Data in Brief 38 (2021) 107314

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Amplicon metabarcoding data of prokaryotes and eukaryotes present in 'Kalamata' table olives packaged under modified atmosphere

Sofia Michailidou[a], George Economou Petrovits[b], Mary Kyritsi[c], Anagnostis Argiriou[a,c,*]

[a] Institute of Applied Biosciences / CERTH, P.O. Box 60361, Thermi, Thessaloniki 57001, Greece
[b] Pelopac S.A., Block 38, NB1A Street, P.O. Box 1298, Thessaloniki Industrial Area, Sindos 57022, Greece
[c] Department of Food Science and Nutrition, University of the Aegean, Lemnos 81400, Greece

ARTICLE INFO

Article history:
Received 5 August 2021
Accepted 18 August 2021
Available online 20 August 2021

Keywords:
Metagenomics
Taxonomy
16S rRNA
18S rRNA
Olives
Modified atmosphere

ABSTRACT

Evaluation of food microbiome is of major importance since it accounts for the product's organoleptic characteristics and their nutritional value. In this dataset, microbes present in olive samples ('Kalamata' variety) stored under modified atmosphere and throughout different time-points of the shelf life of the product are presented, originated after 16S and 18S rRNA sequencing. The different time-points analyzed were: T0 (immediately after packaging), T6 (six months of storage), T12 (12 months of storage) and T18 (six months after the end of shelf life). Sequencing was performed on a MiSeq platform with the MiSeq Reagent Kit v3 (600 cycles). The raw sequence data used for analysis are available in NCBI under the Sequence Read Archive (SRA), with BioProject ID PRJNA688686. Raw reads were analyzed using the QIIME2 pipeline, clustered into Operational Taxonomic Units (OTU) and aligned against SILVA 132 reference database. OTUs are presented in different taxonomic levels for each time-point. These data present valuable information on the microbial communities of table olives, a dynamic niche that affect the final product quality. The data presented are related to the research article "Insights into the evolution of Greek style

DOI of original article: 10.1016/j.foodcont.2021.108286
* Corresponding author at: Institute of Applied Biosciences / CERTH, P.O. Box 60361, Thermi, Thessaloniki 57001, Greece.
E-mail address: argiriou@certh.gr (A. Argiriou).

https://doi.org/10.1016/j.dib.2021.107314

table olives microbiome stored under modified atmosphere: biochemical implications on the product quality" [1].

Specifications Table

Subject	Food Science: Food Microbiology
Specific subject area	Amplicon metabarcoding analysis of table olives
Type of data	Tables and Figures
How data were acquired	Sequencing of 16S rRNA and 18S rRNA genes was conducted on an Illumina MiSeq platform using the MiSeq® reagent kit v3 (paired end sequencing).
Data format	Raw and analyzed
Parameters for data collection	Operational Taxonomic Units (OTU) clustering analysis was conducted using the QIIME2 pipeline and the VSEARCH tool. Sequences were clustered into OTUs at 99% sequence similarity against the SILVA 132 database.
Description of data collection	Samples of 'Kalamata' olives packaged under modified atmosphere were analyzed throughout different time-points of storage (T0, T6, T12) and after the end of shelf life (T18).
Data source location	Institution: Institute of Applied Biosciences – Centre for Research and Technology Hellas City: Thessaloniki Country: Greece Latitude and longitude for collected samples/data: 40.56806, 22.99713
Data accessibility	Repository name: NCBI SRA Data identification number: PRJNA688686 Direct URL to data: http://www.ncbi.nlm.nih.gov/bioproject/688686, https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA688686%20
Related research article	S. Michailidou, F. Trikka, K. Pasentsis, G. Economou Petrovits, M. Kyritsi, A. Argiriou, Insights into the evolution of Greek style table olives microbiome stored under modified atmosphere: biochemical implications on the product quality, Food Control. In Press.

Value of the Data

- This dataset provides information on the microbiome present in olives packaged under modified atmosphere and monitors the changes that take place based on 16S and 18S rRNA amplicon sequencing.
- The data provide useful information on the microbial species present in food, thus, industry and other stakeholders can benefit from the identified microorganisms, by monitoring and evaluating the final product offered to consumers.
- This dataset can serve as a threshold for scientific community regarding the evolution of microorganisms in olive samples stored under modified atmosphere.

1. Data Description

The data reported here, refer to raw reads obtained after sequencing the V3-V4 hypervariable regions of the 16S rRNA gene and the V7-V8 hypervariable regions of the 18S rRNA gene. The raw sequence data (.fastq files) used for analysis are accessible through NCBI's Sequence Read Archive (SRA), under the BioProject ID PRJNA688686. For prokaryotes, sequencing resulted in 698,220 raw reads (Table 1). After quality and chimera filtering 294,978 reads were obtained for OTU clustering. Details on reads per time-point are presented in Table 1. Filtered reads were

Table 1
Number of raw reads, filtered reads and reads remained for OTU (Operational Taxonomic Unit) clustering, for each time point after 16S rRNA amplicon sequencing.

Sample ID	Total reads	Filtered reads	Reads remained for OTU clustering	Observed OTUs
T0	212,735	114,122	44,861	1246
T6	288,414	225,804	167,255	1426
T12	137,399	94,473	60,193	946
T18	59,672	31,966	22,669	540
Total	698,220	466,365	294,978	2715

Table 2
Number of raw reads, filtered reads and reads remained for OTU (Operational Taxonomic Unit) clustering, for each time point after 18S rRNA amplicon sequencing.

Sample ID	Total reads	Filtered reads	Reads remained for OTU clustering	Observed OTUs
T0	310,840	236,920	213,985	520
T6	270,284	155,574	138,854	713
T12	107,308	94,819	81,369	1071
T18	100,485	84,918	74,603	302
Total	788,917	572,231	508,811	1778

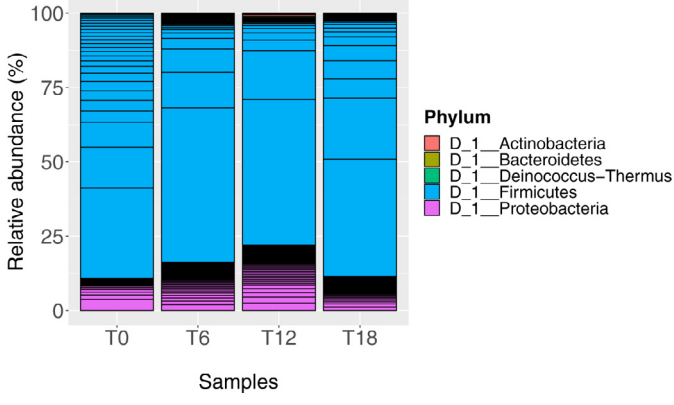

Fig. 1. Analysis of prokaryotic communities in olive samples throughout the different time-points. Distribution of the major Phyla. The scale in the y axis reflects the normalized relative abundance percentages (%). Black lines within each bar separate each Phylum into lower taxonomic levels.

clustered into 2715 unique OTUs. The number of unique OTUs was gradually decreased during storage, with T0 being the most enriched time-point in terms of bacterial diversity ($N = 1246$), whereas T18 was the least diverse, presenting 540 unique bacterial OTUs.

For eukaryotes, sequencing resulted in 788,917 raw reads, which after quality and chimera filtering were reduced to 572,231 and 508,811, respectively (Table 2). Assignment and clustering of these sequences against SILVA 132 database resulted in 1778 unique OTUs. Unlike prokaryotes, the number of observed OTUs was gradually increased from T0 ($N = 520$) to T12 ($N = 1072$), but at T18 a steep decrease was observed ($N = 302$).

Classification of the identified bacterial OTUs showed that Firmicutes was the dominant Phylum with a relative percentage above 71.8% in all time-points (Fig. 1). Proteobacteria, was the second major Phylum identified, with its relative percentages among time-points ranging from 12.84% (T0) to 25.83% (T12). At the Class level, Firmicutes were mainly classified as Bacilli, whereas Proteobacteria were mainly represented by Gammaproteobacteria (Fig. 2).

At the Order taxonomic level bacterial community of olive samples was mainly represented by Lactobacillales with relative percentages ranging from 50.0% at T12 to 65.8% at T0 (Fig. 3).

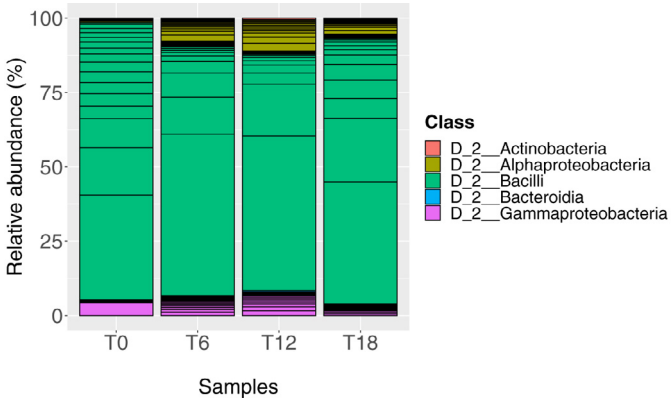

Fig. 2. Analysis of prokaryotic communities in olive samples throughout the different time-points at Class level. The scale in the y axis reflects the normalized relative abundance percentages (%). Black lines within each bar separate each class into lower taxonomic levels.

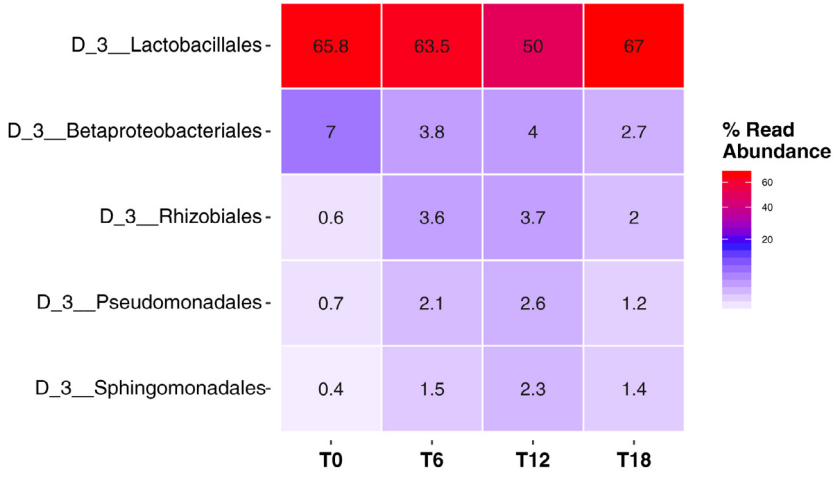

Fig. 3. Heatmap of relative abundances averaged for olive samples at Order level after 16S rRNA amplicon metabarcoding analysis.

At the onset of packaging (T0) bacterial profile was slightly differentiated from the subsequent time-points. In particular, although Betaproteobacteriales were found at 7.0% at T0, they were gradually displaced by Rhizobiales, Pseudomonadales and Sphingomonadales at the following time-points.

At the Family level, *Lactobacillaceae* dominated the bacterial communities with relative percentages being above 71.7% for all time-points (Fig. 4). *Burkholderiaceae* was found as the second most abundant family with relative percentages being on average 5.6%. Other families that were identified included *Sphingomonadaceae, Caulobacteraceae, Rhodobacteraceae, Stappiaceae, Moraxellaceae, Rhizobiaceae* and *Pseudomonadaceae*. Interestingly, *Bacillaceae* was only identified in T0.

Comparison of the different time points through Principal Component Analysis revealed that T0 demonstrates a well differentiated bacterial profile. The first two principal components (PC1+PC2) accounted for the 98.3%, capturing most of the total genetic variation (Fig. 5). T6 and T12 samples grouped together and in great distance from the other time points, while at the beginning of storage (T0) and after the end of product's shelf life (T18), samples were found

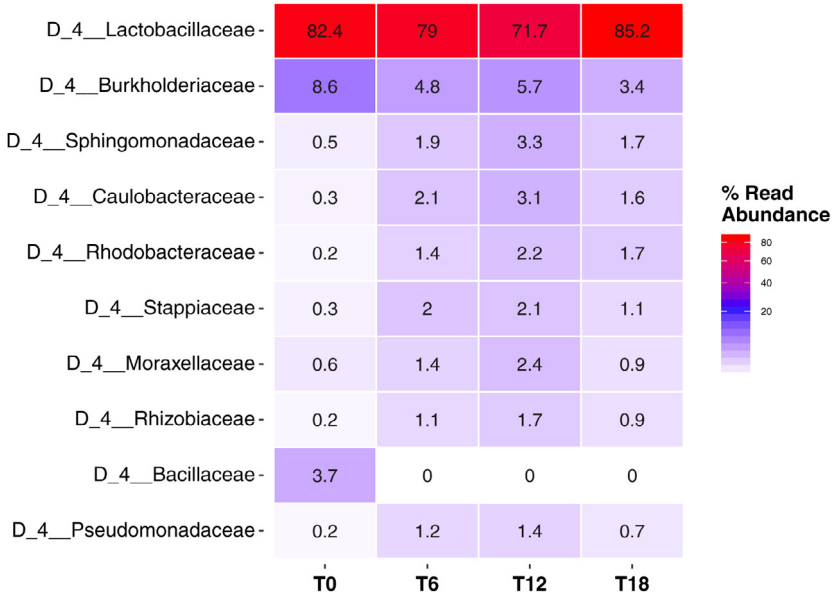

Fig. 4. Heatmap of relative abundances averaged for olive samples at Family level after 16S rRNA amplicon metabarcoding analysis.

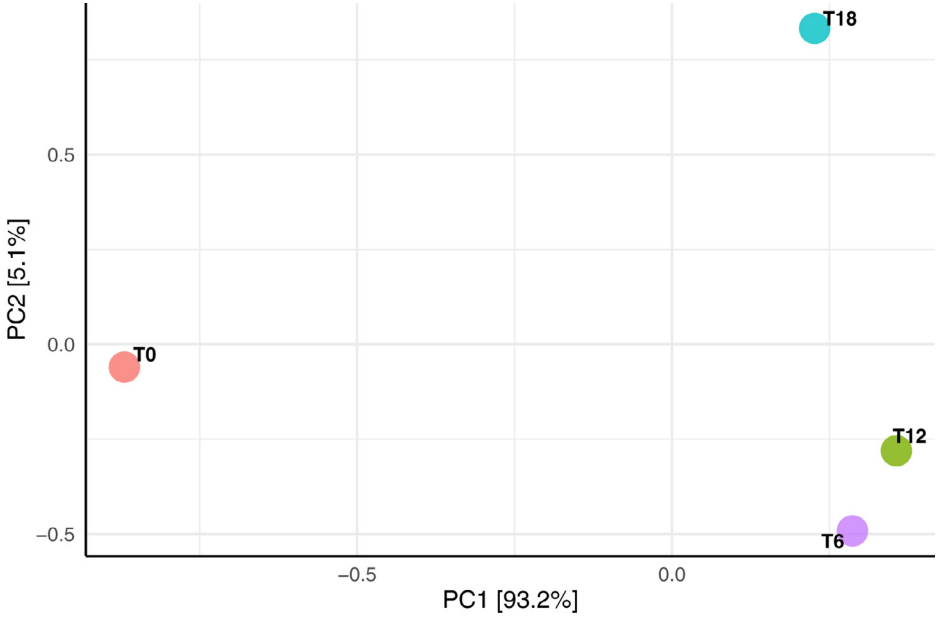

Fig. 5. Principal component analysis for the four time points (T0, T6, T12 and T18) for olive samples based on 16S rRNA amplicon sequencing.

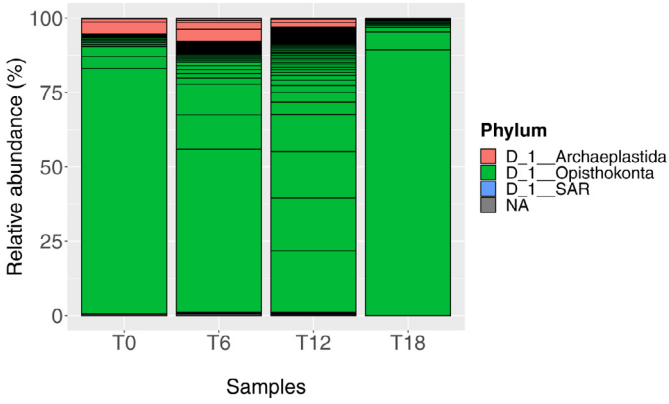

Fig. 6. Analysis of eukaryotic load in olive samples throughout the different time-points. Distribution of the major Phyla. The scale in the y axis reflects the normalized relative abundance percentages (%). Black lines within each bar separate each Phylum into lower taxonomic levels.

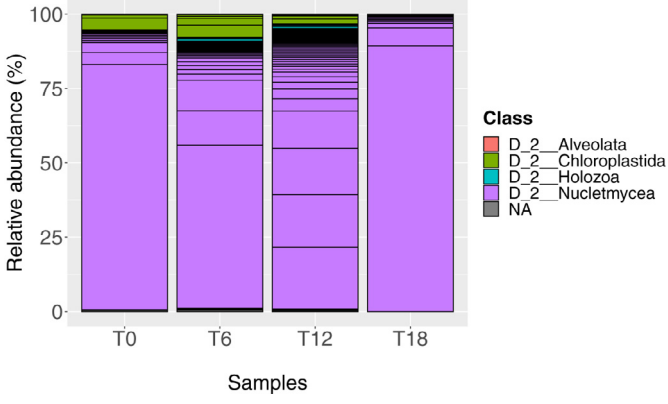

Fig. 7. Analysis of eukaryotic communities in olive samples throughout the different time-points at class level. The scale in the y axis reflects the normalized relative abundance percentages (%). Black lines within each bar separate each Class into lower taxonomic levels.

dispersed in axes, demonstrating a differentiated genetic profile based on the 16S rRNA amplicon sequencing.

Evaluation of eukaryotic OTUs revealed that Opisthokonta was the major Phylum identified in all samples with its relative abundance being over 99,67% across all time-points (Fig. 6). Likewise, Nucletmycea dominated the eukaryotic load at Class level (Fig. 7), whereas for the Order level, Fungi were identified as the dominant eukaryotes (Fig. 8).

At the Family level eukaryotes were mainly represented by members of the *Pichiaceae* family. In particular, from the onset of packaging (*Pichiaceae* at T0: 93.1%) a progressive decline on the relative abundance is observed until members of this family are found with a relative abundance of 36.1% at T12. After the end of shelf life (T18) *Pichiaceae* suddenly increase and dominate in MAP packaging with relative percentage of 93.4% (Fig. 9). Following *Pichiaceae, Saccharomycetaceae* is the second most abundant family with its members found at 4.7%, 13.1%, 16.4% and 6.6% for T0, T6, T12 and T18, respectively. At similar relative abundances to *Saccharomycetaceae*, members of *Cladosporiacea* family are present but only for T6 (11.8%) and T12 (19.6%). Overall, T0 and T18 show similar profile concerning the eukaryotic load of olive samples compared to T6 and T12. In particular, for T6 and T12 families that are present, but detected in traces at T0 and

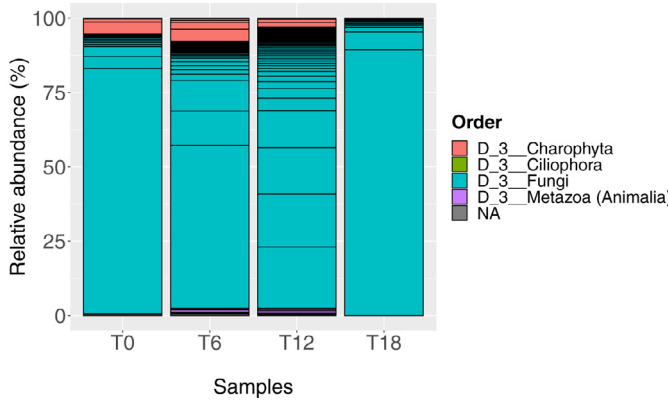

Fig. 8. Analysis of eukaryotic communities in olive samples throughout the different time-points at the level of Order. The scale in the y axis reflects the normalized relative abundance percentages (%). Black lines within each bar separate each Order into lower taxonomic levels.

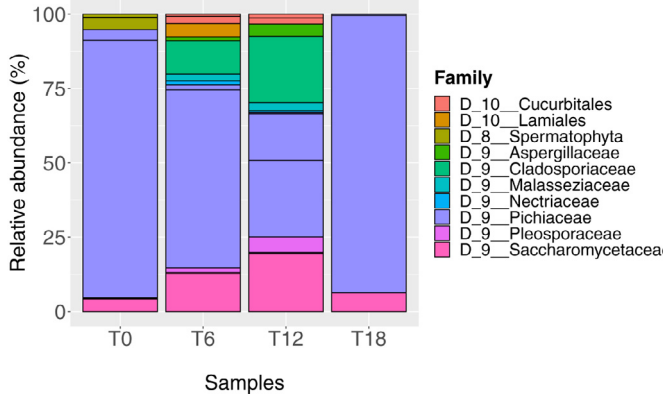

Fig. 9. Analysis of eukaryotic load in olive samples throughout the different time-points. Distribution of the major fungal families. The scale in the y axis reflects the normalized relative abundance percentages (%). Black lines within each bar separate each family into lower taxonomic levels.

T18, are *Aspergilaceae, Malasseziacea, Pleosporaceae* and *Nectriaceae*. In Fig. 9 Cucurbitales, Lamiales and Spermatophyta are presented, since they could not be classified to the Family level.

Genetic relatedness of samples through PCA based on the identified eukaryotic OTUs did not form any particular clusters among samples. The first two principal components (PC1+PC2) accounted for the 93.9%, capturing most of the total genetic variation (Fig. 10).

2. Experimental Design, Materials and Methods

2.1. Sample description and DNA extraction

In this dataset samples of 'Kalamata' table olives are presented packaged under modified atmosphere (30% CO_2 - 70% N_2). Samples were supplied by a private company (Pelopac S.A., Thessaloniki, Greece). Olives were left to ripe and darken naturally on the tree, and then they were harvested at the stage of full ripeness. 'Kalamata' table olives were further processed according to the Greek style; olives were fermented in their natural brine (natural fermentation). Although

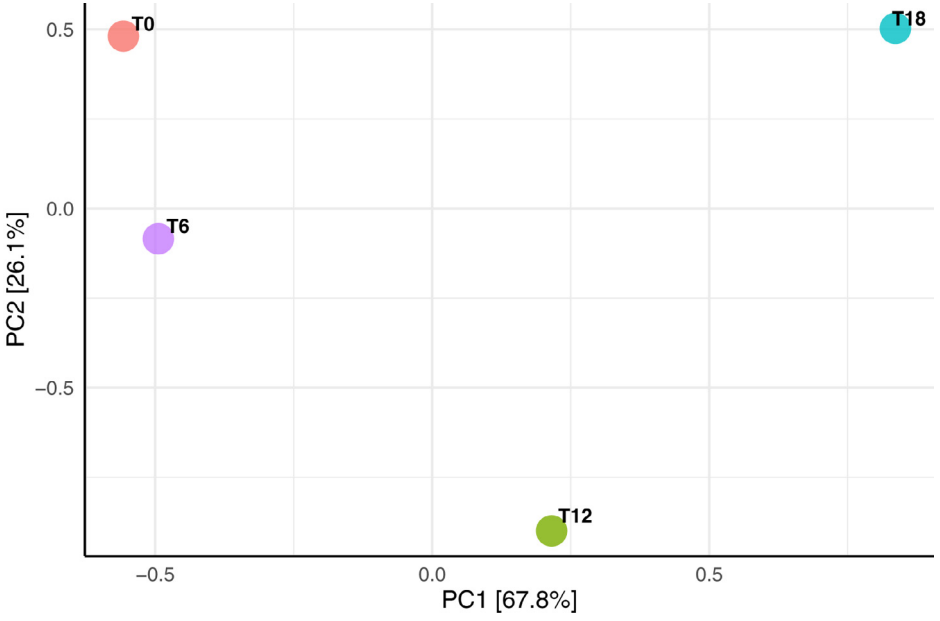

Fig. 10. Principal component analysis for the four time points (T0, T6, T12 and T18) based on 18S rRNA amplicon sequencing.

the shelf life of MAP pouches was 12 months, analysis was also performed six months after the end of the product's shelf life (18 months) to document changes in microbial communities after the expiration date. Hence, the four different time points of storage and analysis were: T0 = beginning, T6 = 6 months, T12 = 12 months and T18 = 18 months.

Upon arrival at the laboratory, samples were stored at 4 °C. From each pouch, 20 g of olive tissue were placed in sterile disposable plastic containers and homogenized in a polytron homogenizer (Polytron PT-MR 6100, Kinematica AG, Littau, Switzerland). To enable uniformity in stirring and obtain a homogenous pulp, apart from their natural brine, 5 ml of ddH_20 were also added to the container prior to homogenization. Microbial DNA was extracted from ~400 µl of the homogenized pulp, using the ZymoBIOMICS DNA Miniprep Kit (ZYMO RESEARCH; Irvine, CA, USA) according to the manufacturer's instructions. For sample disruption, the pulp was beaten using the ZR bashing beads provided by the kit, on a TissueLyser (Qiagen, Hilden, Geermany) for 5 min at 30Hz (1800 oscillations/minute). Elution was performed in the minimum volume allowed according to the protocol for highly concentrated DNA (V elution = 50 µl). DNA concentration was measured on a Qubit 4.0 Fluorimeter using the Qubit® dsDNA BR assay kit (Invitrogen, Carlsbad, CA, USA).

2.2. Library construction and amplicon sequencing

In line with the method previously described in Michailidou et al. [1], assessment of microbial species was conducted by amplifying different genes. For the identification of bacterial load the V3,V4 region of the 16S rRNA gene (≈460 bp) was amplified and sequenced, whereas fungal diversity was assessed by amplifying and sequencing the V7-V8 hypervariable regions of the 18S rRNA gene (≈350 bp). For the amplification of the 16S rRNA sequences, primers S-D-Bact-0341-b-S-17 and S-D-Bact-0785-a-A-21 were selected from on Klindworth et al. [2], whereas for the amplification of V7-V8 regions, universal primers FR1 and FF390 were selected from Chemidlin Prevost-Boure et al. [3]. For each primer used, an Illumina overhang adapter nucleotide sequence

was added at the 5' end of the selected primer. The sequences of the primers used were 16S_F: 5′- TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCT ACG GGN GGC WGC AG-3', 16S_R: 5′- GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGA CTA CHV GGG TAT CTA ATC C-3', 18S_F: 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CGA TAA CGA ACG AGA CCT-3' and 18S_R: 5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAN CCA TTC AAT CGG TAN T-3'. All libraries were constructed following the Illumina's 16S Metagenomic Sequencing Library Preparation (15044223 B) protocol with minor modifications. All PCR reactions were performed on a Rotor-Gene Q thermocycler (Qiagen) and monitored on real-time by adding in each sample a green fluorescent nucleic acid stain. Each PCR reaction was conducted in a final volume of 20 μl, containing 10 μl of 2x KAPA HiFi HotStart Ready enzyme mix (KAPA BIOSYSTEMS, Woburn, MA, U.S.A.), 0.80 μl of 10μM forward and reverse primer mix, 2.50 μl (∼5ng/μl) of microbial DNA, 0.25 μl 50 μM SYTO9 (Thermo Fisher Scientific, USA) and 6.45 μl ddH_2O. All PCR products and libraries were purified to remove unincorporated primers and primer-dimer species using NucleoMag® NGS Bead Suspension (Macherey-Nagel, Düren, Germany) using different ratios of beads (ratio = volume of paramagnetic beads to PRC volume), depending on the target region size. Libraries were initially quantified with a fluorometric quantification using Qubit® dsDNA BR assay kit and their quality was automatically assessed on a Fragment Analyzer system (Agilent Technologies Inc. Santa Clara, United States) using the DNF-477-0500 kit. The final molarity of libraries was evaluated by a quantitative PCR (qPCR), conducted on a Rotor-Gene Q thermocycler (Qiagen, Hilden, Germany) with the KAPA Library Quantification kit for Illumina sequencing platforms (KAPA BIOSYSTEMS, Woburn, MA, U.S.A.). For library quantification, each sample was analyzed in triplicates. Final molarity of libraries was calculated in relation to the size of DNA amplicons after indexing, based on the following equation:

$$C = \text{pM after qPCR} * \frac{452\ (\text{size of DNA standard in bp})}{\text{Average fragment length of library (bp)}} * \text{relevant dilution factor}$$

Finally, 16S and 18S rRNA libraries were pooled at 12.5pM, mixed at equal percentages and sequenced on a MiSeq platform using the MiSeq® reagent kit v3 (2 × 300 cycles) (Illumina, San Diego, California).

2.3. Bioinformatics and data analysis

Raw sequences (fastq files) were analyzed using the Quantitative Insights into Microbial Ecology 2 (QIIME2) pipeline [4]. Analyzes were implemented on a Linux/based HPC cluster assigning one node with 32 cores and 256 GB RAM. Adapters were trimmed from raw sequences using cutadapt plugin [5] with the *trim-paired* function, joined using the *join-pairs* function and filtered with the q*uality-filter q-score-joined* command with minimum quality score 28 (*p-min-quality*). Dereplication of sequences was performed with the *dereplicate-sequences* command and sequences were clustered into operational taxonomic units (OTUs) at 99% sequence similarity, using the open-reference method and the VSEARCH tool [6]. In addition, chimeras and "borderline chimeras" were excluded from downstream analysis, by applying the *uchime-denovo* tool and sequences were further aligned against the SILVA 132 reference database [7]. Taxonomy classification was performed with the *feature-classifier* plugin using the *classify-consensus-blast* command with a percent identity threshold of 0.99. Finally, from the resulted OTU table Archaea and chloroplastic or mitochondrial sequences were excluded from downstream analyzes.

Further analysis and data visualization was performed in R version 3.6.0 [8]. OTU tables and .biom files were imported and merged in R environment using the *import_biom* command of the Phyloseq R package [9]. Ampvis2 [10] and ggplot2 [11] R packages were used to visualize OTU abundances as barplots; phyloseq objects were sorted with the *sort* function, pruned for top OTUs or the desired taxonomy with the *prune_taxa* function and transformed to percentages with the *transform_sample_counts* function. Heatmaps and Principal Component Analyzes (PCA) were conducted using the *amp_heatmap* and amp_*ordinate* commands of ampvis2 package, respectively, by applying the Euclidean distance method.

Ethics Statement

The work did not involve the use of human subjects, animals, cell lines and endangered species of wild fauna and flora.

CRediT Author Statement

Sofia Michailidou: Conceptualization, Methodology, Software, Writing – review & editing; **George Economou Petrovits:** Resources, Funding acquisition; **Mary Kyritsi:** Resources, Funding acquisition; **Anagnostis Argiriou:** Conceptualization, Funding acquisition, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

This research received funding from the Emblematic Action Olive Roads from the General Secretariat for Research and Innovation (GSRI) under Grant Number 2018ΣΕ01300000.

References

[1] S. Michailidou, F. Trikka, K. Pasentsis, G.E. Petrovits, M. Kyritsi, A. Argiriou, Insights into the evolution of Greek style table olives microbiome stored under modified atmosphere: biochemical implications on the product quality, Food Control 130 (2021), doi:10.1016/j.foodcont.2021.108286.
[2] A. Klindworth, E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, F.O. Glöckner, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies, Nucleic Acids Res. 41 (2013), doi:10.1093/nar/gks808.
[3] N. Chemidlin Prévost-Bouré, R. Christen, S. Dequiedt, C. Mougel, M. Lelièvre, C. Jolivet, H.R. Shahbazkia, L. Guillou, D. Arrouays, L. Ranjard, Validation and application of a PCR primer set to quantify fungal communities in the soil environment by real-time quantitative PCR, PLoS One 6 (9) (2011), doi:10.1371/journal.pone.0024166.
[4] E. Bolyen, J.R. Rideout, M.R. Dillon, N.A. Bokulich, C.C. Abnet, G.A. Al-Ghalith, H. Alexander, E.J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J.E. Bisanz, K. Bittinger, A. Brejnrod, C.J. Brislawn, C.T. Brown, B.J. Callahan, A.M. Caraballo-Rodríguez, J. Chase, J.G. Caporaso, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, Nat. Biotechnol. 37 (2019) 852–857, doi:10.1038/s41587-019-0209-9.
[5] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (1) (2011), doi:10.14806/ej.17.1.200.
[6] T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics, PeerJ 10 (2016), doi:10.7717/peerj.2584.
[7] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F.O. Glöckner, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Res. 41 (2013), doi:10.1093/nar/gks1219.
[8] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018 Available online athttps://www.R-project.org/.
[9] P.J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data, PLOS One 8 (2013) e61217, doi:10.1371/journal.pone.0061217.
[10] M. Albertsen, S.M. Karst, A.S. Ziegler, R.H. Kirkegaard, P.H. Nielsen, Back to basics - the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities, PLoS One 10 (2015), doi:10.1371/journal.pone.0132783.
[11] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, 2016.