural difference in catechins concentration in green tea leaves in various origins.

Among other important features quercetin 3-O-(2,6-di-O-rhamnosyl-hexoside) 7-O-rhamnoside, kaempferol 3-O-glucoside, apigenin 6-C-hexosyl-8-C-pentoside, kaempferol 3-O-rutinoside, quercetin-3-O-glucoside and rutin **(Figure 4)** were identified. These flavonoid glycosides not only contribute to tea beverage bitter taste but can also be classified as fermentation "bystanders", compounds not affected by TPO. These derivatives were mostly decreased in Nepal and Azores teas. Another bystander product identified as origin marker was 3-*p*-coymaroylquinic acid, which is increased in teas from Darjeeling and Assam. Products of fermentation were not present on the MIF diagram for *m/z* range 100-1000.

The same approach was applied for *m/z* range 500-1000 to access important variables resulting from fermentation. Best separation of the sample groups was achieved on the Component1/Component 3 score plot with 54% of the explained variance **(Figure 5).** Most important compounds from the loading plot are represented in **table 3.** The MIF diagram in the range *m/z* 500-1000 has a similar content if compared to the whole range, where rutin and kaempferol 3-O-rutinoside had coefficients above 50, as well as kaempferol, quercetin, and apigenin derivatives. Kaempferol and quercetin rutinosides showed similar profiles amongst different countries, where compound concentration is low in Nepal, Assam and Darjeeling teas, and higher in Azores, Ceylon, and Kenya tea. Boxplots, graphically illustrating

variations in sample classes for the most important compounds are shown in **figure S13**, **Appendix A.**

3.7 PLS-DA analysis of *Camellia sinensis* plant variety used for black tea production

The same routine was applied for PLS-DA of samples differing in plant variety. Almost complete separation between *assamica* and *sinensis* samples on the score plot was achieved on the Compound1/Compound2 graph with explained variance 48,3% (**Figure 6**). Amongst important features that contribute to the separation in the mass range of *m/z* 100-1000 are quercetin derivatives (3-O-rhamnoside and 7-O-glucosides), Quinic acid, kaempferol-3-O glucoside, Theaflavin and catechins (EC, GC, ECg, EGCg). Two quercetin derivatives are increased in C. *sinensis* var. *assamica* plant, and their profile is correlating well with previous PCA results, where these compounds are increased in Ceylon, Assam, and Kenya teas if compared to Darjeeling and Nepal teas. Intensity of theaflavin and quinic acid was higher in *assamica* plant variety, which as well correlates with the origin-wise comparison, where BT samples made of C. *sinensis* var. *assamica* (Assam, Ceylon, and Kenya), showed higher intensity of theaflavin. The most important feature for assamica/sinensis evaluation is EGCG, which is increased in *C. sinensis* var. *sinensis* plants. The same trend was observed for other catechins with Darjeeling and Nepal tea displaying a higher amount of catechins if compared to other origins.

Similarly, best separation for data for the *m/z* 500-100 range was achieved on the Component1/Component2 score plot with 53.8% of the explained variance **(Figure 7**). The content of the MIF diagram is varying considerably from the corresponding full range diagram. For instance, among important features two kaempferol derivatives were identified (3-O-rutinoside and glucosyl-rhamnosyl-galactoside), quercetin rutinoside and apigenin derivative,

as well as products of fermentation, e.g. theaflavin, theaflavin-3-O-gallate, theacitrin A and theasinensin B. Fermentation products were increased in *assamica*, whereas kaempferol 3-O-glucosyl-rhamnosyl-galactoside and compound with *m/z* 555 and Rt 22.4 min (ID 204) tentatively assigned as EGCG derivative were increased in *sinensis* plant variety. **Figure S14, Appendix A** illustrates variations in samples according to the botanical variety with **table 4** showing compound assignments. 3D PLS-Da loading plots are shown in **Figure S16, A and B.**

3.8 PLS-DA analysis of black tea processing method

Untargeted multivariate analysis for samples different in processing did not show significant separation on PCA score plot and thus targeted approach was again applied to reveal key marker compounds.

As in the case of plant variety at *m/z* 100-1000 range, catechins (EgC, EC, ECg) were among most important features (**Figure 8**) and showed increased levels in samples processed by the Orthodox method. Kaempherol-3-O-glucoside showed increased concentration in CTC samples, in contrast to Kaempferol 3-glucosyl-(1->3)-rhamnosyl-(1->6)-galactoside and Quercetin 3-O-(rhamnosyl-(1α→6)-O-glucoside) hexoside being elevated in Orthodox processed samples. Another important feature was a 4-O-caffeoylquinic acid which is increased in CTC samples.

PLS-DA on the data set within 500-1000 Da *m/z* range resulted in a well-separated score plot Compound 1/Compound 2 with 51% of the explained variance (**Figure 9**). Most important features in this *m/z* range were identified, whereas TF and TF gallates, kaempferol and quercetin derivatives were identified. Theaflavins were increased in CTC processed black teas resembles PLS-DA result of assamica plant variety, which agrees with most *assamica* black tea being processed by CTC and *sinensis* is mostly produced by orthodox method respectively.

Hence these two variables are interconnected. Another kaempferol glucoside was increased in CTC processed tea. Compound with *m/z* 555 and Rt 22.4 min (ID 204) was identified as EGCG derivative and unknown compound with *m/z* 580, Rt 19.7 min (ID501) showed increased concentration in Orthodox processed teas.

Tentative identification of compounds and p-values for t-test are shown in **table 5**. Boxplots for the most important compounds are shown in **figure S15, Appendix A.** 3D PLS-Da loading plots are shown **in Figure S16, C and D.**

Structures of all identified compounds are shown on the **Figure S17 (Appendix A)**

Our results can be used for optimisation of desired black tea organoleptic properties, especially colour and taste. Among the most important variables are bitter compounds such as flavonoid glucosides, astringent phenolics such as catechins and theaflavins and their oxidation products (Scharbert, Holzmann, & Hofmann, 2004). Theaflavin content is an important quality parameter in black tea industry with theaflavin defining the desired reddish orange colour of a tea infusion. Our results suggest selection of certain plant varieties and processing methods to increase TFs quantity. These results are complementary to findings of Zhang et al., showing increase of TF content by selection of tea grown at low altitude or from small leaves (Zhang et al., 2018). Similarly, educated choice based on detailed chemical insight of plant variety and processing, would allow for a reduction of undesirable bitterness and astringency.

Alternative processing methods would allow optimization of theasinensin or theacitrin content, although much less detailed information on the chemical and biological properties of these key black tea constituents is available yet.

Commercial samples of tea blends are grouped in our statistical analysis. Hence conclusions on the origin, plant varieties used, and processing method most likely applied can be readily drawn.

4 Conclusion

Application of HPLC-MS together with chemometric methods allows differentiation of commercial black teas according to the origin, botanical variety, and processing method. The use of purified thearubigin fractions, as well as the application of ANOVA techniques, was shown to be unsuitable, leading to uninformative analysis outputs in a complex material such as black tea. In contrast, PLS-DA was demonstrated to cope well with black tea complexity resulting in clear identification of marker compounds. We could as well show that a bias introduced through an educated selection of mass ranges allows for identification of fermentation biomarkers. Biomarkers identified include typical tea processing phenolics as well as flavonoids as "bystander" phenolics in the oxidative cascade process. Among most important variables for PLS-DA Origin we observed derivatives of quercetin, apigenin, kaempferol, catechins (EGCG, EC, GC), quinic acid derivatives (4-CQA, 3-CQA, 3-*p*-CoQA) and fermentation products such as theaflavin, theaflavin-3-O-gallate and theacitrin A.

With respect to botanical variety of *Camellia sinensis*, most important variables were identified as GC and EGCG, kaempferol, quercetin and apigenin glucosides, theaflavin and theaflavin gallate.

PLS-DA analysis discovered more products of black tea fermentation as the most significant features; e.g. theaflavin, theasinensin, theacitrin and their gallates. Among other important variables, kaempferol and quercetin glucosides, EGCG, and kaempferol rutinoside were identified.

Acknowledgments

Authors are thankful to Sabur Badmos for performing TPC assays, Jacobs University Bremen and UNILEVER Bangalore R&D for financial support of these studies.

References

Ahmed, S., & Stepp, J. R. (2013). Chapter 2 - green tea: The plants, processing, manufacturing and production. In V. R. Preedy (Ed.), *Tea in health and disease prevention* (pp. 19-31) Academic Press.http://dx.doi.org/10.1016/B978-0-12-384937-3.00002-1

Astill, C., Birch, M., Dacombe, C., Humphrey, P., & Martin, P. (2001). Factors affecting the caffeine and polyphenol contents of black and green tea infusions. *Journal of Agricultural and Food Chemistry, 49*(11), 5340-5347. 10.1021/jf010759+

Bhattacharyya, R., Tudu, B., Das, S. C., Bhattacharyya, N., Bandyopadhyay, R., & Pramanik, P. (2012). Classification of black tea liquor using cyclic voltammetry. *Journal of Food Engineering, 109*(1), 120-126. 10.1016/j.jfoodeng.2011.09.026

Chen, Q., Zhao, J., Zhang, H., & Wang, X. (2006). Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration. *Analytica Chimica Acta, 572*(1), 77-84. 10.1016/j.aca.2006.05.007

Chen, Y., Jiang, Y., Duan, J., Shi, J., Xue, S., & Kakuda, Y. (2010). Variation in catechin contents in relation to quality of 'huang zhi xiang' oolong tea (camellia sinensis) at various growing altitudes and seasons. *Food Chemistry, 119*(2), 648-652. 10.1016/j.foodchem.2009.07.014

Diniz, P. H. G. D., Barbosa, M. F., De Melo Milanez, K. D. T., Pistonesi, M. F., & De Araújo, M. C. U. (2016). Using UV-vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. *Food Chemistry, 192*, 374-379. 10.1016/j.foodchem.2015.07.022

Drynan, J. W., Clifford, M. N., Obuchowicz, J., & Kuhnert, N. (2010). The chemistry of low molecular weight black tea polyphenols. *Natural Product Reports, 27*(3), 417-462. 10.1039/b912523j

D'Souza, R. N., Grimbs, S., Behrends, B., Bernaert, H., Ullrich, M. S., & Kuhnert, N. (2017). Origin-based polyphenolic fingerprinting of theobroma cacao in unfermented and fermented beans. *Food Research International, 99*, 550-559. 10.1016/j.foodres.2017.06.007

Fernández-Cáceres, P. L., Martín, M. J., Pablos, F., & González, A. G. (2001). Differentiation of tea (camellia sinensis) varieties and their geographical origin according to their metal content. *Journal of Agricultural and Food Chemistry, 49*(10), 4775-4779. 10.1021/jf0106143

Fujiwara, M., Ando, I., & Arifuku, K. (2006). Multivariate analysis for1H-NMR spectra of two hundred kinds of tea in the world. *Analytical Sciences, 22*(10), 1307-1314. 10.2116/analsci.22.1307

Graham, H. N. (1992). Green tea composition, consumption, and polyphenol chemistry. *Preventive Medicine, 21*(3), 334-350. 10.1016/0091-7435(92)90041-F

Guo, X., Long, P., Meng, Q., Ho, C. -., & Zhang, L. (2018). An emerging strategy for evaluating the grades of keemun black tea by combinatory liquid chromatography-orbitrap mass spectrometry-based untargeted metabolomics and inhibition effects on a-glucosidase and a-amylase. *Food Chemistry, 246*, 74-81. 10.1016/j.foodchem.2017.10.148

He, W., Hu, X., Zhao, L., Liao, X., Zhang, Y., Zhang, M., & Wu, J. (2009). Evaluation of chinese tea by the electronic tongue: Correlation with sensory properties and classification

according to geographical origin and grade level. *Food Research International, 42*(10), 1462-1467. 10.1016/j.foodres.2009.08.008

http://faostat.fao.org. Faostat.

Ivarsson, P., Holmin, S., Höjer, N. -., Krantz-Rülcker, C., & Winquist, F. (2001). Discrimination of tea by means of a voltammetric electronic tongue and different applied waveforms. *Sensors and Actuators, B: Chemical, 76*(1-3), 449-454. 10.1016/S0925-4005(01)00583-4

Kuhnert, N. (2010). Unraveling the structure of the black tea thearubigins. *Archives of Biochemistry and Biophysics, 501*(1), 37-51. 10.1016/j.abb.2010.04.013

Kuhnert, N., Clifford, M. N., & Mueller, A. (2010). Oxidative cascade reactions yielding polyhydroxy-theaflavins and theacitrins in the formation of black tea thearubigins: Evidence by tandem LC-MS. *Food & Function, 1*(2), 180-199. 10.1039/c0fo00066c

Kuhnert, N., Dairpoosh, F., Yassin, G., Golon, A., & Jaiswal, R. (2013). What is under the hump? mass spectrometry based analysis of complex mixtures in processed food - lessons from the characterisation of black tea thearubigins, coffee melanoidines and caramel. *Food & Function, 4*(8), 1130-1147. 10.1039/C3FO30385C Retrieved from http://dx.doi.org/10.1039/C3FO30385C

Kuhnert, N., Drynan, J. W., Obuchowicz, J., Clifford, M. N., & Witt, M. (2010). Mass spectrometric characterization of black tea thearubigins leading to an oxidative cascade hypothesis for thearubigin formation. *Rapid Communications in Mass Spectrometry, 24*(23), 3387-3404. 10.1002/rcm.4778

Laddi, A., Prakash, N. R., & Kumar, A. (2014). Quality evaluation of black CTC teas based upon seasonal variations. *International Journal of Food Science and Technology, 49*(2), 493-500. 10.1111/ijfs.12327

McDowell, I., Feakes, J., & Gay, C. (1991). Phenolic composition of black tea liquors as a means of predicting price and country of origin. *Journal of the Science of Food and Agriculture, 55*(4), 627-641. 10.1002/jsfa.2740550414

Moreda-Piñeiro, A., Fisher, A., & Hill, S. J. (2003). The classification of tea according to region of origin using pattern recognition techniques and trace metal data. *Journal of Food Composition and Analysis, 16*(2), 195-211. 10.1016/S0889-1575(02)00163-1

Nunes, C. A., Alvarenga, V. O., de Souza Sant'Ana, A., Santos, J. S., & Granato, D. (2015). The use of statistical software in food science and technology: Advantages, limitations and misuses. *Food Research International, 75*, 270-280. 10.1016/j.foodres.2015.06.011

Panda, B. K., & Datta, A. K. (2016). Quantitative analysis of major phytochemicals in orthodox tea (camellia sinensis), oxidized under compressed air environment. *Journal of Food Science, 81*(4), C858-C866. 10.1111/1750-3841.13265

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *Bmc Bioinformatics, 11*, 395. 10.1186/1471-2105-11-395

Roberts, E. A. H. (1957). Oxidative condensation of flavanols in tea fermentation. *Chemistry & Industry, 41*, 1355-1356.

Roberts, E. A. H., Cartwright, R. A., & Oldschool, M. (1957). The phenolic substances of manufactured tea. I.?fractionation and paper chromatography of water-soluble substances. *Journal of the Science of Food and Agriculture, 8*(2), 72-80. 10.1002/jsfa.2740080203

Scharbert, S., Holzmann, N., & Hofmann, T. (2004). Identification of the astringent taste compounds in black tea infusions by combining instrumental analysis and human bioresponse. *Journal of Agricultural and Food Chemistry, 52*(11), 3498-3508. 10.1021/jf049802u

Serpen, A., Pelvan, E., Alasalvar, C., Mogol, B. A., Yavuz, H. T., Gökmen, V., . . . Özçelik, B. (2012). Nutritional and functional characteristics of seven grades of black tea produced in turkey. *Journal of Agricultural and Food Chemistry, 60*(31), 7682-7689. 10.1021/jf302058d

Togari, N., Kobayashi, A., & Aishima, T. (1995). Pattern recognition applied to gas chromatographic profiles of volatile components in three tea categories. *Food Research International, 28*(5), 495-502. 10.1016/0963-9969(95)00029-1

Van Dorsten, F. A., Daykin, C. A., Mulder, T. P. J., & Van Duynhoven, J. P. M. (2006). Metabonomics approach to determine metabolic differences between green tea and black tea consumption. *Journal of Agricultural and Food Chemistry, 54*(18), 6929-6938. 10.1021/jf061016x

Xia, J., & Wishart, D. S. (2016). Using metaboanalyst 3.0 for comprehensive metabolomics data analysis. *Current Protocols in Bioinformatics, 2016*, 14.10.1-14.10.91. 10.1002/cpbi.11

Yassin, G. H., Koek, J. H., Jayaraman, S., & Kuhnert, N. (2014). Identification of novel homologous series of polyhydroxylated theasinensins and theanaphthoquinones in the SII fraction of black tea thearubigins using ESI/HPLC tandem mass spectrometry. *Journal of Agricultural and Food Chemistry, 62*(40), 9848-9859. 10.1021/jf502220c

Yassin, G. H., Koek, J. H., & Kuhnert, N. (2014). Identification of trimeric and tetrameric flavan-3-ol derivatives in the SII black tea thearubigin fraction of black tea using ESI-tandem and MALDI-TOF mass spectrometry. *Food Research International, 63, Part C*(0), 317-327. http://dx.doi.org/10.1016/j.foodres.2014.04.010

Yassin, G. H., Koek, J. H., & Kuhnert, N. (2015). Model system-based mechanistic studies of black tea thearubigin formation. *Food Chemistry, 180*, 272-279. 10.1016/j.foodchem.2015.01.108

Zhang, C., Suen, C. L. -., Yang, C., & Quek, S. Y. (2018). Antioxidant capacity and major polyphenol composition of teas as affected by geographical location, plantation elevation and leaf grade. *Food Chemistry, 244*, 109-119. 10.1016/j.foodchem.2017.09.126

Zheng, X. -., Nie, Y., Gao, Y., Huang, B., Ye, J. -., Lu, J. -., & Liang, Y. -. (2018). Screening the cultivar and processing factors based on the flavonoid profiles of dry teas using principal component analysis. *Journal of Food Composition and Analysis, 67*, 29-37. 10.1016/j.jfca.2017.12.016

Figure 1. Hierarchical cluster analysis using *m/z* range 500-1000. A – origin, B – plant variety (*Assamica* red, *Sinensis* green)

Figure 2. PCA using *m/z* range 100-1000. Pareto scaling, no transformation. Colour-coded according to A – Origin; B – plant variety; C – Processing; D – Loadings plot

Figure 3. Box-and-whisker plots of the most important variables (A) and their structures (B)

Figure 4. PLS-DA Origin using *m/z* range 100-1000. A – score plot; B – loading plot; C – MIF diagram

Figure 5. PLS-DA Origin using *m/z* range 500-1000. A – score plot; B – loading plot; C – MIF diagram

Figure 6. PLS-DA Plant variety using *m/z* range 100-1000. A – score plot; B – loading plot; C – MIF diagram

Figure 7. PLS-DA Plant variety using *m/z* range 500-1000. A – score plot; B – loading plot; C – MIF diagram

Figure 8. PLS-DA Processing using *m/z* range 100-1000. A – score plot; B – loading plot; C – MIF diagram

Figure 9. PLS-DA Processing using *m/z* range 500-1000. A – score plot; B – loading plot; C – MIF diagram

Table 1. Samples details

Origin	Plant variety	Processing	Number of samples
India, region Assam	*Camelia sinensis* var. *assamica*	CTC	10
India, region Darjeeling	*Camelia sinensis* var. *sinensis*	CTC	10
Nepal	*Camelia sinensis* var. *sinensis*	Orthodox	4
Sri Lanka	*Camelia sinensis* var. *assamica*	Orthodox	14
Sri Lanka	*Camelia sinensis* var. *assamica*	CTC	4
Portugal	*Camelia sinensis* var. *assamica*	Orthodox	4
Kenya&Nigeria	*Camelia sinensis* var. *assamica*	CTC	6

Table 2. PCA 100-1000 *m/z* and 500-1000. Origin, plant variety, processing

ID	*m/z* theoretical	Rt, min	Identification	*m/z* measured $[M-H]^-$	Mol, formula	Error,ppm	P value Origin (ANOVA)	P-value Plant Variety (t-test)	P value Processing (t-test)	100-1000 or 500-1000 *m/z*
699	463.0882	33.7	Quercetin 3-O-glucoside	463.0858	$C_{21}H_{20}O_{12}$	4.7	2.75E-09	8.39E-09	-	+/-
4	609.1426	31.8	Rutin	609.1496	$C_{27}H_{30}O_{16}$	0.1	1.46E-15	9.17E-06	-	+/+
6	447.0933	38.3	Kaempferol 3-O-glucoside	447.0933	$C_{21}H_{20}O_{11}$	0.0	3.38E-11	3.72E-05	-	+/-
15	533.172	2.9	Quinic acid dihexoside	533.1731	$C_{19}H_{34}O_{17}$	-1.5	7.08E-07	-	-	+/+
74	479.0814	27.7	Myricetin 3-O-hexoside	479.0835	$C_{21}H_{20}O_{13}$	-4.4	0.012676	-	0.0061626	+/-
937	289.0718	19.7	Epicatechin	289.0711	$C_{15}H_{14}O_6$	2.2	0.042416	0.00088534	0.0016862	+/-
648	441.0827	31.2	Epicatechin gallate	441.0833	$C_{22}H_{18}O_{10}$	-1.4	>0.05	-	-	+/-
794	457.0776	22.4	Epigallocatechin gallate	457.0799	$C_{21}H_{20}O_{11}$	-4.9	0.023646	0.028763	-	+/-
53	305.0667	8.3	Gallocatechin	305.0673	$C_{15}H_{14}O_7$	-2.0	0.00026657	1.13E-05	0.016166	+/-
56	493.0538	22.4	Epigallocatechin gallate dihydrate	493.0550	-	2.0	1.62E-08	2.29E-05	0.0082866	+/-
19	305.0725	12.7	EGC	305.0713	$C_{15}H_{14}O_7$	4.0	-	-	-	+/-

2	343.0671	4.5	5-galloylquinic acid	343.0666	$C_{14}H_{16}O_{10}$	1.3	0.037877	-	-	+/-
1	191.0561	2.9	Quinic acid	191.0590	$C_7H_{12}O_6$	-1.5	3.48E-07	4.13E-08	0.013754	+/-
61	715.1363	50.9	Theaflavin 3-O-gallate	715.1343	$C_{29}H_{32}O_{21}$	2.8	4.62E-05	1.19E-13	1.15E-05	+/+
43	563.1406	23.7	Apigenin 6-C-pentosyl-8-C-hexoside	563.1385	$C_{26}H_{28}O_{14}$	3.7	1.99E-08	0.00035537	-	+/+
5	593.1512	36.3	Kaempferol 3-O-rutinoside	593.1506	$C_{27}H_{30}O_{15}$	1.0	2.81E-13	0.0032481	-	+/+
259	577.1351	18	(Epi)catechin-4,8'-(epi)catechin	577.1330	$C_{30}H_{26}O_{12}$	3.7	-	-	-	-/+
39	609.1250	4.7	Theasinensin C	609.1256	$C_{30}H_{26}O_{14}$	-1.1	1.48E-05	6.67E-06	-	-/+
33	563.1195	47.7	Theaflavin	563.1171	$C_{29}H_{24}O_{12}$	4.3	1.15E-10	6.62E-13	0.00076171	+/+
13	387.1297	11.2	unknown	387.1268	$C_{17}H_{24}O_{10}$	7.4	1.34E-10	3.94E-09	7.97E-09	+/-
14	337.0929	12.4	3-pCoQA	337.0917	$C_{16}H_{18}O_8$	3.6	3.22E-09	-	-	+/-
204	555.0442	22.4	EGCG derivative	555.0462	$C_{18}H_{20}O_{20}$	2.3	1.69E-07	0.00011088	0.0011014	-/+
231	539.0491	31.2	unknown	539.0501	$C_{25}H_{16}O_{14}$	4.7	1.67E-05	0.0015625	0.003751	-/+
612	507.0728	28.5	unknown	507.0686	$C_{29}H_{16}O_9$	6.2	-	4.43E-07	1.40E-08	-/+

Table 2. Identification of compounds. PLS-DA Origin

ID	*m/z* theoretical	Rt, min	Identification	*m/z* measured	Mol. Formula, [M]	Error, ppm	P-value, ANOVA	*m/z* 100-1000 or 500-1000
6	447.0933	38.3	Kaempferol 3-O-glucoside	447.0933	$C_{21}H_{20}O_{11}$	0.0	3.38E-11	+/-
699	463.0882	33.7	Quercetin 3-O-glucoside	463.0858	$C_{21}H_{20}O_{12}$	4.7	2.75E-09	+/-
37	447.0933	36.8	Kaempferol 7-O-glucoside	447.0930	$C_{21}H_{20}O_{11}$	0.6	1.71E-05	+/-
1	191.0561	2.9	Quinic acid	191.0615	$C_7H_{12}O_6$	-1.5	3.48E-07	+/-
4	609.1461	31.8	Rutin	609.1492	$C_{27}H_{30}O_{16}$	-5.0	1.46E-15	+/+
5	593.1512	36.3	Kaempferol 3-O-rutinoside	593.1543	$C_{27}H_{30}O_{15}$	6.7	2.81E-13	+/+
794	457.0776	22.4	Epigallocatechin gallate	457.0799	$C_{21}H_{20}O_{11}$	-4.9	2.36E-02	+/-
648	441.0827	31.2	Epicatechin gallate	441.0833	$C_{22}H_{18}O_{10}$	-1.4	-	+/-
937	289.0718	19.7	Epicatechin	289.0711	$C_{15}H_{14}O_6$	2.2	4.24E-02	+/-
17	353.0878	13.9	4-O-caffeoylquinic acid	353.0878	$C_{16}H_{18}O_9$	0.0	2.72E-02	+/-
68	353.0878	8.8	3-O-caffeoylquinic acid	353.0874	$C_{16}H_{18}O_9$	1.3	4.80E-06	+/-
111	316.0330	16.3	HHDP-galloyl-glucose	316.0336	$C_{27}H_{22}O_{18}$	-1.8	9.74E-11	+/-
259	577.1363	18	(Epi)catechin-4,8'-(epi)catechin	577.1344	$C_{30}H_{26}O_{12}$	3.3	-	+/-

74	479.0814	27.7	Myricetin 3-O-hexoside	479.0835	$C_{21}H_{20}O_{13}$	-4.4	1.27E-02	+/-
55	289.0718	15	Catechin	289.0714	$C_{15}H_{14}O_6$	1.3	-	+/-
58	477.0592	31.2	Epicatechin gallate dihydrate	477.0602	-	-	5.71E-06	+/-
475	471.0933	28.5	Epigallocatechin-3-(3'-O-methyl) gallate	471.0931	$C_{23}H_{20}O_{11}$	0.4	-	+/-
53	305.0667	8.3	Gallocatechin	305.0664	$C_{15}H_{14}O_7$	0.8	2.67E-04	+/-
14	337.0929	12.4	3-*p*-coumaroylquinic acid	337.0930	$C_{16}H_{18}O_8$	-0.3	3.22E-09	+/-
33	563.1195	47.7	Theaflavin	563.1204	$C_{29}H_{24}O_{12}$	-1.6	1.15E-10	-/+
61	715.1363	50.9	Theaflavin 3-O-gallate	715.1343	$C_{29}H_{32}O_{21}$	2.8	4.62E-05	-/+
84	759.1203	12.9	Theacitrin A	759.1227	$C_{37}H_{28}O_{18}$	-3.2	4.82E-12	-/+
15	533.1723	2.9	Quinic acid dihexoside	533.1731	$C_{19}H_{34}O_{17}$	-1.4	7.08E-07	-/+
96	755.2040	31.5	Quercetin 3-O-(6-O-rhamnosylglucosyl)-7-O-rhamnoside	755.2069	$C_{33}H_{40}O_{20}$	-3.8	1.97E-10	-/+
242	755.2099	33.9	Kaempferol 3-hexosyl-rhamnosyl-hexoside	755.2090	$C_{26}H_{44}O_{25}$	1.1	4.62E-08	-/+
25	593.1512	27.5	Apigenin 6,8-di-C-hexoside	593.1521	$C_{27}H_{30}O_{15}$	-1.5	6.13E-03	-/+
28	450.1163	49	Quercetin 3-O-(2,6-di-O-rhamnosyl-hexoside) 7-O-rhamnoside	450.1197	$C_{42}H_{46}O_{22}$	-6.5	4.39E-10	+/-
13	387.1297	11.2	unknown	387.1292	$C_{17}H_{24}O_{10}$	1.1	1.34E-10	+/-
204	555.0442	22.4	EGCG derivative	555.0462	$C_{18}H_{20}O_{20}$	2.3	1.69E-07	+/+

Table 3. Identification of compounds. Plant variety PLS-DA.

ID	*m/z*, theoretical	Rt, min	Identification	*m/z* measured	Mol. Formula	Error, ppm	P-value, t-test	*m/z* 100-1000 or 500-1000
19	305.0725	12.7	EGC	305.0713	$C_{15}H_{14}O_7$	4.0	-	+/-
2	343.0671	4.5	5-galloylquinic acid	343.0666	$C_{14}H_{16}O_{10}$	1.3	-	+/-
242	755.2040	33.9	Kaempferol 3-glucosyl-rhamnosyl-hexoside	755.2032	$C_{33}H_{40}O_{20}$	1.1	8.82E-10	+/+
53	305.0667	8.3	Gallocatechin	305.0673	$C_{15}H_{14}O_7$	-2.0	1.61E-05	+/-
475	471.0992	28.5	Epigallocatechin 3-(3'-O-methyl) gallate	471.0970	$C_{32}H_{46}O_{32}$	4.5	4.88E-03	+/-
43	563.1406	23.7	Apigenin 6-C-pentosyl-8-C-hexoside	563.1494	$C_{26}H_{28}O_{14}$	2.1	0.00073367	-/+
55	289.0718	15	Catechin	289.0714	$C_{15}H_{14}O_6$	1.3	-	+/-
74	479.0814	27.7	Myricetin 3-O-hexoside	479.0835	$C_{21}H_{20}O_{13}$	-4.4	-	+/-
937	289.0718	19.7	Epicatechin	289.0711	$C_{15}H_{14}O_6$	2.2	-	+/-
648	441.0827	31.2	Epicatechin gallate	441.0833	$C_{22}H_{18}O_{10}$	-1.4	-	+/-
794	457.0776	22.4	Epigallocatechin gallate	457.0799	$C_{21}H_{20}O_{11}$	-4.9	3.36E-03	+/-
699	463.0882	33.7	Quercetin 3-O-glucoside	463.0893	$C_{21}H_{20}O_{12}$	-2.4	7.78E-08	+/-
6	447.0933	38.3	Kaempferol 3-O-glucoside	447.0933	$C_{21}H_{20}O_{11}$	0.0	8.09E-06	+/-
4	609.1461	31.8	Rutin	609.1492	$C_{27}H_{30}O_{16}$	-5.0	6.99E-06	+/+
1	191.0561	2.9	Quinic acid	191.0615	$C_7H_{12}O_6$	-1.5	3.30E-08	+/-
17	353.0878	13.9	4-O-caffeoylquinic acid	353.0893	$C_{16}H_{18}O_9$	-4.1	-	+/-
15	533.1723	2.9	Quinic acid dihexoside	533.172	$C_{19}H_{34}O_{17}$	0.6	-	+/+
5	593.1512	36.3	Kaempferol 3-O-rutinoside	593.1508	$C_{27}H_{30}O_{15}$	0.7	0.0024864	-/+
262	771.1989	30.3	Quercetin 3-O-glucosylrutinoside	771.1999	$C_{33}H_{40}O_{21}$	-1.2	0.00010161	-/+
88	593.1512	34.1	Kaempferol 3-O-rutinoside 7-O-rhamnoside	593.1513	$C_{27}H_{30}O_{15}$	0.2	-	-/+
33	563.1195	47.7	Theaflavin	563.1192	$C_{29}H_{24}O_{12}$	2.6	1.15E-10	-/+
61	715.1363	50.9	Theaflavin 3-O-gallate	715.1343	$C_{29}H_{32}O_{21}$	2.8	2.6308E-07	-/+

84	759.1203	12.9	Theacitrin A	759.1227	$C_{37}H_{28}O_{18}$	-3.2	-	-/+
40	593.1512	20.1	Apigenin 6,8-di-C-hexoside	593.1503	$C_{27}H_{30}O_{15}$	1.4	1.7423E-05	-/+
156	633.0733	16.3	HHDP-galloyl-glucose	633.072	$C_{27}H_{22}O_{18}$	2.1	0.01016	-/+
56	493.0538	22.4	Epigallocatechin gallate dihydrate	493.0548	-	-	3.41E-07	+/-
57	533.1301	28.4	Apigenin 6,8-di-arabinosyl	533.1259	$C_{25}H_{26}O_{13}$	0.2	0.00021055	-/+
589	501.0675	20.7	unknown	501.0670	$C_{23}H_{18}O_{13}$	1.0	2.9913E-11	-/+
204	555.0442	22.4	EGCG derivative	555.0462	$C_{18}H_{20}O_{20}$	2.3	1.0253E-07	-/+

Table 4. Identification of compounds. PLS-DA Processing.

ID	*m/z, theoretical*	Rt, min	Identification	*m/z* measured	Mol. Formula	Error, ppm	P-value, t-test	*m/z* 100-1000 or 500-1000
2	343.0671	4.5	5-galloylquinic acid	343.0666	$C_{14}H_{16}O_{10}$	1.3	-	+/-
5	593.1512	36.3	Kaempferol 3-O-rutinoside	593.1508	$C_{27}H_{30}O_{15}$	0.7	0.002486	+/+
4	609.1461	31.8	Rutin	609.1492	$C_{27}H_{30}O_{16}$	-5.0	6.99E-06	+/+
699	463.0882	33.7	Quercetin 3-O-glucoside	463.0893	$C_{21}H_{20}O_{12}$	-2.4	7.78E-08	+/-
1	191.0561	2.9	Quinic acid	191.0615	$C_{7}H_{12}O_{6}$	-1.5	3.3E-08	+/-
6	447.0933	38.3	Kaempferol 3-O-glucoside	447.0933	$C_{21}H_{20}O_{11}$	0.0	8.09E-06	+/-
33	563.1195	47.7	Theaflavin	563.1192	$C_{29}H_{24}O_{12}$	2.6	9.92E-13	+/+
475	471.0933	28.5	Epigallocatechin-3-(3'-O-methyl) gallate	471.0931	$C_{23}H_{20}O_{11}$	0.4	0.004876	+/-
937	289.0718	19.7	Epicatechin	289.0711	$C_{15}H_{14}O_{6}$	2.2	-	+/-
55	289.0718	15	Catechin	289.0714	$C_{15}H_{14}O_{6}$	1.3	-	+/-
56	493.0933	22.4	Epigallocatechin gallate dihydrate	493.0548	-	-	3.41E-07	+/-
648	441.0827	31.2	Epicatechin gallate	441.0833	$C_{22}H_{18}O_{10}$	-1.4	-	+/-
794	457.0776	22.4	Epigallocatechin gallate	457.0799	$C_{21}H_{20}O_{11}$	-4.9	0.003362	+/-
43	563.1406	23.7	Apigenin 6-C-pentosyl-8-C-hexoside	563.1494	$C_{26}H_{28}O_{14}$	2.1	-	-/+
15	533.1723	2.9	Quinic acid dihexoside	533.1731	$C_{19}H_{34}O_{17}$	-1.4	-	-/+
61	715.1363	50.9	Theaflavin 3-O-gallate	715.1343	$C_{29}H_{32}O_{21}$	2.8	1.4E-05	-/+
242	755.2040	33.9	Kaempferol 3-hexosyl-rhamnosyl-hexose	755.2032	$C_{33}H_{40}O_{20}$	1.1	-	-/+
262	771.1989	30.3	Quercetin 3-O-glucosylrutinoside	771.1999	$C_{33}H_{40}O_{21}$	-1.2	-	-/+
84	759.1203	12.9	Theacitrin A	759.1227	$C_{37}H_{28}O_{18}$	-3.2	0.000453	-/+
91	761.1359	13.7	Theasinensin B	761.1322	$C_{37}H_{30}O_{18}$	4.9	0.006044	-/+
17	353.0878	13.9	4-O-caffeoylquinic acid	353.0893	$C_{16}H_{18}O_{9}$	-4.1		
28	450.1163	49	Quercetin 3-O-(2,6-di-O-	450.1197	$C_{42}H_{46}O_{22}$	-6.5	0.004574	+/-

			rhamnosyl-hexoside) 7-O-rhamnoside					
79	739.2091	32.8	Kaempferol 3-O-(6-O-rhamnosyl-hexoside)7-O-rhamnoside	739.2076	$C_{33}H_{40}O_{19}$	2.1	7.28E-05	-/+
501	580.1508	19.7	unknown	579.1476	$C_{30}H_{28}O_{12}$	5.6	0.010148	-/+
13	387.1297	11.2	unknown	387.1292	$C_{17}H_{24}O_{10}$	1.1	3.94E-09	+/-
204	555.0442	22.4	EGCG derivative	555.0462	$C_{18}H_{20}O_{20}$	2.3	0.000592	-/+

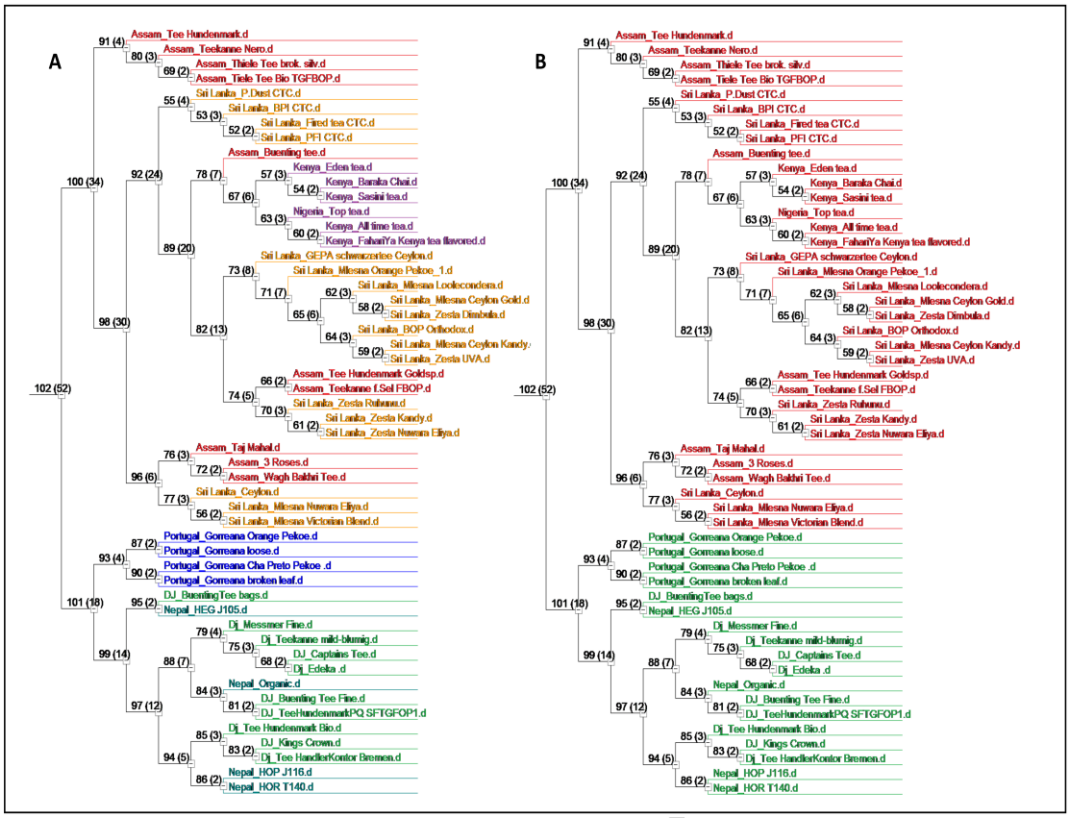


fig. 1

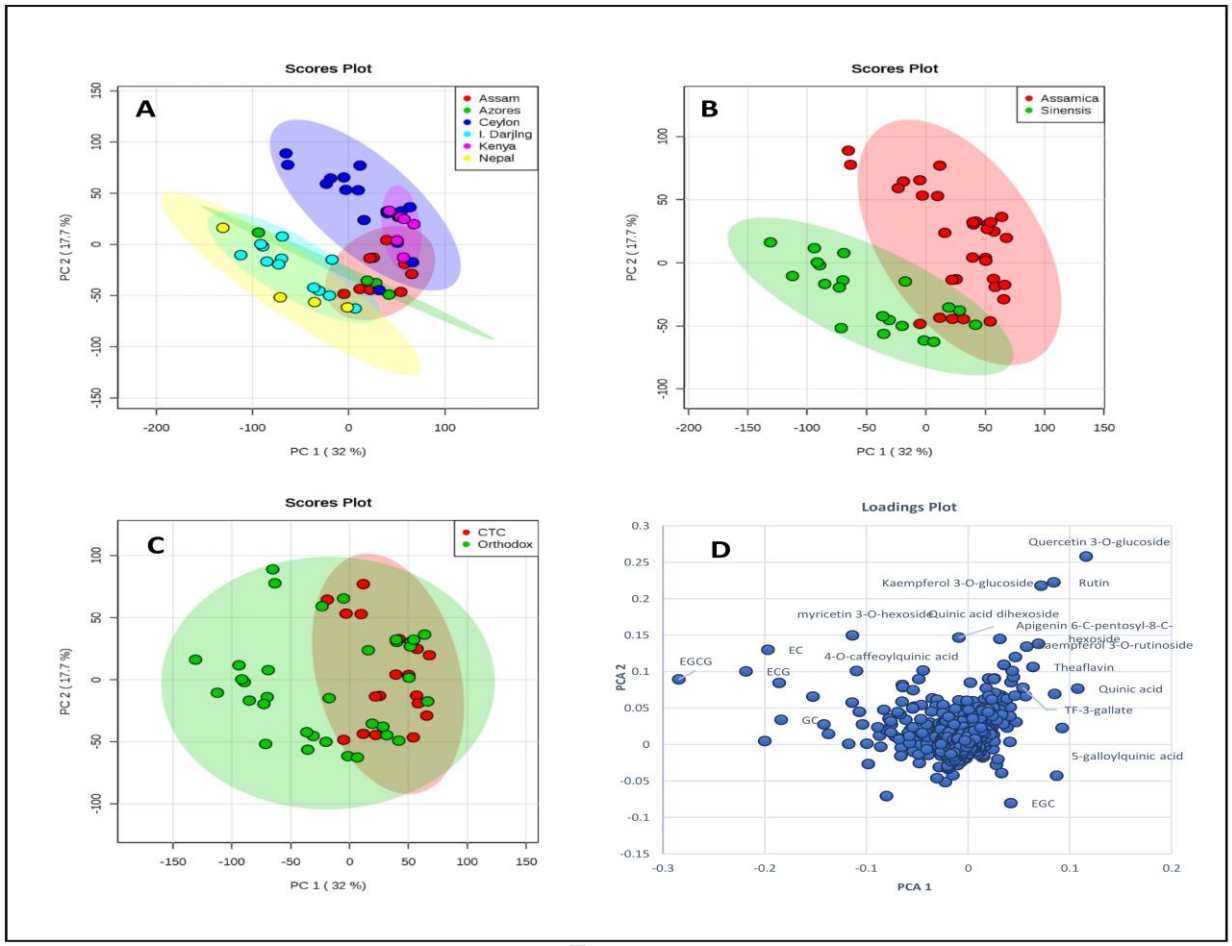


Fig. 2

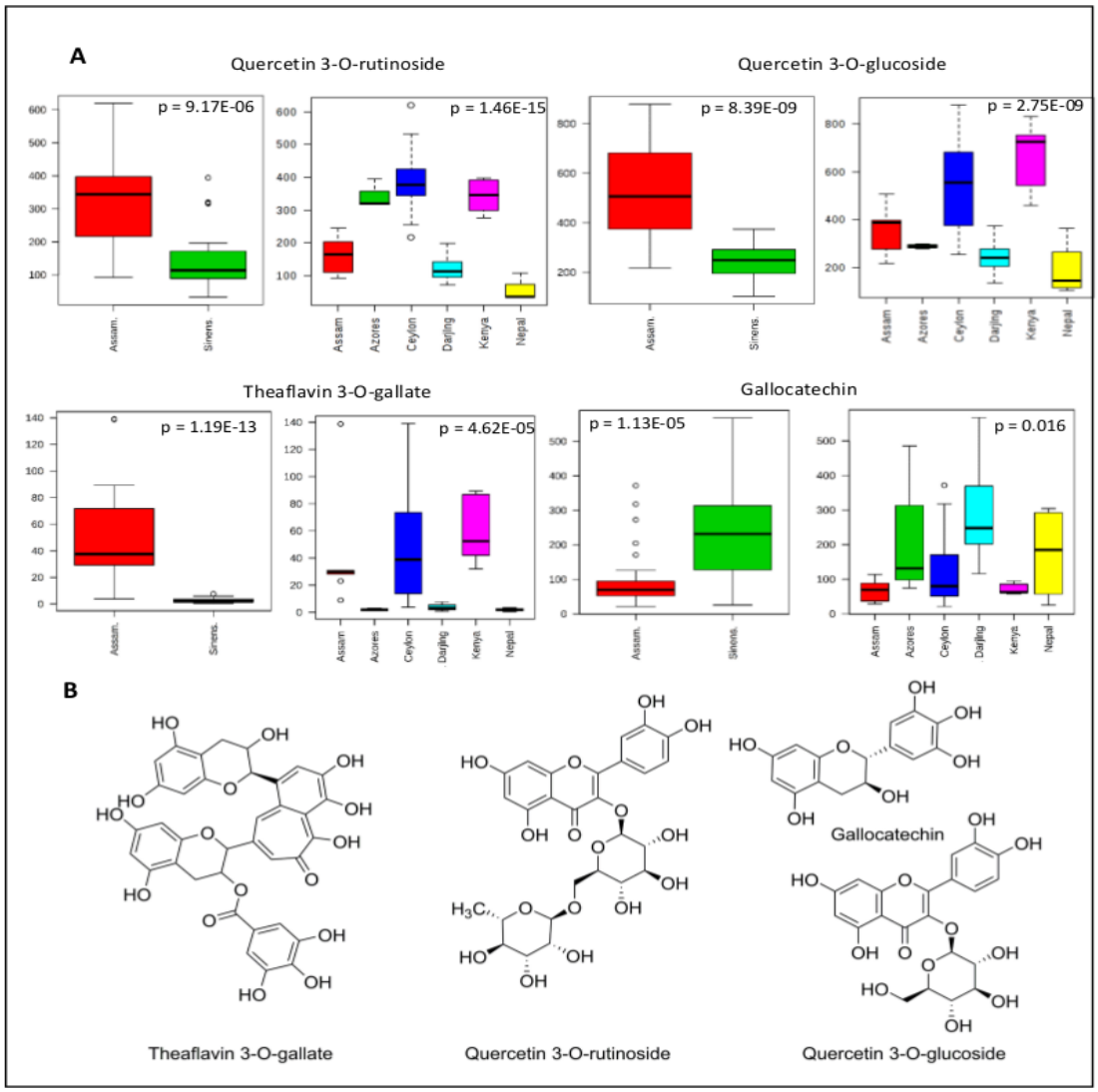


fig. 3

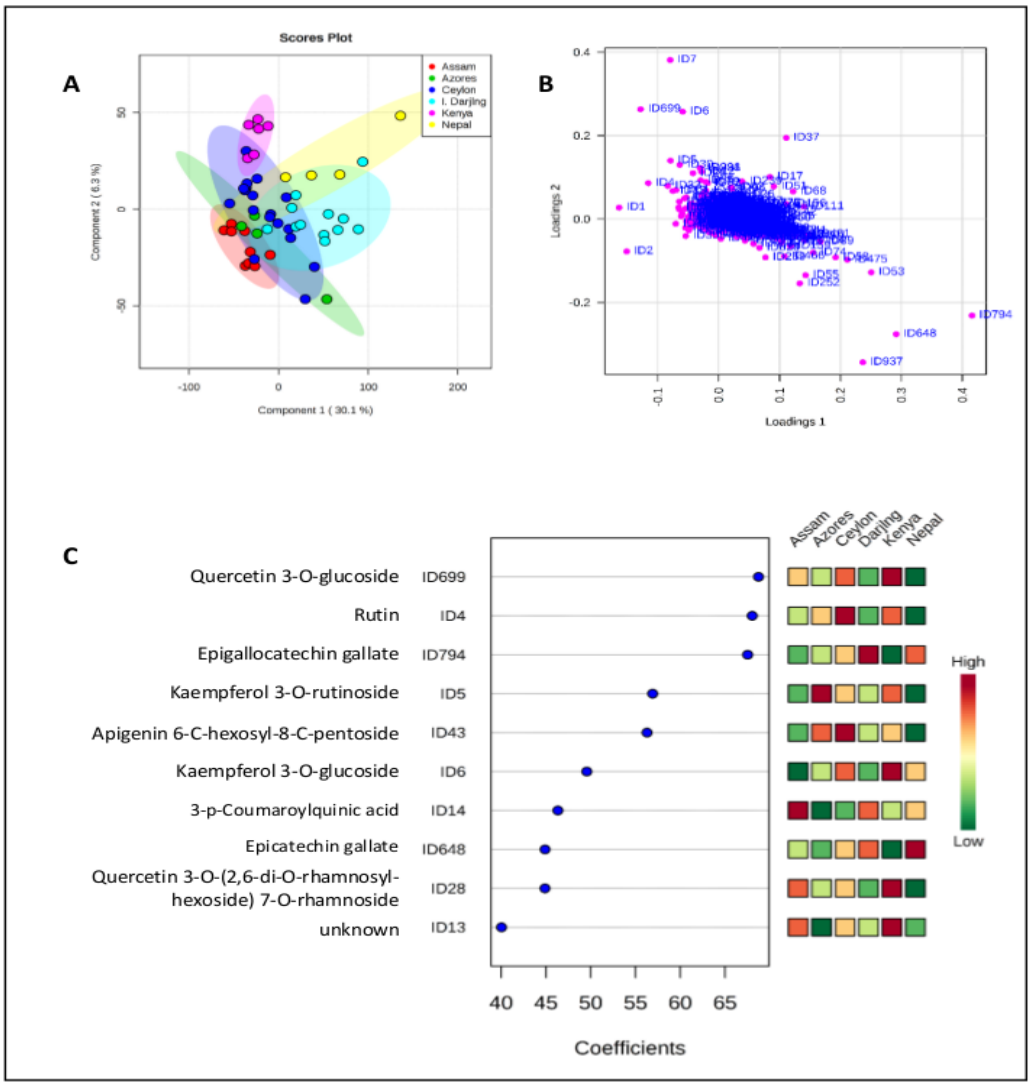


Fig. 4

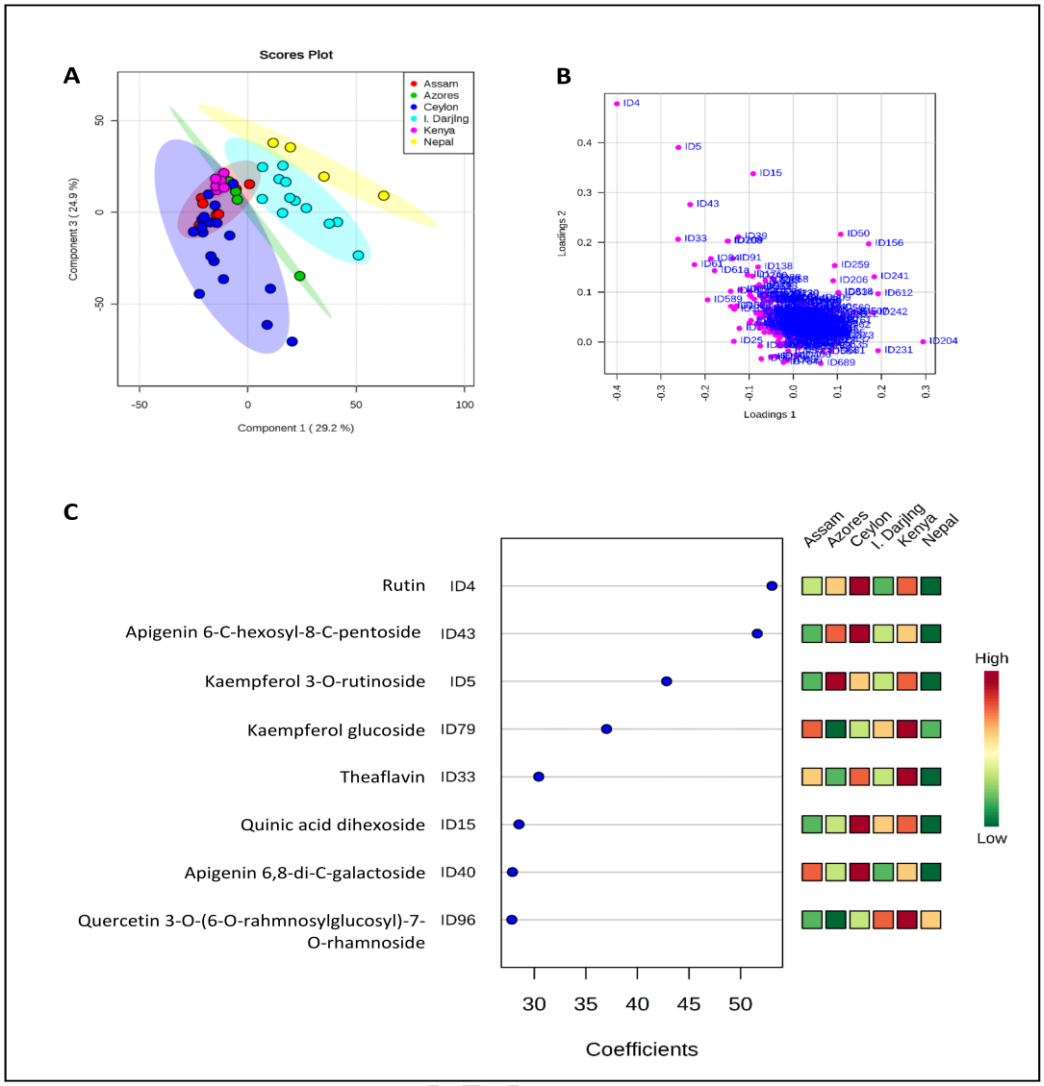


fig. 5

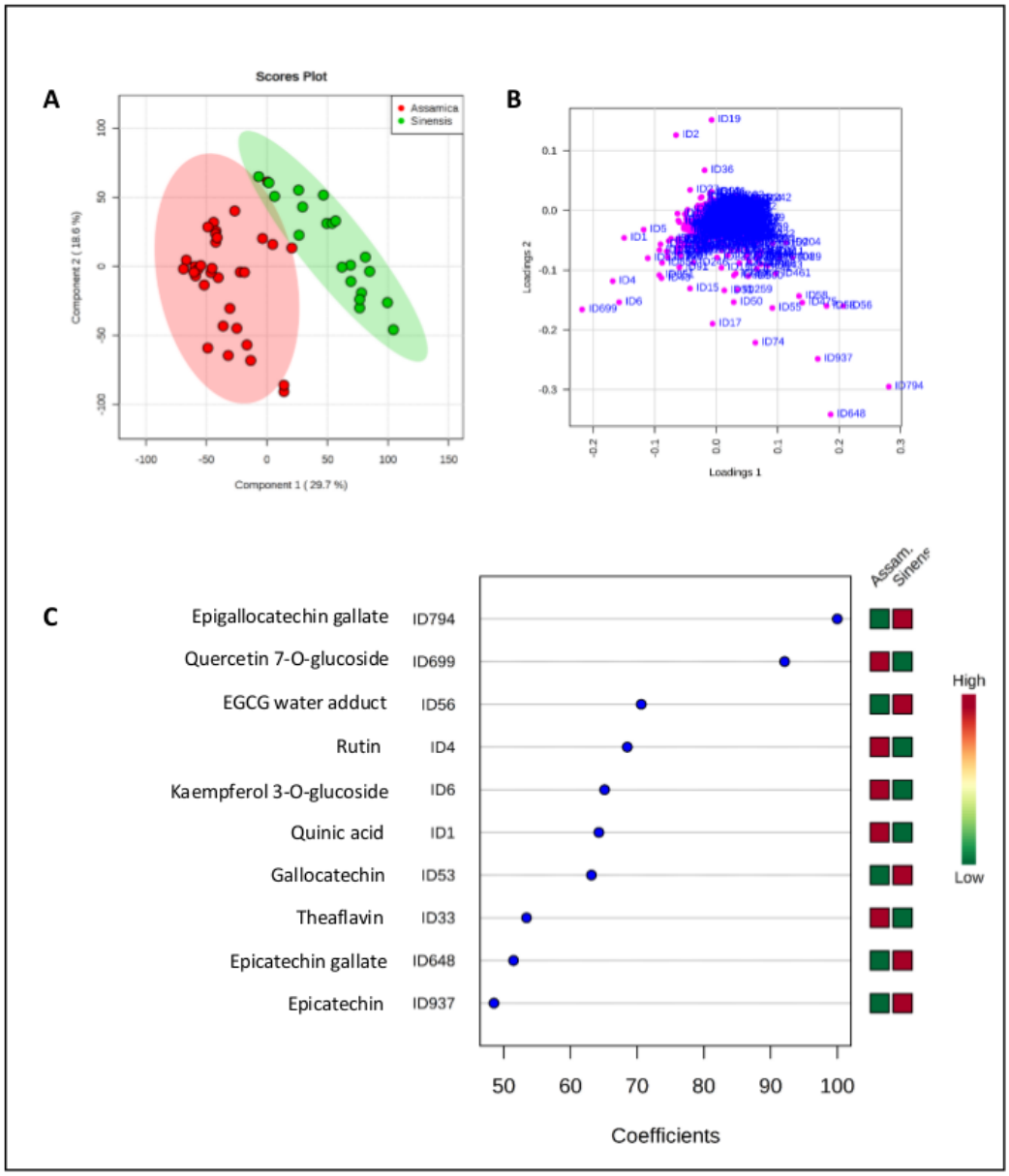


fig. 6

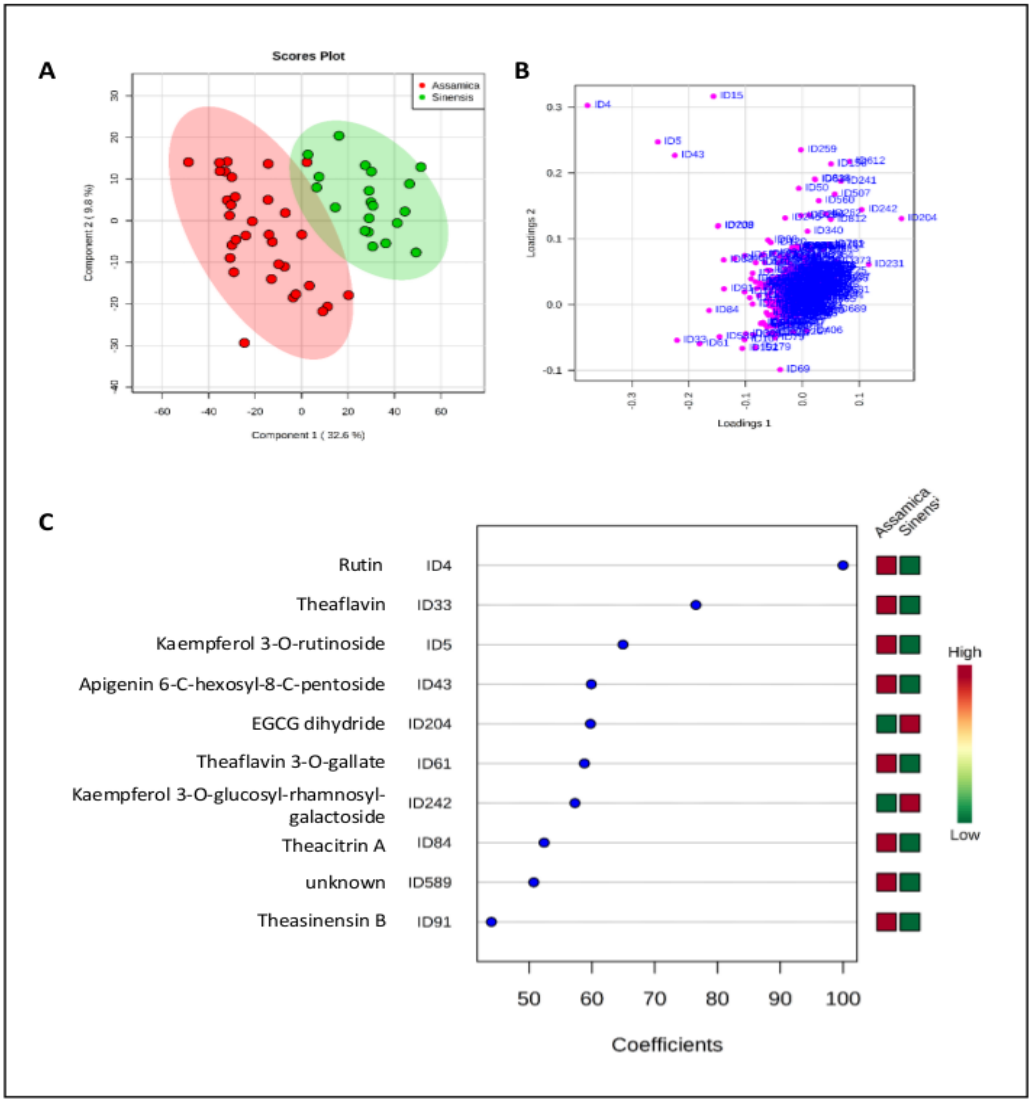


fig. 7

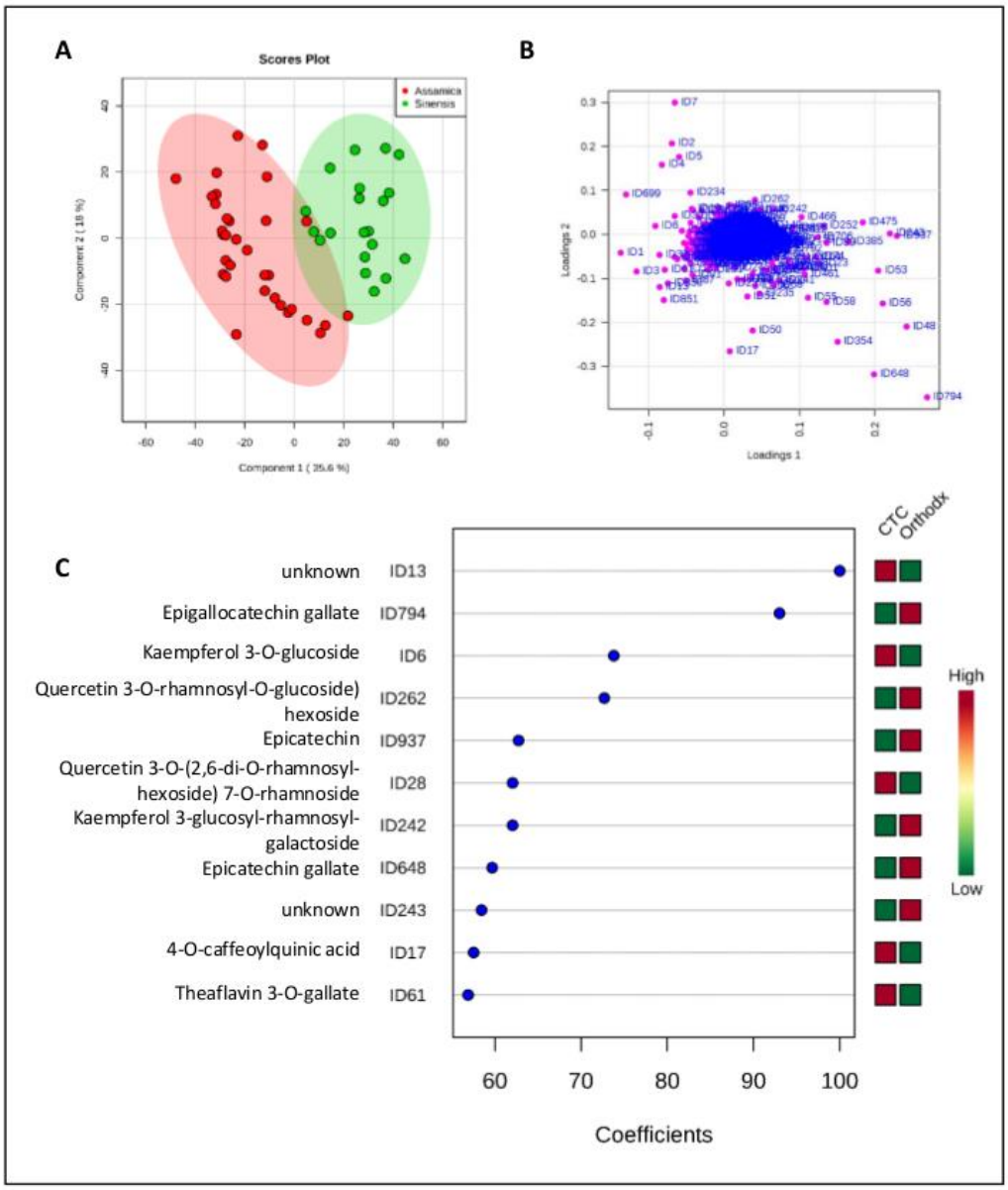


fig. 8

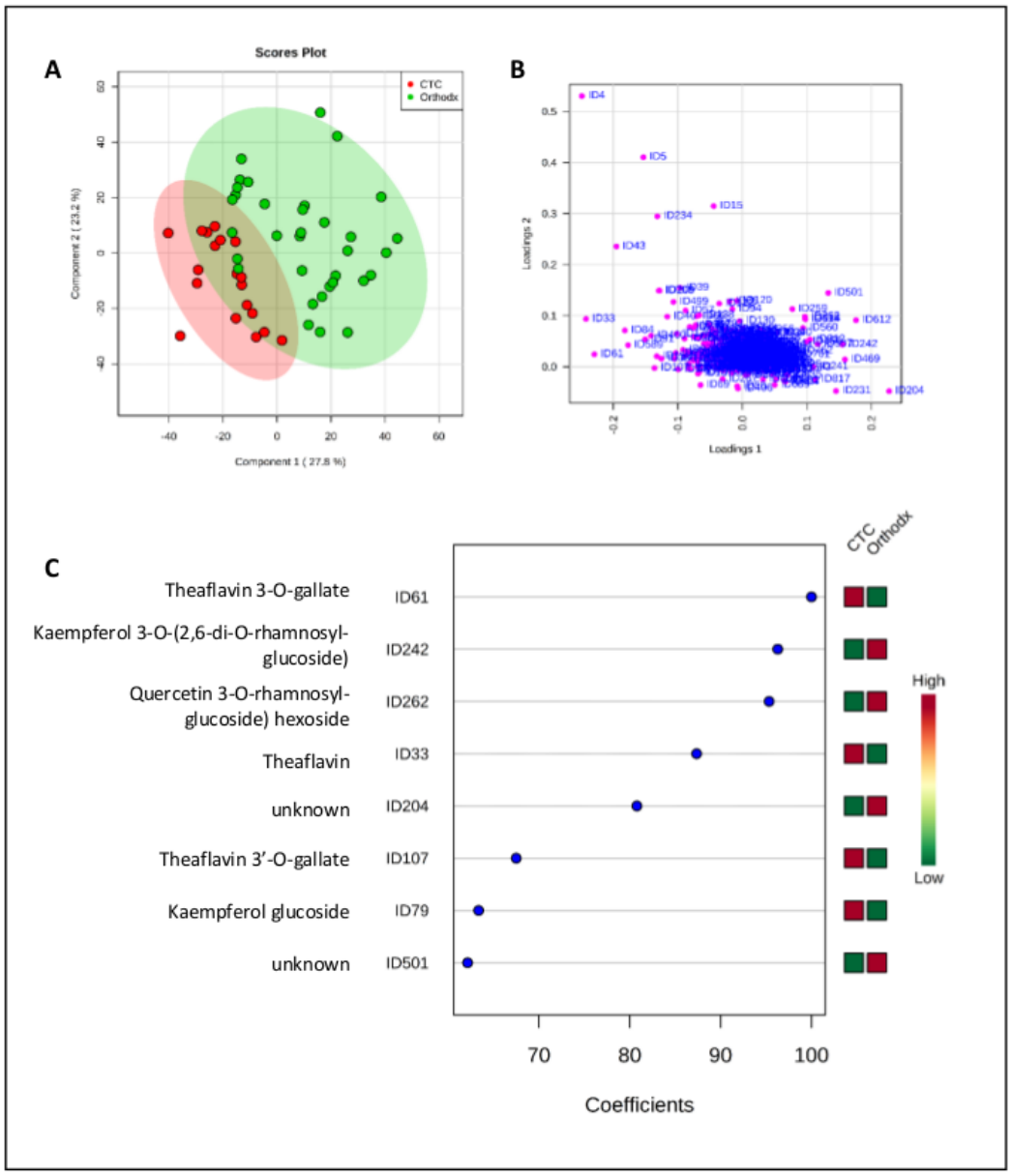
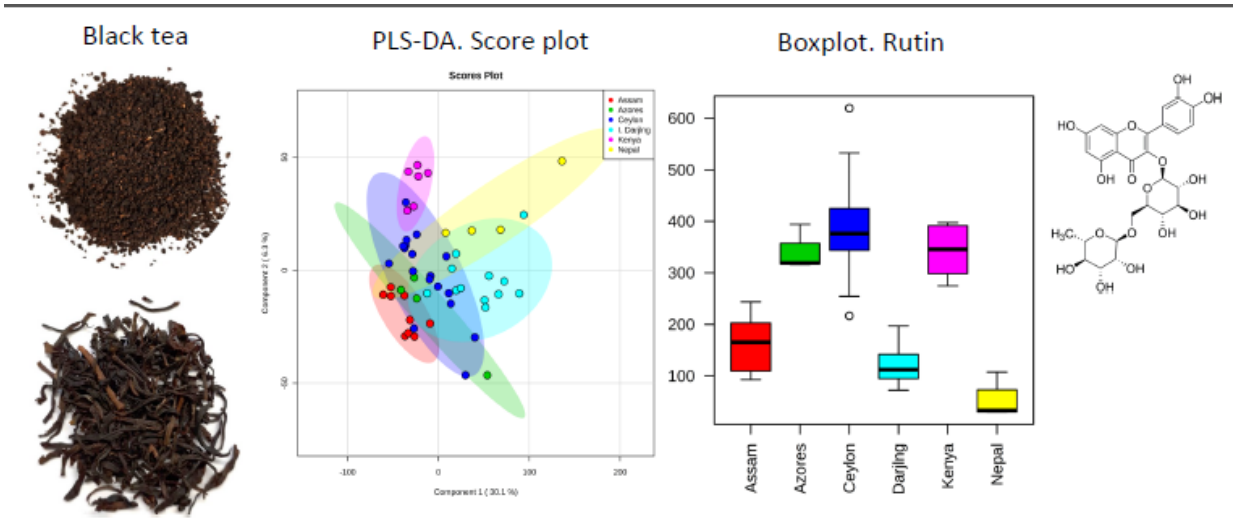


fig. 9

Graphical abstract



Highlights

- Multivariant statistical analysis of LC-MS data allows discrimination of black teas
- Products of black tea fermentation contribute significantly to the variation
- Origin, botanical variety and processing affect black tea composition
- More than 50 biomarkers responsible for variations were identified
- The high chemical complexity of black tea requires dedicated statistical methods