Postprint of Garrido Fernández, A., Benítez-Cabello, A., Rodríguez-Gómez, F., Jiménez-Díaz, R., Arroyo-López, Francisco Noé, Morales, M.L., Relating starter cultures to volatile profile and potential markers in green Spanish-style table olives by compositional data analysis, Food Microbiology (2020), doi: https://doi.org/10.1016/j.fm.2020.103659.

Relating starter cultures to volatile profile and potential markers in green Spanish-style table olives by Compositional Data Analysis

Antonio Garrido Fernández[1], Antonio Benítez-Cabello[1,*], Francisco Rodríguez-Gómez[1], Rufino Jiménez-Díaz[1], Francisco Noé Arroyo-López[1], M. Lourdes Morales[2]

[1]Instituto de la Grasa (CSIC). Departamento de Biotecnología de Alimentos. Campus Universitario Pablo de Olavide. Building 46. Ctra. Sevilla-Utrera, km 1. 410013. Seville, Spain.

[2]Área de Nutrición y Bromatología, Dpto. Nutrición y Bromatología, Toxicología y Medicina Legal Facultad de Farmacia, Universidad de Sevilla, C/P. García González, nº 2, 41012 Seville, Spain

Running title: CoDa analysis of table olive volatiles profile

***Corresponding author:** Antonio Benítez-Cabello. E-mail address: abenitez@ig.csic.es

Abstract

This work relates native lactic acid bacteria (LAB) (*Lactobacillus pentosus* LPG1, *L. pentosus* Lp13, and *Lactobacillus plantarum* Lpl15) and yeast (*Wickerhamomyces anomalus* Y12) starters to the volatile components (VOCs) produced in green Spanish-style table olives. For this aim, the VOC profile was considered as compositional data (CoDa). The CoDa analysis generated new information on the relationship among inocula and VOCs through the tetrahedral plot, CoDa-biplot, variation array matrix, and CoDa dendrogram. The *ilr* (which includes *pivot*) *coordinates* (Euclidean space) from VOCs produced more reliable starters' clustering than the original data. The potential VOC markers, identified by a test based on the pairwise comparison of the logratio variation arrays from the whole data set and the individual groups, were (starters in the parenthesis): 2-phenylethyl acetate (LPG1, Y12, Y12+LAB), methanol (Lpl15), *cis*-2-penten-1-ol (LPG1, Y12, Y12+LAB), 2-methyl-3-hexanol (LPG1, Y12), U (non-identified) C (m/z 83-112-97) (Y12) and UF (m/z 95-154-110) (LPG1, Y12+LAB). Besides, some VOCs were partial/totally inhibited by specific starters: 2-methyl-1-propanol (Lp13, Y12+LAB), 2-phenyl ethanol (Lp13), furfuryl methyl ether (Y12+LAB), purpurocatechol (Y12, Y12+LAB), 4-ethyl guaiacol (Lp13, Lpl15), 4-ethyl phenol (Lpl15), 5-*tert*-butylpyrogallol (Lp13, Lpl15), and UE (m/z 111-198) (Lp13). A better understanding of the relationship between starters and their VOC may facilitate modelling the flavour and quality of Spanish-style green table olive fermentations.

Keywords: fermentation; inoculation; segregation; clustering; classification; compositional data analysis.

1. Introduction

Green Spanish-style represents 50-60% of the world table olives, estimated as $3.26 \cdot 10^6$ tonnes/year by the International Olive Council (IOC, 2019). Its processing consists of debittering of fruits with lye (NaOH solution), washing with tap water, and brining. Then, a spontaneous lactic fermentation produces numerous metabolites (Garrido-Fernández et al., 1997). Apart from the lactic, acetic and other minor acids, the volatile compounds (VOC) play an essential role in the sensory characteristics of the product. The introduction of the of GC/MS stimulated studies on the VOC profile, particularly on the effect of cultivar, growing area, packaging conditions or influence of inoculation (Cortés-Delgado et al., 2016; Sánchez et al., 2017; López-López et al., 2018; Sánchez et al., 2018; Benítez-Cabello et al., 2019). Several of these compounds were related to the "zapatería" spoilage (de Castro et al., 2018). These studies have systematically involved the application of standard statistics and multivariate methods. Moreover, the influence of starter cultures on the sensory characteristics of fermented olives was always obviated. However, those strains associated with the most favourable components could be used for improving the flavour and quality of the final products.

Compositional Data (CoDa) Analysis is a recent statistical methodology proposed initially by Aitchison (1986) to treat data expressed in proportions (e.g. mg/kg, or percentage) of the whole sample. Pawlowsky-Glahn et al. (2015) have also defined them as vectors with strictly positive components that carry relative information. Such structure has specific geometrical connotations because the same absolute difference may not reflect the real (relative) changes. Therefore, its study by multivariate tools, developed for data expressed in absolute values, may lead to useless conclusions (van den Boogaart and

Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018). For treating these data, Aitchison (1986) proposed the use of logratios, although other alternatives like additive (*alr*), centred (*clr*), or isometric logratio (*ilr*) transformations (Egozcue et al., 2003) are also suggested. Recently, *pivot coordinates*, a particular case of *ilr* transformation has also been introduced (Filzmoser et al., 2018). Simultaneously, tools for their treatment in-the-simplex (the sample space for compositions) was also developed. Nowadays, the proper application of CoDa analysis to these data includes stay-in-the-simplex techniques and their transformation into *clr or ilr coordinates*, followed by the study of these coordinates by the standard multivariate tools (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018).

The CoDa analysis is common in geology (Tolosana-Delgado et al., 2011), genetic (Pierotti and Martín-Fernández, 2011), spatial exploration (Lammer et al., 2011), or lipid dynamics in pelagic amphipods (Kraft et al. 2015). Nevertheless, its use in foods is still scarce and related to wine (Hron et al., 2012), pig fat (Ros-Freixedes and Estany, 2014; Garrido Fernández and León Camacho, 2019) or table olives (Garrido Fernández et al., 2018). Recently, the standard multivariate techniques did not adequately segregate among Manzanilla treatments (Benítez-Cabello et al., 2019). In the study of the VOCs of coffee, compounds like acetic acid, 2-methyl pyrazine, furfural, 2-furfuryl alcohol, 2-6-dimethyl hydrazine, and 5-methyl furfural were chosen as relevant markers (Korhoňová et al., 2009). Therefore, the use of the new CoDa statistic to characterise the VOCs produced in green Spanish-style processing is challenging.

The work aims to relate the starter cultures used for the fermentation of green Spanish-style Manzanilla table olives to the formed VOCs, the selection of the most

characteristic components, and the tentative identification of potential markers, using CoDa analysis.

The use of selected microorganisms may represent a good strategy for controlling the flavour of table olives and standardise their quality.

2. Material and Methods

2.1. Olive processing

The olives were from the Manzanilla cultivar, harvested at the green maturation stage. Processing was carried out in cylindrical fermentation vessels (9.5 kg olives/5 L liquid) where the fruits were debittered using a lye solution containing: 32.4 g/L NaOH lye, 21.9 g/L NaCl and 8.9 g/L $CaCl_2$ (97% purity). When the alkali reached 2/3 of the flesh (7 h), the olives were washed with fresh water for 5 h and, finally, brined in a solution having, per litre, 100 g NaCl, 14.2 g $CaCl_2$ and 0.012 L of 35% HCl.

2.2 Treatments

The strains used as starters were: *L. pentosus* LPG1 (onwards LPG1), *L. pentosus* Lp13 (Lp13), *L. plantarum* Lpl15 (Lpl15), and yeast *Wickerhamomyces anomalus* Y12 (Y12), all of them belonging to the Table Olive Microbial Collection (TOMC) of Instituto de la Grasa (CSIC). They were isolated from the surface of fermented table olives and selected because of their technological and probiotic properties (Benítez-Cabello et al. 2019). The experiment consisted of six duplicate fermentation processes (treatments) inoculated with LPG1 (T1), Lp13 (T2), Lpl15 (T3), Y12 (T4), a sequential use of Y12 and a combination of every LAB (T5), and the usual spontaneous process (T6) (Fig. 1). Despite

the initial HCl acid added to the brine, the optimum pH for the LAB inoculation (approx. 6.0-7.0 units) was not reached until the 9th day after brining.

2.3 Inoculation

The LAB were grown on Man, Rogosa and Sharpe (MRS) broth (Oxoid, Basingstoke, Hampshire, England) at 37 ºC for 24h, while yeast was grown on YM broth (Difco) at 28 ºC for 48 h. Cultures were then washed and re-suspended in 0.9% saline buffer. The inoculum sizes were prepared to reach in the cover brine approximately 6 $\log_{10}$ CFU/mL and 5 $\log_{10}$ CFU/mL for LAB and yeasts, respectively. LAB strains were inoculated on the 9th day of fermentation (once the optimum pH was reached), while the yeast was inoculated on the first day after brining. In T5 treatment, the inoculation was sequential, and the mix with all LAB strains was incorporated eight days after inoculating the yeast. The vessels were kept for fermentation (65 days) in the pilot plant facilities of the Instituto de la Grasa (CSIC, Seville, Spain), at room temperature (22±3 ºC). At the end of the process, the samples for analysing the VOCs were withdrawn.

2.4. Analysis of the volatile compounds

The VOCs were obtained by sequential sorptive extraction of brines with Twisters® (TW), using two polydimethylsiloxane TW in each sample. The operation was carried out first in immersion (SBSE), for semi volatile, and then in the head space (HSSE), for highly VOCs (Ubeda et al., 2016). The procedure improves the sensitivity of the just HSSE extraction (Úbeda et al., 2016). Six mL of brine, 1.8 gr of NaCl (30% w/v), and 8 µL of the internal standard 4-methyl-2-pentanol (1,044 mg/L final concentration) were placed in a 20 mL vial. SBSE was performed for 1 h and continue stirring with a conventional (non-

112 coated) magnetic stir bar at 200 rpm and room temperature, using a Twicester® to keep the
113 TW immersed. TW was removed, rinsed with Milli-Q water and dried with tissue paper.
114 After this, a new TW was placed in an open glass inserted in the same vial for HSSE
115 extraction. The vial was again tightly capped and heated in a thermostatic bath at 62 ºC for
116 1 h. Then, the TW was cleaned and dried with tissue paper. Both TWs were simultaneously
117 desorbed in the GC/MS by introducing them into the same desorption tube.

118 The analyses were performed in an Agilent 6890 GC system coupled up to an
119 Agilent 5975 inert quadrupole mass spectrometer equipped with a Gerstel Thermo
120 Desorption System (TDS2), a Cooling Injector System CIS-4 PTV inlet (Gerstel, Müllheim
121 an der Ruhr, Germany), aJ&W CPWax-57CB column (50 m x 0.25 mm and 0.20 μm film
122 thickness) (Agilent, Santa Clara, CA, US). The detector was never saturated. Table S1
123 (supplementary material) shows identification details. The concentrations were expressed
124 as relative peak area to an internal standard of the target ion of each compound.

125 ***2.5 CoDa analysis***

126 VOCs are usually studied by standard multivariate methods developed for data in
127 the Euclidean space but, due to the estimation method, such as profiles contain only relative
128 information (Aitchison, 1986) that directly affects the covariance structure. On the contrary,
129 the CoDa methodology preserves their relative scale property (Filzmoser et al., 2018). The
130 Appendix in supplementary material contains succinct information on the most relevant
131 techniques used in this work. For detailed explanations, readers should consult specialized
132 texts (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018).

The CoDa analysis was performed using the packages CoDaPack (Comas-Cufí and Thió-Henestrosa, 2011), robCompositions R (Templ et al., 2011), and the plug in XLSTAT v.2017 for Excel (Addinsoft, Paris, France).

3. Results and discussion

3.1 Data set

The data set consisted of 12 rows (duplicate treatments) and a sub-composition of VOCs with significant differences between at least two treatments (Benítez-Cabello et al. (2019). Compounds not conclusively identified yet (21) are reported just as m/z values (see Table S1 in supplementary material). The profiles included acetates (3), acids (1), alcohols (19), aldehydes (2), sulfoxide (1), C_{13}-norisoprenoid (1), ethyl ester (3), furan (1), ketones (3), methyl esters (3), phenols (8), terpenes (3), and other (1) as well as several unknown (U) compounds. Therefore, apart from the lactic and acetic acid production (data not shown), the formation of alcohols and their esters characterised the fermentations. Several compounds were not detected (n.d.) in some treatments (cells with zeros). CoDa analysis considers them as rounded zeros (presence below the detection limit, common in analytical chemistry) and recommends their replacement by a reasonable low value (65% of the detection limit) (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018). The treatments with the largest number of n.d. compounds (in parenthesis) were T2 (9), inoculated with Lp13, and T3 (5), inoculated with Lpl15. The presence of non-identified VOCs in table olive studies is frequent (Sánchez et al., 2018; de Castro et al., 2018) because of high microbial diversity during the current fermentation conditions.

3.2 Central tendency and dispersion

In CoDa, the central tendency and dispersion of components are represented by their geometric means (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018) and percentiles, respectively (Table S1), although noticing that the latter rely on the concrete scale used. Values (0-100%) ranged between 0.1185 (ethanol) and 0.0006 (*cis*-3-hexenyl acetate). The parts with the highest dispersion, supposedly due to the effect of treatments, could be the most appropriate to segregate among starters. However, variability associated with determination errors of components in low concentrations should not be underestimated (Korhoňová et al., 2009).

3.3 Variation array

In CoDa analysis, the so-called variation array presents the variances of the logratios of each part over the others (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018) in the upper diagonal (Table S2 supplementary material). As the matrix is symmetric, the lower diagonal shows the averages of their matching logratios. The highest logratio variances (upper diagonal) were found for ln(UF/purpurochatechol) (23.3285), followed by ln(*cis*-penten-1-ol/purpurochatechol) (22.1243). Nevertheless, in practice, the dispersion within each component is evaluated by its *clr* variance, i.e. the variance of its *clr* transformed *coefficients* across fermentation processes (Table S2, last column) as the *clr* coefficients aggregate all logratios with a given component. The most relevant were: UF (with 6.5329) (n.d. in T2, T3, T6); *cis*-2-penten-ol (6.4804); 4-ethylguaiacol (6.3033) (n.d. in T2, T3); 2-methyl-1-propanol (6.1982); 2-ethynyl-2-butenal (6.0914); purpurocatechol (6.0677) (n.d. in T4, and T5); 2-phenylethyl acetate (5.6235); 5-*tert*-butylpyrogallol (5.0818) (n.d. in T2, T3, and one replicate of T5); 2-methyl-3-hexanol (4.4568) (n.d in T2, T3, T6); 1-butanol (3.5247) (T6); furfuryl methyl ether (3.4963) (n.d in T5); UC (3.3188)

(n.d. in one replicate of T1, T2, T3, T5, and T6), and UE (3.1994) (n.d. in T2). Together, they represent about 83.84% of the total *clr* variance. Several of the cases of high variance corresponded to components below the detection limit/low central values in some fermentation processes; however, their variances could also respond to relevant differences between bacterial performances) and makes pertinent to their considering.

3.4 Tetrahedral plot

The association of inocula with VOCs can be visualised in the simplex as a function of, at maximum, four components, usually chosen among those with the highest variance (i.e., the greatest segregation power) (Fig. 2). T3 (inoculated with Lp13) and T6 (spontaneous) treatments are different due to their high and moderate contents of 2-methyl-1-propanol (I), respectively, but both are poor in *cis*-2-penten-1-ol (O) and UF (BE). T1 (LPG1) was also different due to its low level of 2-methyl-1-propanol (I) and modest concentrations of the remaining VOCs. Besides, T2 (Lp13) is very low (or below detection limits) in 2-methyl-1-propanol (I) and 4-ethylen guaiacol (AP). T4 (Y12) and T5 (Y12+LAB) are relatively close and have a moderate presence of the four components. Furthermore, the plot also includes the three Principal Components (PCs), which are used to detect possible linear relationships between treatments. However, in this case, fermentation processes did not follow any trend. Then, the plot highlighted the peculiar VOC profiles of the spontaneous fermentation (T6), T3 (Lpl15), and T1 (LPG1) and prevents against any linear evolution of processes (at least as a function of these four compounds).

3.5 CoDa-biplot

The CoDa biplot (Aitchison and Greenacre, 2002), based on *clr coefficients* and PCA, explained 72.1% of the total variance and required particular interpretation. The covariance option (Fig. 3 A) allows studying the relationships among VOCs. The distances between the ends of the rays (links) are proportional to their logratio variances. The largest values were observed between *clr*BE (UF) or *clr*D (2-phenylethyl acetate) and either *clr*AI (purpurocatechol), *clr*Y (2-ethenyl-2-butenal), *clr*AF (furfuryl methyl ether), or *clr*I (2-methyl-1-propanol), with progressive lower values. On the contrary, *clr* components following similar trends and adjacent rays lead to almost constant logratios, indicating a strong correlation, and redundant information: e.g. *clr*BE (UF), *clr*O (*cis*-2-penten-1-ol), and *clr*D (2-phenylethyl acetate); *clr*AT (5-*tert*-butylpyrogallol) and *clr*BD (UE); or *clr*AI (purpurocatechol), and *clr*Y (2-ethenyl-2-butenal). Such relationships may be interpreted as parallel productions. VOCs situated close to the centre can indicate low relevance or poor representation on the PC1/PC2 plane. Nonetheless, the additional contribution of PC3 was reduced (9.27% total variance), and only *clr*J (1-butanol), associated to PC3, was well represented on the PC2/PC3 plane.

In form biplot (Fig. 3 B), the distances between symbols are an approximation of the distances between processes. In the plot, the replicates were close, indicating that they followed similar trends, particularly those fermented with individual strains (T1, LPG1; T2, Lp13; T3, Lpl15; and T4, Y12); however, those inoculated with Y12+LAB (T5) and the spontaneous (T6) were moderately distant, situation compatible with their less rigid processing conditions. The projections of processes onto PC2/PC3 plane did not improve the interpretation.

Regardless of the type of biplot, there are some vertices lying in a straight line. For example, *clr*I (2-methyl-1-propanol), *clr*BD (UE), *clr*S (2-methyl-3-hexanol) and any of *clr*D (2-phenylethyl acetate), *clr*O (*cis*-2-penten-1-ol), or *clr*BE(UF)) reveal logratios of high correlation (e.g. VOCs produced in parallel) which could deserve further studies. Finally, parts forming a rectangle (a,b,c,d) reveal a simple logratio contrast of the form: ln(a)-ln(b)+ln(c)-ln(d)=constant. An example could be *clr*BB (UC), *clr*I (2-methyl-1-propanol), *clr*AF (furfuryl methyl ether), and any of the *clr* components close to the origin (e.g. *clr*AD (ethyl 5,6-dimethylnicotinate)). Therefore, the CoDa biplot had the striking ability to display the relationships among the most relevant components, and their logratios, which condense the data structure. Also, it made evident the clear differences between the VOCs from the diverse starters and, even, some particularities between replicates in case of lax microbial control (T5, a combination of Y12+LAB, and T6, spontaneous,).

3.6 Sequential binary partition, ilr transformation (coordinates) and dendrogram of balances

For transforming the original CoDa data set into the Euclidean space, one possibility to obtain *ilr coordinates* is to construct them using the sequential binary partition (SBP). Apart from the standardization factor, it consists of dividing the parts successively into two non-overlapping subgroups and estimating their balances (Egozcue, and Pawlowsky-Glahn, 2005). In this work, the SBP compares successively (in order of descending variances) each of the following compounds (numerator) over de geometric means of the remaining components (denominator): UF, *cis*-2-penten-1-ol, 4-ethylguaiacol, 2-methyl-1-propanol, 2-ethenyl-2-butenal, purpurocatechol, 2-phenylethyl acetate, 5-*tert*-butylpyrogallol, 2-methyl-3-hexanol, 1-butanol, furfuryl methyl ether, UC, UE, 3-

methylbutanoic acid, and methyl acetate. After the 14th, the balances were successively formed as the logratio between the first still not used component over the remaining ones. The process ended after estimating the logratio between the last two parts. The SBP matrix (Table S3, supplementary material) summarizes the successive steps. There, 1, -1, and 0 denote the components used in the numerator, denominator, or not participating in the partition, respectively. For improving understanding, the means of balances and their variances are also included in this matrix (Table S3, last two columns). To highlight the presence of both positive and negative logratio balances (*ilr coordinates*), as in the Euclidean sampling space.

The CoDa dendrogram is the graphical presentation of balances. There, the mean values are represented in the horizontal axis (Fig. 4) while the vertical lines stand for the variances of the overall balances. The first 14 balances account for 91.33% of the total variance (Table S3), which could be a good approximation for representing the data structure. The information from the remaining balances looks like mere noise (Fig. 4).

The *coordinates* obtained by this SBP are somewhat similar to the *pivot coordinates* (Filzmoser et al., 2018), which is a particular form of balance. Both are essential for the transformation of data into *coordinates* in the Euclidean space, where can be analyzed by standard multivariate tools.

3.7 Effect of the clr and ilr transformations on the fermentation processes' segregation power

In CoDa analysis, the input for clustering is not the original dataset but dissimilarities; that is, the matrix of distances (for observations) or the variation matrix (for

variables). The Euclidean distances of the original data are not reliable (they do not follow geometrical properties of CoDa) and are quite different from those estimated according to CoDa analysis principles using the Aitchison distance (Table S4, supplementary material). Furthermore, this Aitchison distance is preserved even when transforming the original data into the Euclidean space as *clr coefficients* or *ilr coordinates* (Table S4). Therefore, clustering using the original data set can mislead grouping, as occurred in this case (Fig. 5 A) where replicates of the same fermentation process were assigned to different groups. However, clustering using the *ilr coordinates* grouped, on the left, the three rich in VOCs inoculated treatments LPG1 (T1), Y12 (T4) and Y12+LAB (T5), and on the right those with moderate volatile contents Lp13 (T2), Lpl15 (T3), and spontaneous (T6). Besides, there was no incorrect assignation of replicates of their corresponding treatments (Fig. 5 B). *Pivot coordinates* led to the same result (Fig. 5 C) than another choice of *ilr coordinates* because, as demonstrated previously, the distances between cases (processes) in CoDa do not depend on the transformation used. Furthermore, the 14th first *ilr coordinates* also led to similar association (Fig. 5 D), indicating that the remaining balances might mainly contribute with noise, in agreement with Fig. 4. In Spanish-style Gordal fermentations, the fatty acid data in their original units also led to the worst grouping of processing steps than using *ilr coordinates* (Garrido Fernández et al., 2018).

The improving of the segregation power also was observed when PCA was applied. Using the original data led to a poorer representation and segregation (Fig. 6 A) than in case of *clr coefficients* (Fig. 6 B) and *ilr coordinates* (Fig. 6 C), which show a more realistic separation of processes according to starters. Furthermore, the first fourteen balances of the whole set of *ilr coordinates* was also as efficient as *pivot* or *ilr coordiantes*

290 since the results (Fig. 6 D) were comparable, corroborating the noise from non-influential
291 VOCs (Fig. 4).

292 Definitively, using *pivot coordinates* or relevant general *ilr coordinates* led to clear
293 segregation among treatments (starters) than with the original VOCs, in agreement with the
294 CoDa hypothesis. In contrast, the standard multivariate tools directly applied to
295 compositional data may lead to misleading results.

296 Clustering can also be achieved according to variables (or Q-mode) (van den
297 Boogaart and Tolosana-Delgado, 2013; Filzmoser et al., 2018). In the Euclidean geometry,
298 the association between the components is measured by the Pearson correlation coefficient,
299 while in CoDa, the relationship can be deduced from the variation array matrix. CoDa Q-
300 clustering, based on the variation array matrix and using both classic and robust (preferable
301 because allow suppressing the influence of possible outliers) methods segregated two main
302 groups (Fig. 7). The first, on the left, consisted of: 2-methyl-3-hexanol (S); UF (BE); 2-
303 phenylethyl acetate (D); *cis*-2-penten-1-ol (O); UE (BD); 4-ethyl guaiacol (AP); and 5-*tert*-
304 butylpyrogallol (AT). It also included UC (BB) in case of the robust option. Besides, there
305 was a second common group (classic and robust options) on the right, which included 2-
306 ethenyl-2-butenal (Y), purpurocatechol (AI), 2-methyl-1-propanol (I), and furfuryl methyl
307 ether (AF). Interestingly, these components also showed the highest variances in the
308 variation matrix; i.e. could have the greatest segregation power. However, the largest group
309 (in the centre) was somewhat different in the two methods, with the robust option showing
310 a very close relationship among components (Fig. 7, bottom panel), in agreement to
311 previous observations.

These clustering results regarding treatments were also in agreement with those observed in other works on green Spanish-style table olives according to cultivars and growing area, which were always more accurate when following CoDa techniques (Garrido Fernández et al., 2018). Despite these eviences, standard multivariate methods, using the original VOCs dataset, was applied in stoned Spanish-style table olives for segregating compounds by chemical classes (Malheiro et al., 2011), studying the evolution of VOCs during olive processing (Dabbou et al., 2011), differentiating normal from spoiled products (De Castro et al., 2018), or relating sensory analysis to volatile composition (López-López, et al., 2018).

3.8 Identification of potential markers vs the spontaneous fermentation process

For this purpose, the Walach et al. (2017) method was used. Briefly, it consisted of comparing the pairwise logratio variation array matrix corresponding to the two groups (full data set) with those estimated from each one separately. The result is expressed in terms of V_j^* (Appendix; Walach et al., 2017). Compounds which V_j^* exceeded the 1.96 cut-off limit ($p<0.05$) were considered significant and potential markers. The methodology was applied for obtaining the V_j^* values for the whole set of VOC comparisons between the inoculated (starters) and the spontaneous process (Fig. 8, for the case of T4 (Y12) vs T6 (spontaneous). The significant compounds were identified by their respective indexes (Table 1). Several significant compounds (high/low contents) were not exclusive for a specific inoculum but common to various (Table 1).

According to Table 1, the formation of the following VOCs was promoted by the respective strains (in parenthesis) and could then be considered as potential markers for

them: 2-phenylethyl acetate (LPG1, Y12, Y12+LAB), methanol (Lpl15), cis-2-Penten-1-ol (LPG1, Y12, Y12+LAB), 2-methyl-3-hexanol (LPG1, Y12), UC (Y12), and UF (LPG1, Y12+LAB). 1-butanol (LPG1, Lp13, Lpl15, Y12, Y12+LAB) would also be included in this group, but its wide distribution in previous studies (Cortés-Delgado et al., 2016; Sánchez et al., 2017; de Castro et al., 2018; López-López et al., 2018; Sánchez et al., 2018) and its formation by all LAB and yeast strain fermentations prevents its consideration as a marker; however, it seems to be characteristic of the inoculated processes.

Besides, some starters (in parenthesis) can reduce/inhibit the formation of others VOCs: 2-methyl-1-propanol (Lp13, Y12+LAB), 2-phenyl ethanol (Lp13), furfuryl methyl ether (Y12+LAB), purpurocatechol (Y12, Y12+LAB), 4-ethyl guaiacol (Lp13, Lpl15), 4-ethyl phenol (Lpl15), 5-*tert*-butylpyrogallol (Lp13, Lpl15), and UE (Lp13). In this case, 4-ethyl guaiacol (Lp13, Lpl15) and 4-ethyl phenol (Lpl15) have been mentioned in other works (Cortés-Delgado et al., 2016; Sánchez et al., 2017; de Castro et al., 2018; López-López et al., 2018; Sánchez et al., 2018), but their inhibition in some fermentations can be regarded as characteristic of their respective inoculated strains.

Some of these possible markers provide specific aromatic notes. Related to LPG1 and Y12 were 2-phenylethyl acetate, which gives sweet roses (Suárez-Lepe and Morata 2012) or flowery with honey notes (Lilly, Lambrechts, Pretorius, 2000), and *cis*-2-penten-1-ol, associated with green aroma notes (Acree and Arn, 2019). On the contrary, the fermentation by Lpl15 was characterized by the presence of 4-ethyl phenol, which is considered as an off-flavour for its horse stable–like, faecal, and medicinal odour (Czerny et al., 2011); therefore, its formation in high proportion could represent a serious obstacle for the use of this strain as inoculum.

4. Conclusions

This study has demonstrated that applying CoDa analysis introduces new exploratory techniques like tetrahedral plot, biplot, CoDa-dendrogram, or variation array, which were useful for segregating processes according to inocula or studying relationships among VOCs and potential markers. Thus, the study opens the possibility of using specific starter cultures for the production of particular VOCs or the prevention of undesirable compounds in real fermentation conditions, i.e. for modelling the flavour and quality of green Spanish-style table olives. Furthermore, the association of compounds with distinct strains may facilitate the study of the biological pathways of their formation.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The research was funded by the Spanish Government (Project OliFilm AGL-2013-48300-R: www.olifilm.science.com.es) A-BC thanks the Spanish Ministry of Economy and Competitiveness for their FPI grant.

Supplementary material 1. Concise comments on the compositional data analysis techniques used in the work.

Supplementary material 2. Tables S1-S4.

References

376 Acree, T., Arn, H., 2004. Flavornet and human odor space
377 http://www.flavornet.org/flavornet.html. Accessed date: August 2020.

378 Aitchison, J., 1986. The Statistical Analysis of Compositional Data, reprinted in 2003 with
379 additional material by The Blackburn Press. New Jersey (USA) (Chapman & Hall Ltd.)

380 Aitchison, J., Greenacre, M., 2002. Biplots for compositional data. J. Roy. Stat.l Soc., C
381 Appl. Stat. 51, 375-392.

382 Benítez-Cabello, A., Rodríguez-Gómez, F., Morales, M.L., Garrido-Fernández, A.,
383 Jiménez-Díaz, R., Arroyo-López, F.N., 2019. Lactic acid bacteria and yeast inocula
384 modulate the volatile profile of Spanish-style green table olive fermentations. Foods. 8
385 (8), 1–17. doi:10.3390/foods8080280.

386 Comas-Cufí M, Thió-Henestrosa S., 2011. CoDaPack 2.0: a stand-alone, multi-platform
387 compositional software. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI, eds.
388 CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Sant Feliu
389 de Guíxols; 2011.

390 Cortés-Delgado, A., Sánchez, A.H., de Castro, A., López-López, A., Beato, V.M.,
391 Montaño, A., 2016. Volatile profile of Spanish-style green table olives prepared from
392 different cultivars grown at different locations. Food Res. Int. 83, 131-142.
393 dx.doi.org/10.1016/j.foodres.2016.03.005

394 Czerny, M., Brueckner, R., Kirchhoff, E., Schmitt, R., Buettner, A., 2011. The influence of
395 molecular structure on odor qualities and odor detection thresholds of volatile alkylated
396 phenols. Chemical Senses 36, 539–553. doi: 10.1093/chemse/bjr009

Dabbou, S., Issaoui, M., Brahhmi, F., Nakbi, A., Chehab, H., Mechri, B., Hammani, M., 2011. Changes in the volatile compounds during processing of Tunisian-style table olives. J. Am. Oil Chem. Soc. 89, 347-354. doi: 10.1007/s11746-011-1907-8

de Castro, A., Sánchez, A.H., A. López-López, A., Cortés-Delgado, A., Medina, E., Montaño, A., 2018. Microbiota and metabolite profiling of spoiled Spanish-style green table olives. Metabolites 8, 73; PMC6316098. doi: 10.3390/metobo8040073.

de Castro, A., Sánchez, A.H, Cortés-Delgado, A., López-López, A., Montaño, A., 2019. Effect of Spanish-style processing steps and inoculation with *Lactobacillus pentosus* stater culture on the volatile composition of cv. Manzanilla green olives. Food Chem. 271, 543-549. doi.org/10.1016/j.foodchem.2018.07.166

Egozcue, J.J., Pawlowsky-Glahn, V. 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7), 795-828.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateo-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geology, 35, 279-300.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis, with worked examples in R. Springer Nature Switzerland AG. Cham , Switzerland.

Garrido-Fernández, A., Fernández-Díez, M.,J., Adams, R. M., 1997. Table Olives Production and Processing. London: Chapman & Hall.

Garrido Fernández, A., Cortés Delgado, A., López López, A., 2018. Tentative application of compositional data analysis to the fatty acid profiles of green Spanish-style Gordal table olives. Food Chem. 241, 14-22. doi.org/10.1016/j.foodchem.2017.08.064

Garrido Fernández, A., León Camacho, M., 2019. Assessing the effect of season, *montanera* length, and sampling location on Iberian pig fat by compositional data analysis and standard multivariate statistics. Food Chem. 295, 377-386. dx.doi.org/10.1016/j.foodchem.2019.05.123

Garrido-Fernández, A., Montaño, A., Sanchez Gómez, A.H., Cortés-Delgado, A., López-López, A., 2017.Volatile profile of green Spanish-style table olives: Application of compositional data analysis for the segregation of their cultivars and production areas. Talanta, 169, 77-84. http://dx.doi.org/10.1016/j.talanta.2017.03.066

Hron, K., Jelínková, M., Filzmoser, P., Kreziger, R., Bednář, P., Barták, P., 2012. Statistical analysis of wines using a robust compositional biplot. Talanta, 90, 46-50. https://doi.org/10.1016/j.talanta.2011.12.060.

IOC, International Olive Oil Council., 2019. World table olive figures. http://www.internationaloliveoil.org/estaticos/view/132-world-table-olive-figures Last updated: March 2019.

Korhoňová, M., Hron, K., Klimčiková, D., Muller, L., Bednař, P., Barták,P., 2009. Coffee aroma-statistical analysis of compositional data. Talanta. 80 (2), 710-715. https://doi.org/10.1016/j.talanta.2009.07.054

436 Kraft, A, Graeve, M., Janssen, D., Greenacre, M., Falk-Petersen, S., 2015. Artic pelagic
437 amphipods: lipid dynamics and life strategy. J. Plankton Res. 37, 790-
438 807. doi.org/10.1093/plankt/fbv052

439 Lammer, H., Wurz, P., Martín-Fernández, J., Lichtenegger, H.I.M., 2011. Compositional
440 data analysis in planetology: the surfaces of Mars and Mercury. In *Compositional Data*
441 *Analysis: Theory and Practice* 1srt Ed. Pawlowsky-Glahn, V., Cuccianti, A. eds). Willey
442 & Sons. Chichester (UK). pp 267-281.

443 Lilly, M., Lambrechts, M.G., Pretorius, I.S., 2000. Effect of increased yeast alcohol
444 acetyltransferase activity on flavor profiles of wine and distillates. Appl. Environ.
445 Microbiol. 66 (2), 744-753. doi: 10.1128/aem.66.2.744-753.2000

446 López-López, A., Sánchez, A.H., Cortés-Delgado, A., de Castro, A., Montaño, A., 2018.
447 Relating sensory analysis with SPME-GC-MS data from Spanish-style green table olive
448 aroma profiling. LWT-Food Sci. Technol. 89, 725-734.
449 https://doi.org/10.1016/j.lwt.2017.11.058.

450 Malheiro, R., Guedes de Pinho, P., Casal, S., Bento, A., Pereira, J.A., 2011. Determination
451 of the volatile profile of stoned table olives from different varieties by using HS-SPME
452 and GC/IT-MS. J. Sci.Food Agr. 91, 1693-1701. doi:10.1002/jsfa.4372

453 Pawlowsky-Glahn, V., Egozcue, J.J., 2011. Exploring compositional data with the CoDa-
454 dendrogram. Aust. J.Stat. 40, 103-113.

455 Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. *Modeling and analysis*
456 *of compositional data*. John Wiley & Sons Ltd. Chichester, U.K.

457 Pierotti, M., E., R., Martín-Fernández, J., 2011. Compositional analysis in behavioural and
458 evolutionary ecology. In Pawlowsky-Glahn, V., Cuccianti, A. (eds, *Compositional Data*
459 *Analysis: Theory and Practice* (pp 218-234. Willey & Sons., Chichester (UK.

460 Ros-Freixedes, R., Estany, J., 2014. On the compositional analysis of fatty acids in pork. J.
461 Agr., Biol.Env. Stat, 19, 136-155.

462 Sánchez, A.H., de Castro, A. López-López, A., Cortés-Delgado, A., Beato, V.M., Montaño,
463 A., 2017. Retention of color and volatile compounds of Spanish-style green table olives
464 pasteurized and stored in plastic containers under conditions of constant temperature.
465 LWT- Food Sci. Technol. 75, 685-691. https://doi.org/10.1016/j.lwt.2016.10.027

466 Sánchez, A.H., López-López, A., Cortés-Delgado, A., Beato, V.M., Medina, E., de Castro,
467 A., Montaño, A., 2018. Effect of post-fermentation and packaging stages on the volatile
468 composition of Spanish-style green table olives. Food Chem. *239*, 343-353.
469 http://dx.doi.org/10.1016/j.foodchem.2017.06.125

470 Suárez-Lepe, J.A., Morata A., 2012. New trends in yeast selection for winemaking. Trends
471 Food Sci. Technol. 23, 39-50. https://doi.org/10.1016/j.tifs.2011.08.005

472 Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: An R-package for Robust
473 Statistical Analysis of Compositional Data, in Compositional Data Analysis: Theory and
474 Application , Pawlowsky-Glahn Buccianti, Edts. Wiley & Sons, London, U.K.

475 Tolosana-Delgado, R., Eynatten, H.V., Karious, V., 2011. Constructing modal mineralogy
476 from geochemical composition:a geometric Bayesian approach. Math. Geosci. 37 (5),
477 SI, 677-691. doi:10.1016/j.cageo.2010.08.005

478 Ubeda, C., Callejón, R.M., Troncoso, A.M., Peña-Neira, A., Morales, M.L., 2016. Volatile
479 profile characterisation of Chilean sparkling wines produced by traditional and Charmat
480 methods via sequential stir bar sorptive extraction. Food Chem. 207, 261–271. .
481 https://doi.org/10.1016/j.foodchem.2016.03.117.

482 van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing compositional data with
483 Springer-Verlag. R. Berlin Heidelberg, Germany

484 Walach, J., Filzmoser, P., Hron, K., Walczak, B., 2017. Robust biomarker identification
485 based on pairwise log-ratios. Chemon. Intell. Lab.Syst. 171, 277-285.
486 doi:10.1016/j.chemolab.2017.09.003

Figure legends

Figure 1. Scheme of the experimental design for the different fermentation processes performed in the work.

Figure 2. Tetrahedral plot and Principal Components' axes (PCs), according to inocula. The plot is based on the VOCs with the highest *clr* variances. 2-methyl-1-propanol (I); cis-2-penten-1-ol (O); 4-ethyl guaiacol (AP); and UF (BE). The symbol c stands for closure. T1, process inoculated with LPG1; T2, Lp13; T3, Lpl15; T4, Y12; T5, Y12 + LAB; T6, spontaneous.

Figure 3. CoDa-biplot of VOCs according to treatments. Projection onto the plane PC1 vs PC2. A) covariance biplot, and B) form biplot. Identification of the most relevant VOCs for the graph: D, 2-phenylethyl acetate; I, 2-methyl-1-propanol; J, 1-butanol; O, cis-2-penten-1-ol; Y, 2-ethenyl-2-butenal; AF, furfuryl methyl ether; AI, purpurocatechol; AP, 4-ethyl guaiacol; AT, 5-tert-butylpyrogallol; BB, UC(m/z 83-112-97; BD, UE (m/z 111-198; BE, UF (m/z 95-154-110; clr stands for *clr* transformation. For other relationships between CoDa symbols and VOCs, see Table S1.

Figure 4. CoDa dendrogram of VOCs, regardless of treatments. Balance sequences were built (until the 14^{th} balance) based on the progressive decreasing order of the *clr* variance. The complete set of sequential binary partitions is reported in Table S3.

Figure 5. Hierarchical clustering analysis based on A) the original data set, B) the proposed in this work *ilr coordinates*, C) *pivot* (a special case of *ilr*) *coordinates*, and D) the first 14^{th} *ilr coordinates* which accounted for 91.33% of the total variance. T1, inoculated with

LPG1; T2, Lp13; T3, Lpl15; T4, Y12; T5, sequential combination Y12+LAB; T6, spontaneous.

Figure 6. Projection of treatment scores onto the plane of the first two Factors. PCA analysis based on A) the VCOs expressed in their original units, B) the *clr coefficients* (central logratio transformation), C) the *irl coefficients* (isometric logratio transformation, and D) only the first 14th *ilr coordinates* (accounting for the 91.33% of the total variance).

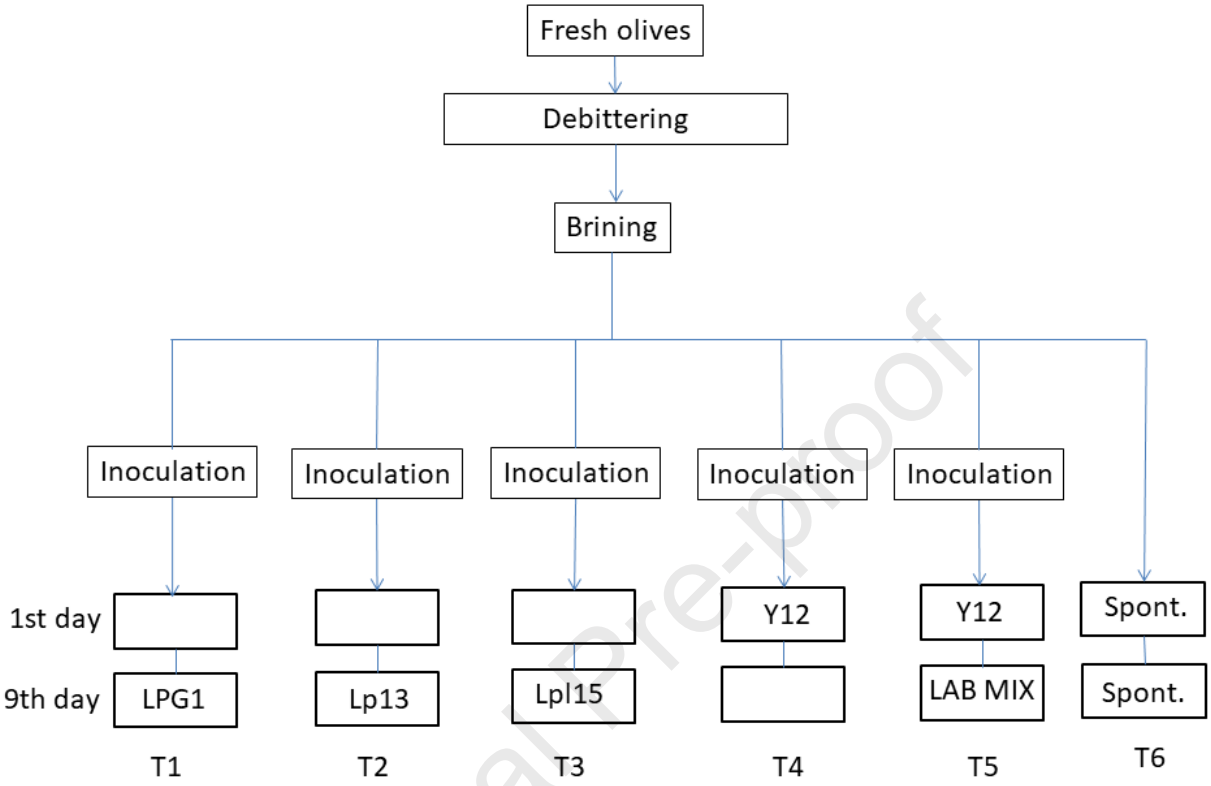
Figure 7. CoDa Q-clustering of the VOCs, based on the original data, using classical method (upper panel) and robust mode (bottom panel). Correspondence between symbols and compounds' names can be found in Table S1.

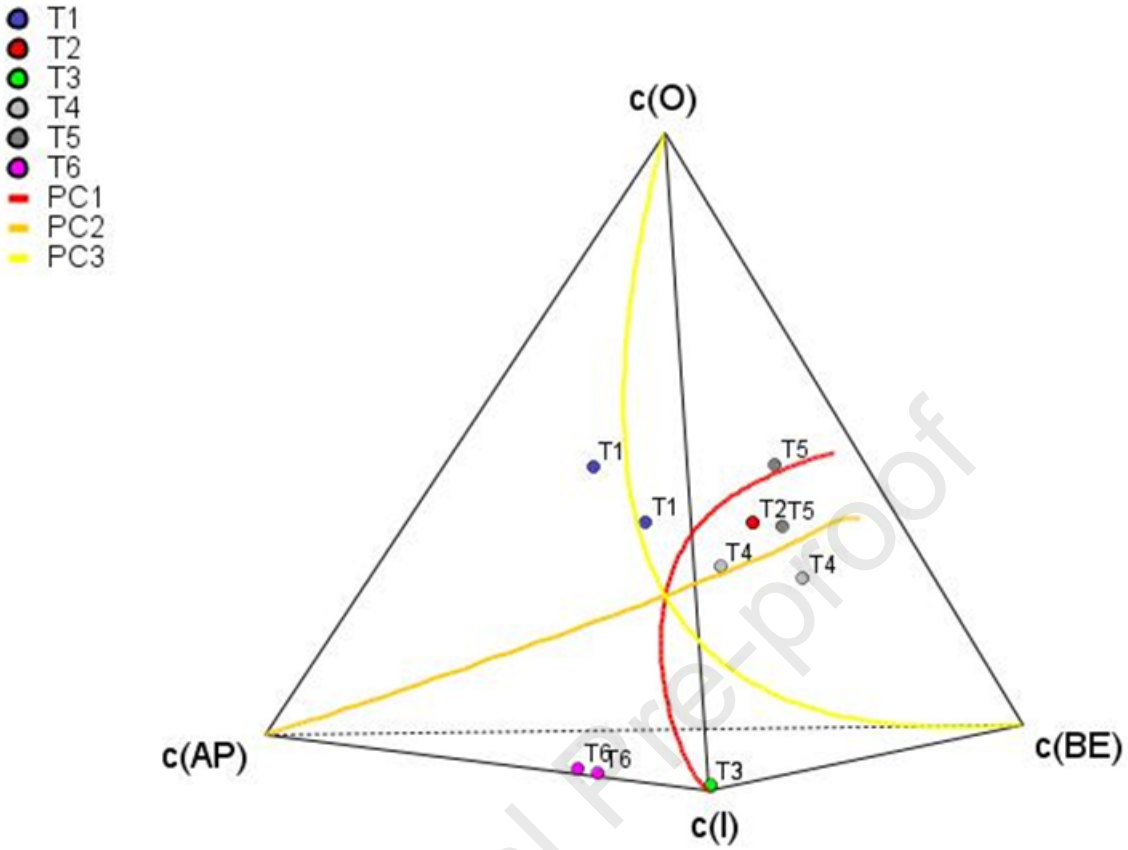
Figure 8. Relating starters with characteristics VOCs. Potential biomarkers revealed by V_j^*, using 1.96 as the cut-off limit (Walach et al., 2017). Case of T4 (Y12) vs T6 (spontaneous).

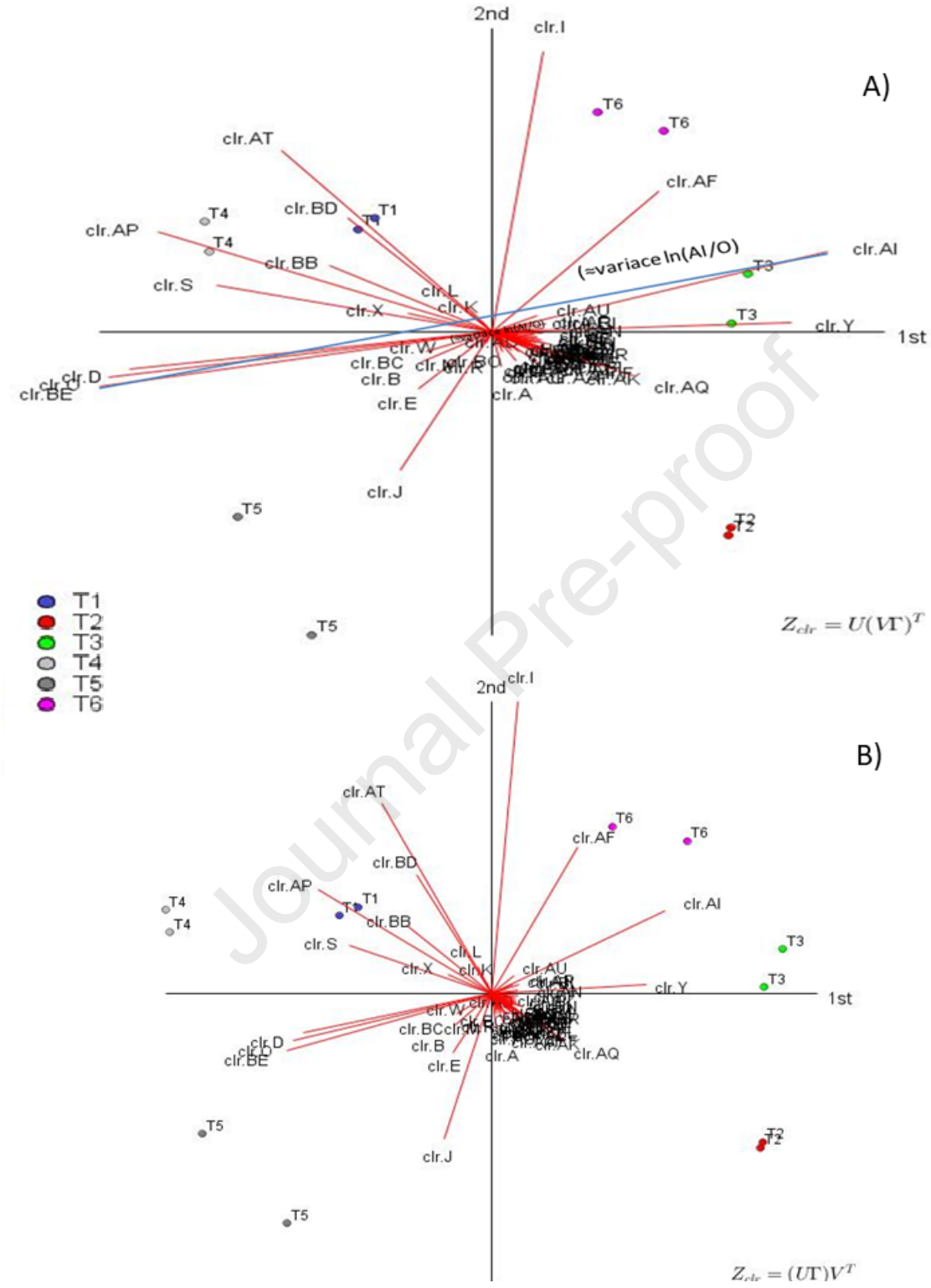
Table 1. Potential (significant) VOC markers for LPG1, Lp13, Lpl15, Y12, and Y12+LAB, using the V_j^* robust statistics (Walach et al., 2017).

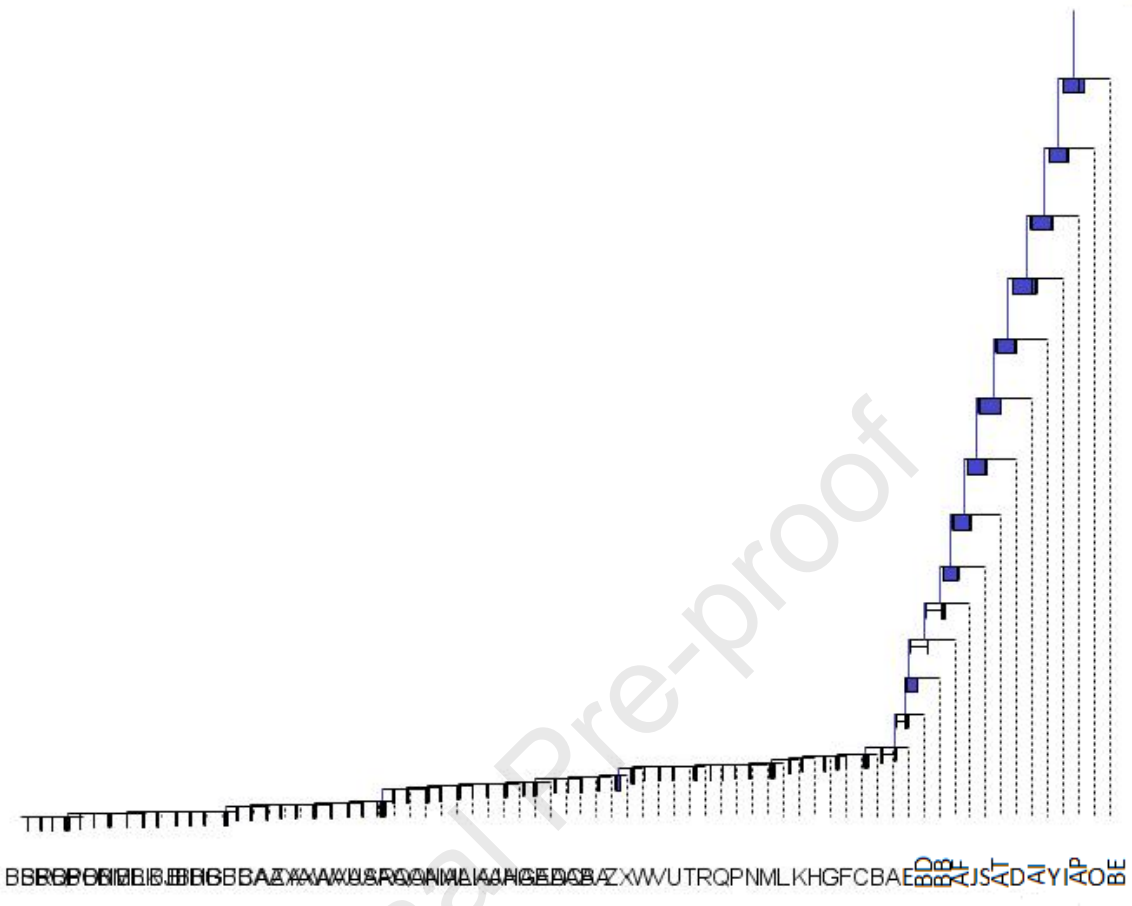
Index	Symbol in CoDa	Volatile compound	LPG1	Lp13	Lpl15	Y12	Y12+LAB
		Acetates					
4	D	2-Phenylethyl acetate	***			***	***
		Alcohols					
6	F	Methanol			***		
9	I	2-Methyl-1-propanol		***L			***L
10	J	1-Butanol	***	***	***	***	***
15	*O*	*cis*-2-Penten-1-ol	***			***	***
19	S	2-Methyl-3-hexanol	***			***	
24	X	2-Phenyl ethanol		***			
		Furans					
32	AF	Furfuryl methyl ether					***L
		Ketones					
35	AI	Purpurocatechol				***L	***L
		Phenols					
42	AP	4-Ethyl guaiacol		***L	***L		
43	AQ	4-Ethyl phenol			***L		
46	AT	5-*tert*-Butylpyrogallol		***L	***L		
		Non-identified					
54	BB	U C (m/z 83-112-97)				***	
56	BD	U E (m/z 111-198)		***L			
57	BE	U F (m/z 95-154-110)	***				***

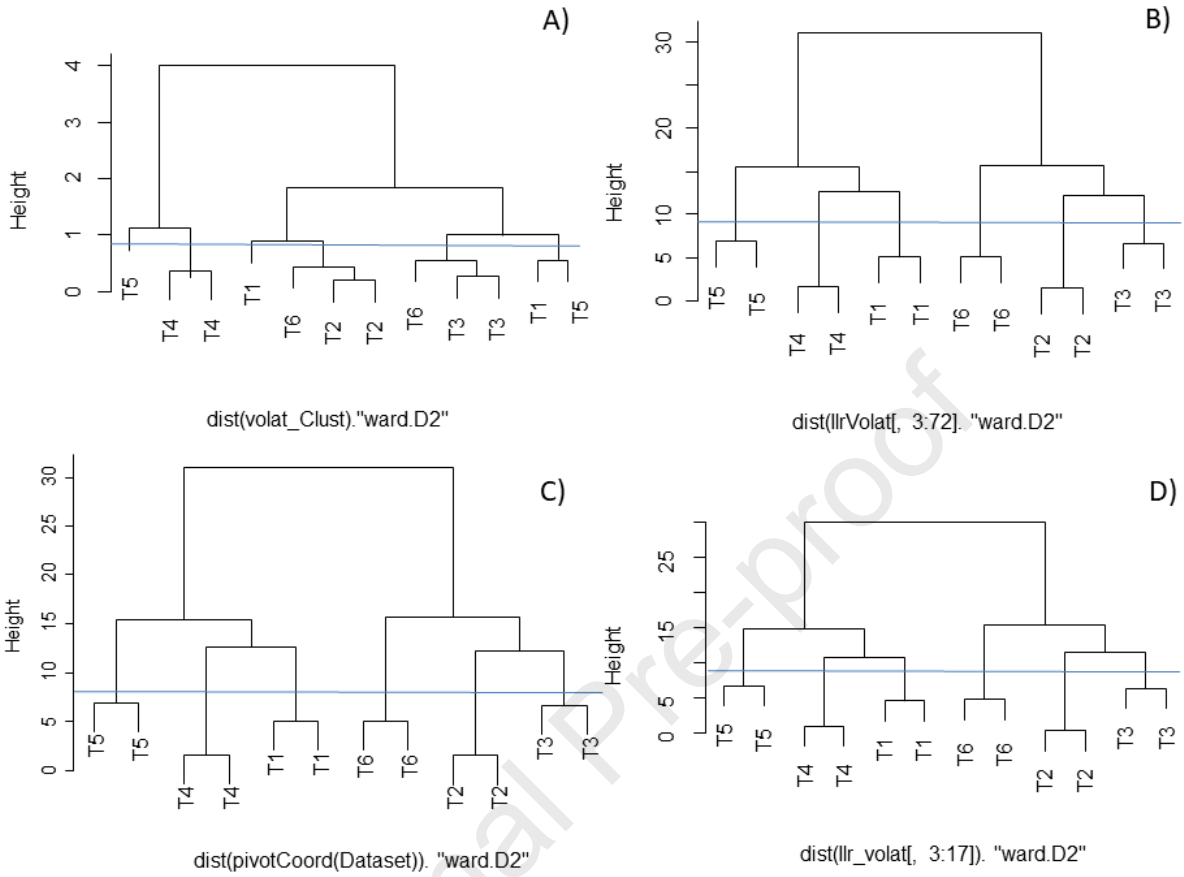
Notes: *** significant at p≤0.0; L, low/n.d. presence of a compound; U, unknown (that is, low probability of right identification according to NIST Mass Spectral Search Program).

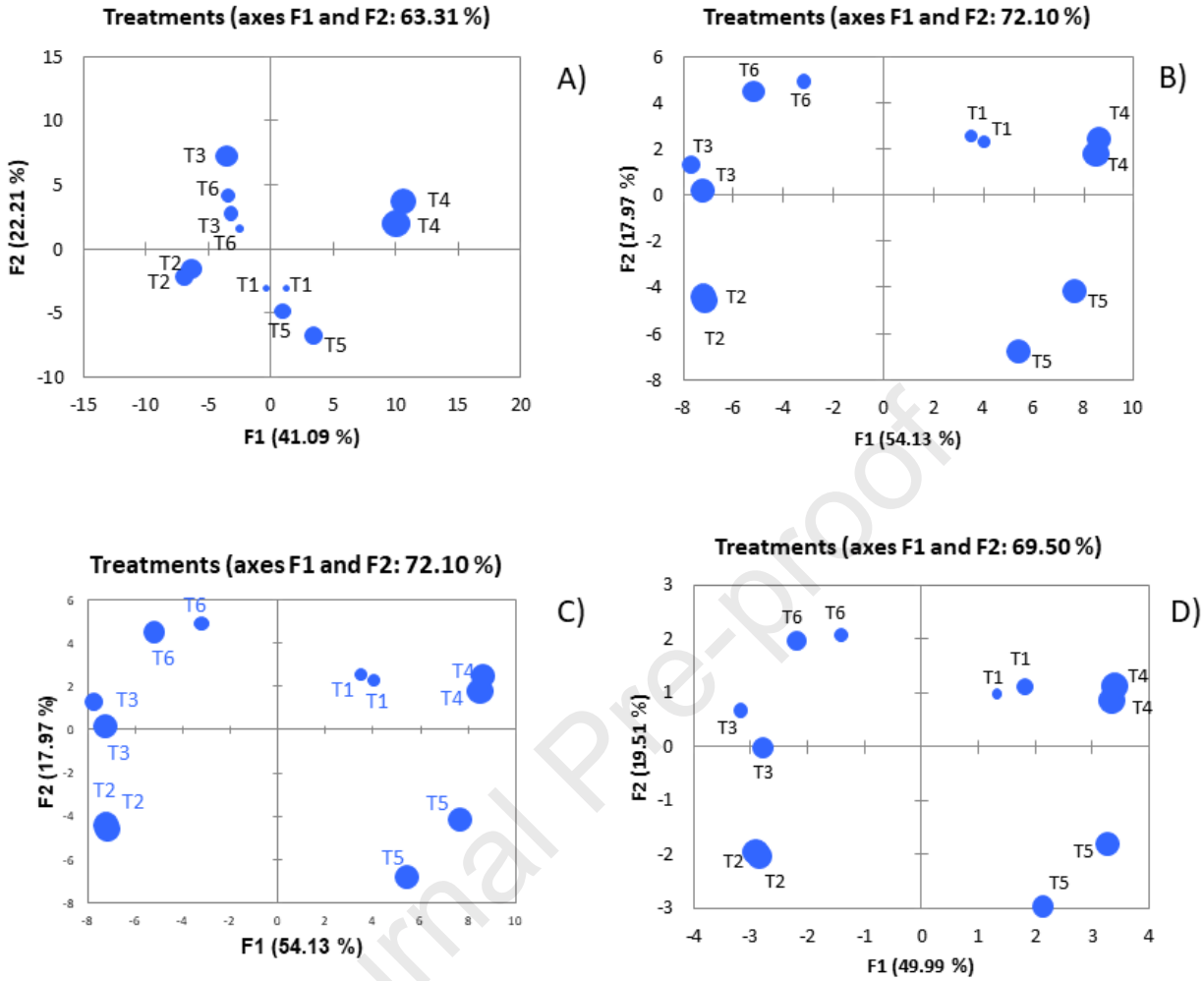


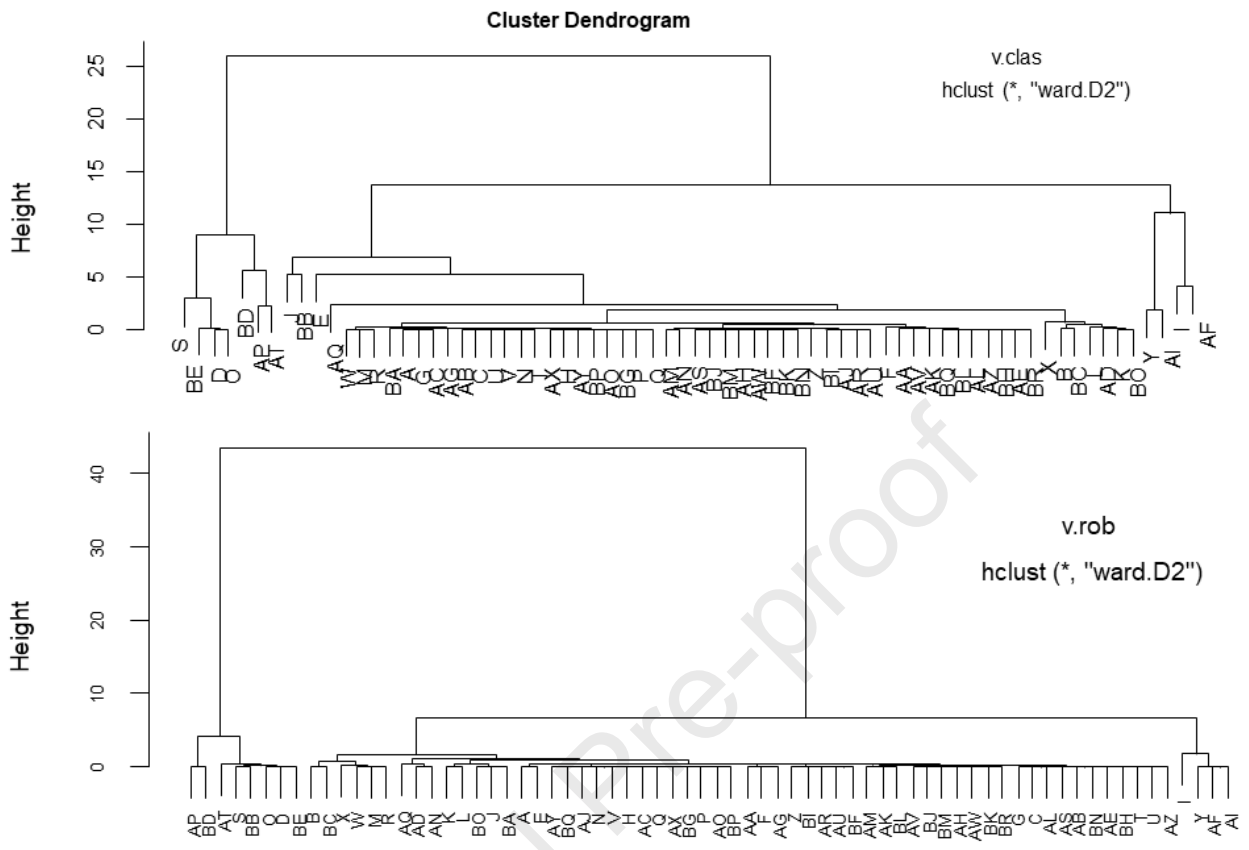


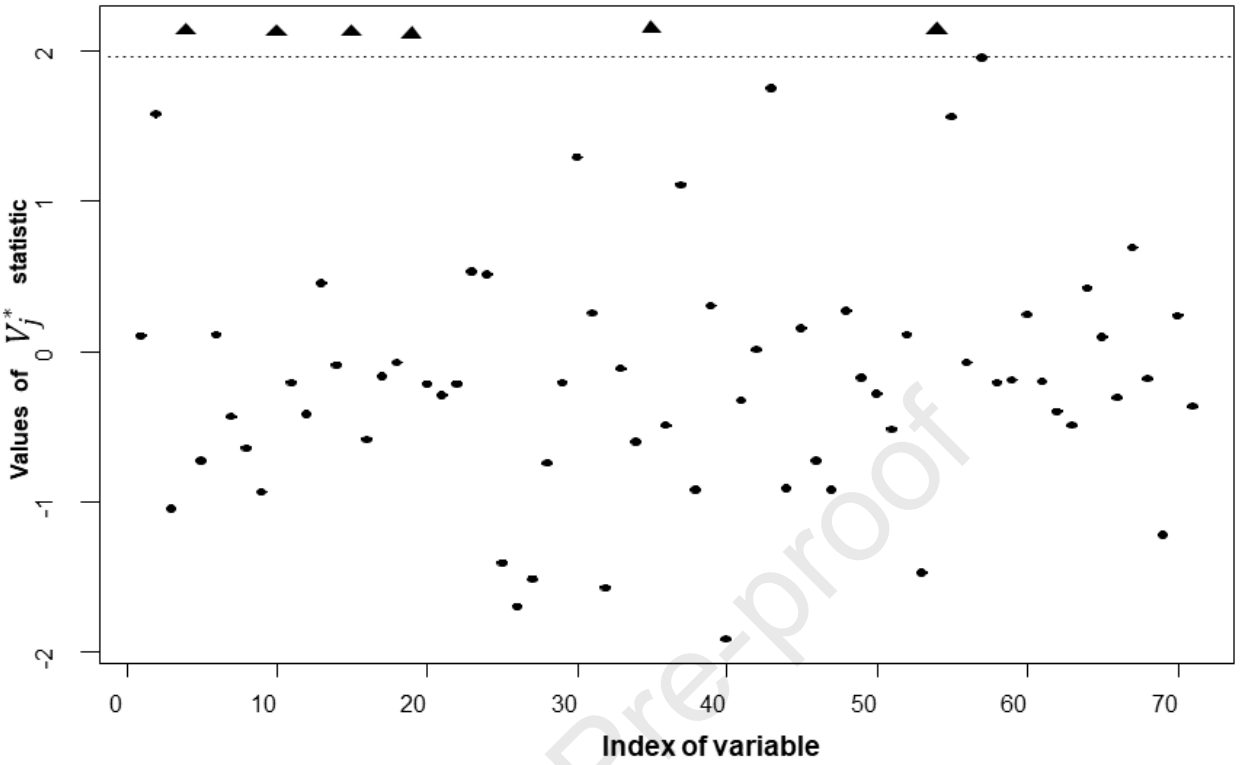












Highlights

-Microbial starters lead to different volatile profiles in concluded fermentations.

-Starters were better related to volatiles by CoDa analysis than by standard techniques.

-Strains were linked to characteristic volatiles and potential markers by CoDa tools.

-Relating starters and volatiles promotes sensory controlled table olive production.