UNIVERSIDADE DE SÃO PAULO

FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

Metagenomic prospection of quorum sensing related bacteria during spontaneous cocoa beans fermentation

Prospecção metagenômica de bactérias relacionadas a quorum sensing durante a fermentação espontânea do cacau

Otávio Guilherme Gonçalves de Almeida

Ribeirão
Preto
2022

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

Otávio Guilherme Gonçalves de Almeida

Metagenomic prospection of quorum sensing related bacteria during spontaneous cocoa beans fermentation

Prospecção metagenômica de bactérias relacionadas a quorum sensing durante a fermentação espontânea do cacau

Doctoral thesis presented to the Graduate Program of Biosciences and Biotechnology of the School of Pharmaceutical Sciences of Ribeirão Preto/USP for the degree of Doctor in Sciences.

Concentration Area: Bioagents and Biotechnology Applied to Pharmacy

Supervisor: Prof. Dr. Elaine Cristina Pereira De Martinis

Versão corrigida da Dissertação de Mestrado ou Tese de Doutorado apresentada ao Programa de Pós-Graduação em Biociências e Biotecnologia em 31/08/2022. A versão original encontra-se disponível na Faculdade de Ciências Farmacêuticas de Ribeirão Preto/USP.

Ribeirão Preto

2022

Almeida, Otávio Guilherme Gonçalves de

Metagenomic prospection of quorum sensing related bacteria during spontaneous cocoa beans fermentation. Ribeirão Preto, 2022.
170p.: il.; 30cm.

Doctoral thesis presented to the Graduate Program of Biosciences and Biotechnology of the School of Pharmaceutical Sciences of Ribeirão Preto/USP for the degree of Doctor in Sciences. Concentration Area: Bioagents and Biotechnology Applied to Pharmacy

Supervisor: De Martinis, Elaine Cristina Pereira.

1. Cocoa Fermentation 2. Metagenomics 3. Quorum sensing 4. *luxS* 5. Lactic acid bacteria

APPROVAL PAGE

Name: Otávio Guilherme Gonçalves de Almeida

Title: Metagenomic prospection of quorum sensing related bacteria during spontaneous cocoa beans fermentation

Doctoral thesis presented to the Graduate Program of Biosciences and Biotechnology School of Pharmaceutical Sciences of Ribeirão Preto/USP for the degree of Doctor in Sciences.

Concentration Area: Bioagents and Biotechnology Applied to Pharmacy

Supervisor: Prof. Dr. Elaine Cristina Pereira De Martinis

Approved on:

Examiners

Prof. Dr. ____

Institution: _____________________________ Signature:____________________

Prof. Dr. ____

Institution: _____________________________ Signature:____________________

Prof. Dr. ____

Institution: _____________________________ Signature:____________________

Prof. Dr. ____

Institution: _____________________________ Signature:____________________

Prof. Dr. ____

Institution: _____________________________ Signature:_____________________

Prof. Dr. ____

Institution: _____________________________ Signature:_____________________

“I dedicate this work to my family, my teachers, Brazilian society and my childhood inspirations”.

Acknowledgments

I always will be grateful to

My wife Tamy, my mother Lilian, and my aunt Leda for the strong support I've received in my life.

The University of São Paulo.

The Faculty of Pharmaceutical Sciences of Ribeirão Preto.

To *Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC)* and *Centro de Pesquisas do Cacau (CEPEC)* of Brazilian Ministry of Agriculture represented by Carlyle Brito Matos for enabling field experimentations.

The São Paulo Research Foundation (FAPESP) for the doctoral fellowship (Process #2017/13759-6), as well as for the funding of a fellowship abroad carried out in Italy (Process 2018/26719-5).

The Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) which funded partially this Project (Finance Code 001).

Professor Elaine Cristina Pereira De Martinis for the support along my studies, incentive, and committed supervision.

Professor Wilson Araújo da Silva Jr. from Faculdade de Medicina de Ribeirão Preto (FMRP) for his support in providing bioinformatics facility.

Kamila Peronni Chagas for her contribution with Next-Generation Sequencing and shotgun metagenomics.

Marita Gimenez Pereira for scientific discussions and collaboration in lab experiments.

The staffs Marília, Jefferson, and Emerson.

Professor Giovanna Felis (*Università Degli Studi di Verona*) for committed supervision abroad.

My friends Diego Frazilio, João Pedro Furlan, Rayana Bighetti-Trevisan, Gabriela Batitucci for fruitful collaborations and high-level scientific discussions.

To Italian friends Eleonora Troiano, Renato L. Binatti, Veronica Gatto, Giulia Bertazzoli, and professor Sandra Torriani for the amazing professional experience.

“My Advice is: Don’t ask for advice”

Ada Yonath, Nobel Prize in Chemistry (2009)

Resumo

Almeida, O. G. G. **Prospecção metagenômica de bactérias relacionadas a quorum sensing durante a fermentação espontânea do cacau**. 2022. 170f. Tese (Doutorado). Faculdade de Ciências Farmacêuticas de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2022.

Para obtenção de chocolate de alta qualidade é necessário que a matéria-prima, isto é, as sementes de cacau, sejam fermentadas. Os microrganismos infiltram-se após o corte dos frutos de cacau e contaminam as sementes estéreis. Como a polpa mucilaginosa é rica em nutrientes e apresenta alta atividade de água, os microrganismos encontram condições propícias para sobrevivência e multiplicação. Destacam-se as leveduras, bactérias láticas (BAL) e bactérias acéticas (BAA), que nesta ordem, realizam uma sucessão microbiana bem definida ao longo do processo fermentativo. Esse processo permite a drenagem da polpa mucilaginosa que recobre as sementes e, ao mesmo tempo, estimula as hidrolases endógenas a metabolizar os substratos de reserva armazenados nas sementes. A atividade microbiana somada à atividade metabólica interna das sementes resulta na formação de precursores característicos do sabor e aroma do chocolate. Muitos autores consideram a alimentação cruzada como o fator determinante para a sucessão microbiana, pois as leveduras despectinizam a polpa liberando açúcares que podem ser metabolizados pelas BAL, em seguida as BAL convertem os açúcares em etanol, manitol e ácido acético. O manitol é metabolizado pelas BAA como fonte de carbono, resultando em ácido acético. Embora a alimentação cruzada explique a sucessão microbiana, do ponto de vista de ecologia de comunidades, pouco se conhece sobre as interações que ocorrem ao longo do processo fermentativo, como o quorum sensing (QS), por exemplo. O QS é um processo de sincronização da expressão gênica por meio da liberação de moléculas autoindutoras em altas densidades celulares. Dessa maneira, visto que a fermentação de cacau apresenta condições para que os microrganismos atinjam altas populações e que estudos têm indicado que bactérias relacionadas ao QS podem apresentar vantagem competitiva e adaptativa a ambientes estressantes, este trabalho teve por objetivo, inicialmente, identificar por meio da metagenômica, a microbiota relacionada com QS por meio da identificação in situ do gene *luxS* ao longo da fermentação e em genomas de bactérias isoladas de fermentação de cacau. O gene *luxS* é tido como um marcador universal de QS, pois caracteriza a comunicação interespecífica célula a célula. No primeiro Capítulo deste trabalho é apresentada a análise metagenômica de uma fermentação espontânea de cacau amostrada por até 144h. Os dados obtidos revelaram que os fungos estavam presentes ao longo de todo o processo fermentativo. Além disso, demonstram também que *reads* relacionados ao gene *luxS* são enriquecidos à medida que aumenta o tempo de fermentação, atingindo um pico máximo de detecção no tempo de 72h de fermentação. Foi também observado que os gêneros *Enterobacter*, "*Lactobacillus*", *Bacillus* e *Pantoea* foram associados aos genes *luxS*, evidenciando as unidades taxonômicas operacionais relacionadas a esse gene. No segundo Capítulo é apresentado um estudo genômico comparativo no qual três cepas de *Lactiplantibacillus plantarum* Lb2, *Limosilactobacillus fermentum* Lb1 e *Pediococcus acidilactici* P1 isoladas de fermentação espontânea de cacau tiveram seus genomas sequenciados e comparados contra todos os genomas públicos pertencentes a essas espécies. Os resultados mostraram que o gene *luxS* estava presente em todas as cepas dessas espécies e que *Lp. plantarum* apresentava seis clusters gênicos para *luxS*, evidenciando um alto número de cópias. Para *Lm. fermentum* foram observados apenas dois clusters gênicos e em *P. acidilactici* um cluster gênico. Análises filogenéticas sugeriram que o segundo cluster do gene *luxS* (denominado luxS_2) foi transferido horizontalmente via transdução de *Lp. plantarum* para *Lm. fermentum*, visto que ambas apresentaram o mesmo cluster e a região flanqueadora desse cluster em *Lm. fermentum* era composta de transposases IS30. Adicionalmente, após a investigação da presença desse gene nessas espécies, primers espécie- e clado-específicos foram desenhados para *screening* rápido de cepas para avaliação do potencial relacionado ao QS e para aplicação em estudos envolvendo qRT-PCR. No terceiro Capítulo é apresentado um estudo comparativo genômico para as BAA isoladas de processo fermentativo espontâneo de cacau, cujas cepas de *Acetobacter senegalensis* MRS7, GYC10, GYC12, GYC19 e GYC27 (com genótipos distintos), tiveram seus genomas sequenciados e comparados com genomas publicamente disponíveis de *A. senegalensis*. O estudo demonstrou que em *A. senegalensis* não havia genes luxS, mas sim genes relacionados ao QS intraespecífico como acil homoserina lactonas (AHLs) e *response regulators*, bem como genes envolvidos com a inibição de QS, codificando para acilases e lactonases, o que corroborou as predições realizadas no trabalho apresentado no primeiro capítulo desta tese. Além disso, visto que as BAA apresentam potencial de aplicação em outros processos industriais, como a produção de celulose e fermentação de vinagre, vias metabólicas envolvidas na adaptação bacteriana a processos fermentativos sob condições de estresse foram também investigadas. Os dados revelaram que essas cepas apresentam um bom potencial de adaptação em ambientes estressantes, com capacidade de sintetizar chaperonas, álcool desidrogenases e proteínas ABC, que conferem tolerância a altas concentrações de etanol e a altas temperaturas. Adicionalmente, as cepas de AAB não apresentaram potencial patogênico e nem genes de resistência a antibióticos. Dessa maneira, os resultados obtidos sugerem que as cepas de *A. senegalensis* isoladas da fermentação espontânea de cacau podem ser aplicadas também em outros processos industriais. Por fim, no quarto Capítulo há o desfecho da tese, com a apresentação de um artigo submetido para publicação, o qual trata da detecção in situ do gene *luxS* em fermentações conduzidas em escala de laboratório. Dessa maneira, seis fermentações foram realizadas em duplicata, sendo denominadas F1, F2 e F3, as quais foram inoculadas com diferentes combinações de coquetéis contendo leveduras, BAL e/ou BAA, com o objetivo de comparar a expressão do gene *luxS* ao longo de 96h de fermentação. Para comparação, foi preparado um controle não inoculado (sem inoculação de microrganismos, sem replicata - denominada F0). Para cada uma das condições estudadas, foram realizadas análises relativas à enumeração de microrganismos, medida de pH e da temperatura de fermentação, bem como análises de metataxonômica (*16S rRNA* e *ITS*), mensuração de atividade enzimática e detecção de metabolites voláteis (VOCs) por cromatografia gasosa acoplada a espectrometria de massas (GC-MS). Os resultados foram analisados para avaliar o potencial funcional microbiano relativo a QS e à qualidade do cacau fermentado. Os resultados revelaram que ocorreu uma sucessão microbiana típica da fermentação de cacau, que foi corroborado pelas análises clássicas de enumeração de microrganismos e de metataxonômica. No entanto, estatisticamente não houve diferenças significativas nas populações microbianas enumeradas na fermentação F0 e nas demais ($p > 0,05$), o que também foi reforçado pela ausência de diferença significativa entre os valores de α-diversidade para as diferentes fermentações, determinada por análises metataxonômica. Não houve correlação significativamente estatística entre atividade enzimática e alterações composicionais da microbiota total, indicando que, provavelmente, a principal atividade enzimática era decorrente das próprias hidrolases endógenas das sementes e não da microbiota per se. Em relação à mensuração do gene *luxS*, para as espécies *Lp. plantarum* houve atividade ao longo da fermentação, enquanto que a expressão do gene *luxS* para *Lm. fermentum* foi detectada apenas nas primeiras 72h. A correlação entre a qualidade da fermentação expressa por meio da quantidade de VOCs detectados e o padrão de expressão dos genes *luxS* evidenciou uma associação positiva entre a expressão desse gene por *Lp. plantarum* e características sensoriais indesejáveis para as sementes fermentadas. Considerando os resultados deste trabalho, é possível afirmar que o gene *luxS* foi ativo ao longo da fermentação de cacau, conforme hipótese do primeiro Capítulo, e houve avanço na compreensão da dinâmica da fermentação de cacau, com demonstração da participação de diferentes enzimas ao longo do processo. Além disso, os resultados dessa pesquisa indicam que *Lp. plantarum* é um protagonista neste tipo de fermentação, com grande potencial de expressão de vias de QS. Futuros estudos poderão elucidar em condições de campo os dados levantados ao longo deste trabalho.

Palavras-chave: Fermentação de cacau, metagenômica, quorum sensing, *luxS*, bactérias láticas.

Abstract

Almeida, O. G. G. **Metagenomic prospection of quorum sensing related bacteria during spontaneous cocoa beans fermentation**. 2022. 170f. Thesis (Doctoral). Faculdade de Ciências Farmacêuticas de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2022.

In order to obtain processed chocolate with high standards it is mandatory the fermentation of cocoa seeds, which are the raw material for chocolate production. After cutting of cocoa fruits, the autochthonous microbiota infiltrates, and enters in contact with the seeds, contaminating them. As the seeds are surrounded by a mucilaginous pulp, which is rich in nutrients and presents high water activity, microorganisms find adequate conditions to proliferate and growth. Among the possible microorganisms related to cocoa fermentation, yeast, lactic acid bacteria (LAB), and acetic acid bacteria (AAB) stand out, as these groups are related to a well-defined microbial succession along cocoa fermentation. This process allows the pulp drainage and stimulates the endogenous hydrolases to metabolize the stored substrates in the seeds. The microbial activity in consonance with the intrinsic metabolic activity of the seeds leads to the releasing of chocolate's flavour precursors. Many authors attribute to cross feeding the main role guiding and shaping microbial succession, since the yeasts depectinize the pulp releasing sugars that can be metabolized by LAB. The LAB convert these sugars into lactic acid, mannitol, and acetic acid. Mannitol is metabolized by AAB as carbon source, resulting in more acetic acid released by bacterial metabolism. Although the cross feeding explains microbial succession in cocoa fermentation, under the point of view of microbial ecology, literature on the microbial interactions in cocoa fermentation is scarce, specifically the occurrence and influence of quorum sensing (QS), for instance. It is known QS is a process of synchronization of gene expression intermediated by autoinducer molecules (AIs) in high cell densities conditions. As cocoa fermentation presents all the conditions for microbial enrichment and some studies have shown that bacteria related to QS may present competitive advantages in harsh environments, this research aimed initially to identify, by metagenomics, the QS related microbiota through the monitoring of the gene *luxS* along fermentation. Besides, bacterial genomes recovered from spontaneous fermentation were also investigated for the presence of this gene. The *luxS* gene is recognized as a universal QS marker because it characterizes the interspecific cell to cell communication. In the **first Chapter** of this work, it is presented a metagenomic analysis of an entire spontaneous cocoa fermentation sampled during 144h of fermentation. The data revealed the fungi were present along the entire fermentative process. Moreover, it was also demonstrated that reads related to the *luxS* gene are enriched as fermentation progresses, reaching a maximum at 72h of fermentation. It was also observed the genera *Enterobacter*, *"Lactobacillus"*, *Bacillus*, and *Pantoea* were associated with *luxS* gene, which allowed to track the operational taxonomic units related to QS in cocoa fermentation. In the **second Chapter**, a

comparative genomic analysis is presented. In that study, three strains of the species *Lactiplantibacillus plantarum* Lb2, *Limosilactobacillus fermentum* Lb1, and *Pediococcus acidilactici* P1 isolated from a spontaneous cocoa fermentation had their genomes sequenced and compared against all publicly available genomes of cognate species. The results shown the gene *luxS* is detected in all strains of this species and *Lp. plantarum* species in particular presents six *luxS* gene clusters, highlighting the high copy number of this gene in *Lp. plantarum* strains. For *Lm. fermentum*, there were only two gene clusters, and in *P. acidilactici* a single gene cluster. Phylogenetic analysis has shown the second gene cluster (named luxS_2) of *Lm. fermentum* was horizontally transferred by transduction from a *Lp. plantarum* strain to a *Lm. fermentum* strain, as both species present the same gene cluster and the flanking region of this cluster in *Lm. fermentum* was composed by IS30 transposases. In addition, *luxS* homologous sequences were determined by multiple alignment analysis to draw species- and clade-specific primers for rapid screening of strains to evaluate their potential related to QS and for qRT-PCR purposes. In the **third Chapter**, it is presented a comparative genomic analysis for AAB isolated from a spontaneous cocoa fermentation process, whose strains MRS7, GYC10, GYC12, GYC19, and GYC27 belonged to *A. senegalensis* species and were genotypically diverse, had their genomes sequenced and compared with public databases available for *A. senegalensis* strains at the time of publication. The study has shown *A. senegalensis* did not carry any *luxS* gene, but it presented genes related to intraspecific QS, such as acylhomoserine lactones (AHLs) and response regulators, as well as genes related to QS inhibition such as acylases and lactonases, corroborating the previous analyses presented in the first Chapter of this thesis. Besides, as AAB present potential to be applied in several industrial processes, such as cellulose production and vinegar fermentation, the metabolic pathways involved with bacterial adaptation to stressing conditions were investigated. The results shown these strains presented a good genetic repertoire related to bacterial adaptation to harsh environments, indicated by the presence of chaperones, alcohol dehydrogenases, and ABC proteins that confer tolerance to high concentrations of ethanol and high temperature. Additionally, the strains did not present pathogenic potential, as indicated by the absence of antibiotic resistance genes. In this way, the results suggested these strains of *A. senegalensis* isolated from cocoa fermentation could be applied also in other industrial processes. Finally, in the **fourth Chapter** the outcome of the Ph.D project is presented in a submitted manuscript showing the *in situ* detection of the *luxS* gene in lab scale fermentation. Thus, seven distinct fermentations were performed. The control fermentation (named F0) was conducted without duplicate, while the remaining fermentations, named F1, F2, and F3 were performed in duplicates, and inoculated with distinct combinations of cocktails containing yeasts, LAB, and/or AAB. The objective of that work was to compare the *luxS* gene expression along 96h of fermentation in each replicate and in the control (non inoculated fermentation – F0). In parallel, analyses for selective microbial enumeration, pH, and temperature monitoring, as well as metagenomics (*16S rRNA* and *ITS*), enzymatic activity dosage, and gas-chromatography coupled with mass spectrometry (GC-MS) for detection of

volatile organic metabolites (VOCs) were performed to correlate QS potential with cocoa fermentation quality. The results revealed that even in laboratory conditions the microbial succession was observed for all fermentations, which was corroborated by microbial enumeration and metataxonomic analysis. However, no statistical difference was observed for presumptive microbial enumeration of the F0 and experimental fermentations ($p > 0.05$), which was also reinforced by the absence of significative difference for α-diversity metrics determined by metataxonomics among the fermentations. Additionally, no statistically significant differences of enzymatic activities were detected, and there were no microbial amplicon sequence variants correlated with enzymatic activities, suggesting the enzymatic activity was mainly shaped by endogenous hydrolases of the seeds and not by the microbial shifts *per se*. Regarding the *luxS* gene measurements for *Lp. plantarum* and *Lm. fermentum* species, it was observed the *luxS* genes of *Lp. plantarum* were active during all fermentation period, while *Lm. fermentum luxS* genes were detected during the first 72h of fermentation. The correlation between quality of fermentation and *luxS* gene expression evidenced a positive association of *Lp. plantarum* with undesirable sensorial attributes for fermented seeds. Based on the results of this work, it is possible to affirm that the *luxS* gene is active during cocoa fermentation, as presented in the first Chapter and there was progress in the understanding of the dynamics of cocoa fermentation, with demonstration of different enzymes acting along fermentation. In addition, the results of this research consolidate the hypothesis that *Lp. plantarum* is a possible protagonist in this type of fermentation, with great potential for expressing QS pathways. Future studies may apply the data collected throughout this study under field conditions.

Keywords: Cocoa fermentation, metagenomics, quorum sensing, *luxS*, lactic acid bacteria.

List of figures

Chapter 1

Supplementary figures

Chapter 2

Supplementary figures

Chapter 3

Chapter 4

Supplementary figures

List of tables

Chapter 1

Supplementary tables

Chapter 2

Supplementary tables

Chapter 3

Chapter 4

List of abbreviations and acronyms

AAB	*Acetic acid bacteria*
ADH	*Alcohol dehydrogenase*
AI	*Auto-inducer*
ALDH	*Aldehyde dehydrogenase*
ANI	*Average Nucleotide Identity*
ARG	*Antibiotic Resistance Gene*
ASV	*Amplicon Sequence Variant*
CAPES	*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*
CDS	*Coding DNA-sequence*
CEPEC	*Centro de Pesquisas do Cacau*
CEPLAC	*Comissão Executiva do Plano da Lavoura Cacaueira*
COG	*Cluster of Orthologues Genes*
DNA	*Deoxyribonucleic acid*
FAPESP	*The São Paulo Research Foundation*
FCFRP	*Faculdade de Ciências Farmacêuticas de Ribeirão Preto*
FFCLRP	*Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto*
GC-MS	*Gas chromatography accoupled with mass spectrometry*
GYC	*Glucose, yeast extract and sodium carbonate medium*
HGT	*Horizontal gene transfer*
ITS	*Internal Transcribed Spacer*
LAB	*Lactic acid bacteria*
MAG	*Metagenome-assembled genome*
MetaPhlan	*Metagenomic Phylogenetic Analysis*
MRS	*De Man, Rogosa and Sharpe medium*
MSLA	*Multiple Sequence Alignments*
NCBI	*National Center for Biotechnology Information*
NGS	*Next-Generation Sequencing*
OTU	*Operational taxonomic unit*
PCA	*Principal componente analysis*
PCR	*Polymerase Chain Reaction*
Qiime	*Quantitative Insights Into Microbial Ecology*
QQ	*Quorum quenching*
qRT-PCR	*Quantitative Real-Time Polymerase Chain Reaction*
QS	*Quorum sensing*
RNA	*Ribonucleic acid*
rRNA	*Ribosomal ribonucleic acid*
UniVr	*Università Degli Studi di Verona*
USP	*Universidade de São Paulo*
VBNC	*Viable but non-culturable*
VOCs	*Volatile organic compounds*

WGS	*Whole-Genome Sequencing*

Summary

Introduction

Introduction

1.1. Cocoa fermentation dynamics

Cocoa seeds are the raw material for chocolate production, and the seeds must be initially fermented, then dried and finally processed to yield the final cocoa product. The most intriguing step is the fermentation, which is characterized by three main phases guided and shaped by a microbial succession of yeasts, lactic acid bacteria (LAB) and acetic acid bacteria (AAB). The source of these microorganisms for the fermentation is attributed to the contamination by the autochthonous microbiota, which reaches the seeds after fruit cutting, and through the in-house microbiota persistent in the fermentation boxes, banana leaves, and other utensils used to handle the fermentation (CHAGAS JUNIOR; FERREIRA; LOPES, 2021; DE VUYST; LEROY, 2020).

Cocoa beans are surrounded by a mucilaginous white pulp, which is a matrix hard to be extracted by mechanical methods, and the fermentation is responsible to consume its constituents, consequently, draining it and allowing the assessment to the seeds content (CASTRO-ALAYO et al., 2019). The pulp is composed by 85% of water, 10%-15% of sugars such as glucose, fructose, and sucrose, which varies according to the fruits' maturation age. Besides, it is also composed by 2%-3% of pentoses, 1%-3% of citric acid, 1.5% of pectin, amino acids, vitamins, such as C vitamin, and minerals (ORDOÑEZ-ARAQUE et al., 2020). Due to these highly nutritious characteristics, the mucilaginous pulp supports the microbial survival and growth.

In Brazil, cocoa fermentation is carried out in wooden boxes, which is also common in Malaysia, while in other countries several other methods can be used (CAMU et al., 2008), such as in heaps (Ghana and Ivory Coast), baskets (Nigeria and Ghana), trays (Ghana), sacks (Ecuador) or in platform fermentation (Ecuador) (PEREIRA; SOCCOL; SOCCOL, 2016). The boxes present interspaced holes in the bottom to drain the cocoa honey, which is the liquid fraction of the cocoa pulp, while in the top the cocoa mass could be capped with banana leaves to keep the heat produced by microbial activity. Fermentation carried out in baskets or heaps are also quilted with banana leaves to keep heat (AFOAKWA; GUDA; GADHE, 2017). Currently, there is no mandatory standard procedure to ferment cocoa seeds but is recommended to be performed for no longer than six days because there is a risk of deterioration, with increased multiplication of filamentous fungi. Besides, longer-time fermentation can become result in over-fermented seeds and produce a putrid ammonia smell due to poor fermentation conditions (GUEHI et al., 2010).

Regarding the microbial dynamics along fermentation, the process could be divided in three main steps: (1) Anaerobic growth of yeasts in the first 24h; (2) microaerophilic growth of LAB in the range of 24h-72h; and (3) aerobic growth of AAB in the range of 48h-112h, dropping after 120h of fermentation (CAMU et al., 2007). Yeasts are the pioneers in the fermentation process, which starts in acidic environment (pH~3.6) and with low levels of oxygen. As yeasts grow, glucose and fructose from the mucilaginous pulp are consumed, but there is also the release of enzymes, such as pectinases, invertases, amylases, cellulases and xylanases, leading to an augment of monosaccharides' availability (DE ARAÚJO et al., 2019; DE VUYST; WECKX, 2016).

Yeasts' activity is responsible for preparing the cocoa fermentation environment for LAB colonization as they produce sugars, ethanol and CO2, providing a microaerophilic microenvironment, which is ideal for LAB growth. At the same time, the ethanol also shapes the microbial succession since it is a substrate for AAB growth. The microbial metabolism will generate heat, which in combination with the metabolic products (ethanol, acetic acid and flavour compounds) will diffuse into the seeds and reach the cotyledon, causing the death of the seed embryo, releasing several chocolate flavour precursors. Besides, yeasts metabolism provides several secondary metabolites that will impact the sensorial traits of the fermented cocoa beans, including higher alcohols, ketones, esters, aldehydes, fatty acid esters, and organic acids (DE VUYST; WECKX, 2016; SCHWAN; PEREIRA; FLEET, 2014).

Once yeasts started the process of pulp's depectinization and the draining of mucilage, the air reaches the interstices of seeds assuring a microaerophilic environment for LAB. Additionally, during this phase, temperature also increases up to 35ºC-45ºC due to the production of ethanol by exothermic reactions as well as the liberation of some organic acids from yeasts' metabolism (for example, succinic and acetic acid), which may confer a buffering effect for microbial metabolism (DE VUYST; WECKX, 2016). *Limosilactobacillus fermentum* and *Lactiplantibacillus plantarum* are the dominant LAB species along fermentation, as they are tolerant to acid and ethanol, besides presenting the ability to ferment citrate. This evidences a fundamental role for LAB in pulp drainage, which is related to the metabolization of citrate and sugars that can be further converted to mannitol, acetic acid, and lactic acid by homo and heterofermentative LAB (FIGUEROA-HERNÁNDEZ et al., 2019). These metabolites aid the growth of AAB which oxidize ethanol and lactic acid into acetic acid, and also mannitol to fructose (MOENS; LEFEBER; DE VUYST, 2014) through exothermic reactions that will favour the embryo's death inside the seeds' cotyledons due to the enhancement of box temperatures up to 50ºC (LEE et al., 2019). Once the cotyledons are reached, several endogenous hydrolases will catalyse the transformation of stored pigments and other compounds, such as proteins, which will be able to integrate complex biochemical pathways that will support the development of sensorial traits involving taste, aroma, and colour in the fermented seeds (CHAGAS JUNIOR; FERREIRA; LOPES, 2021; LEE et al., 2019; OUATTARA; ELIAS; DUDLEY, 2020). Well-performed fermentations are supposed to provide the traditional chocolate flavour and brown-coloured beans (OUATTARA; ELIAS; DUDLEY, 2020).

Taking into account the crucial role of microorganisms for chocolate's flavour formation, several studies have addressed the challenge of obtaining a suitable starter culture to standardize the fermentation (CRAFACK et al., 2013; FARRERA et al., 2021; LEFEBER et al., 2012; VISINTIN et al., 2017). In this sense, the standardization of cocoa fermentation process should still retain the geographical identity of diverse chocolates, while increasing quality, safety and yields.

Some authors have already achieved partial success with microbial consortia of yeast and bacteria for fermentation at different locations. Most of the studies showed the combination of yeasts and bacteria augmented the rates of pulp degradation and ethanol production as well as the ability of AAB to over oxidize ethanol and lactic acid, which are considered desirable properties in well-carried fermentations (MOENS; LEFEBER; DE VUYST, 2014; OUATTARA; ELIAS; DUDLEY, 2020). Other authors evaluated the impact of the addition of starter cultures on the production of volatile organic compounds (VOCs). A study showed the combination of *Saccharomyces cerevisiae* UFLA CCMA 0200, *Lactiplantibacillus plantarum* CCMA 0238, and *Acetobacter pasteurianus* CCMA 0241 could result in a suitable candidate starter culture, considering the fermentations inoculated with the cocktail of microbial strains produced better chocolates in comparison with the non-inoculated control. Those authors also reported the VOCs butanediol and 2,3-dimethylpirazine were determinant for chocolates' sensorial quality (MAGALHÃES DA VEIGA MOREIRA et al., 2017). In a Mexican fermentation study carried out with a mix of *S. cerevisiae* 120, *Lp. plantarum* 14, *Lm. fermentum* 16, *Acetobacter pasteurianus* 98, and *Acetobacter tropicalis*, the authors observed an increase of pyrazines (trimethylpyrazine, 2-5-dimethylpyrazine, and tetramethylpyrazine), as well as a high-content of alcohols, ketones, and aldehydes in inoculated fermentations. These VOCs are precursors of pyrazines that may serve as substrates for Maillard reactions and Streaker synthesis at the later processing steps of fermented cocoa seeds (such as roasting and conching), contributing to flavour development (ALVAREZ-VILLAGOMEZ et al., 2022). Diverse reviews pointed out the influence of VOCs on cocoa seeds' quality as well as the interference of starter culture and VOCs profiles, showing the positive and negative impact of the production of acids, alcohols, aldehydes, ketones, esters, pyrazines, pyrroles, terpenoids, furans, and alkanes (BATISTA et al., 2016; CASTRO-ALAYO et al., 2019; MAGALHÃES DA VEIGA MOREIRA et al., 2017; MOTA-GUTIERREZ et al., 2019; SAUNSHIA et al., 2018). Taken all together, data on microbial composition combined with the determination of VOCs produced during cocoa fermentation are are expected to allow the development of starter cultures for industrial applicability and for a consumer acceptable final product.

1.2. Quorum sensing

Food fermentations are complex processes that highly depend on microbial interactions that allow the exchange of energy sources and lead to the dominance of well-adapted synergistic microbes, with different niche occupations and division of labour. These microbial consortia are important to provide more robustness to environmental variations, in comparison to pure cultures (SMID; LACROIX, 2013).

Some microorganism-microorganism interactions depend on the exchange of diffusible signalling molecules among cells (PARK et al., 2016). This communication process is known as Quorum sensing (QS), and it depends on auto-inducing molecules (AIs) also called "pheromones" that accumulate in situations of high cell densities. The QS systems are basically constituted by an autoinducer, a synthase of the autoinducer and a receptor for this molecule, whose actions consist, in most cases, in the positive regulation (activation) of target genes in the presence of the pheromone (PINTO; PAPPAS; WINANS, 2012). In other words, it is a form of transcriptional regulation dependent on high bacterial cell densities.

Quorum Sensing-dependent systems have been identified in many species of Gram-negative and Gram-positive bacteria, being related to the regulation of diverse functions, such as biofilm formation, bioluminescence, conjugation and even virulence (PAPENFORT; BASSLER, 2016). There are several types of QS systems in bacteria, which are dependent on autoinducers such as acyl homoserine lactone (AHL), prevalent in Proteobacteria. In a simplified way, the family of LuxI enzymes (AI synthases) synthesize AHL signal molecules, which will be detected by the family of receptor proteins named as LuxR (response regulators), that act as transcriptional regulators. Although phylogenetic relationships exist, different chemical signals are used in different bacteria (SCHUSTER et al., 2013). A specific set of AIs is produced and detected by each QS bacterial species but it is reasonable to assume that these species are exposed to non-cognate signal analogs produced by other species in the vicinity (HAWVER; JUNG; NG, 2016).

Gram-negative bacteria can present two basic QS circuits. The first one is characterized by a single-component system, in which AIs are produced by a AI synthase, released to the extracellular medium and, later, diffuse to the interior of cells. The AIs are then recognized by cytoplasmic QS receptors, acting as transcription factors. In the two-component system, the AI synthase produces inducing molecules that will be released into the extracellular environment. These molecules will sensitize membrane receptors, which will trigger an intracellular signalling cascade that regulates the downstream response to QS (HAWVER; JUNG; NG, 2016).

In Gram-positive bacteria, two basic QS circuits also occur, but they are more elaborated. In the circuit encompassing one-component system the ribosome, which represents the AIP synthase, produces self-inducing peptides, which are released into the extracellular environment by a transporter, in most cases, as a longer peptide precursor (HAWVER; JUNG; NG, 2016). It is worth mentioning that, in this case, there is a specific gene that encodes AIP. In the extracellular environment, these peptides are hydrolyzed and return to the intracellular environment through a permease, and then act as a transcription factor after sensitizing intracellular QS receptors (also called response regulators). In the two-component system, the self-inducing peptides synthesized by the ribosome are released into the extracellular environment, where they undergo post-translational modifications and, after sensitizing a transmembrane receptor, trigger an intracellular signalling cascade that regulates the QS-dependent response (HAWVER; JUNG; NG, 2016).

In Gram-negative bacteria, acyl homoserine lactones are the most common classes of self-inducing molecules (PAPENFORT; BASSLER, 2016). *Vibrio harveyi* is a species of Gram-negative marine bacterium, in which all architecture and functioning of QS systems were first described. This bacterium has two QS systems that control functions such as the production of siderophores, metalloproteases, type III secretion system and bioluminescence (HENKE; BASSLER, 2004). Self-inducing system 1 (AI-1) is an intraspecific communication system (Papenfort & Bassler, 2016). The AI-1 molecule is an N-3-hydroxybutanoyl-HSL produced by LuxM family proteins (analogous to LUXI – AI synthase) and recognized by LuxN family receptors (response regulators) (HENKE; BASSLER, 2004).

The autoinducer system 2 (AI-2) is a system related to interspecific communication and its synthesis depends on the *luxS* gene. LuxS proteins catalyse the conversion of S-ribosilhomocysteine (SRH) to the 4,5-dihydroxy-2,3-pentanedione molecule (DPD), which is a precursor of AI-2. Due to the instability of DPD, rearrangements can occur in the molecule resulting in several derivative molecules with analogous activity to AI-2 (TUROVSKIY; CHIKINDAS, 2006).

The AI-2 molecule in high concentration and high cell density interacts with the periplasmic protein-binding receptor LuxP, modifying its conformation, which leads to the dimerization of the cytoplasmic regulatory subunit LuxQ with LuxP. Dimerization results in the change of histidine kinase to phosphatase state in LuxQ, initiating the dephosphorylation of LuxO and making expression of LuxR possible. In *V. harveyi*, LuxR expression induces transcription of *LuxCDABE* and target genes, which leads to luciferase-mediated emission of light photons (MANDABI; GANIN; MEIJLER, 2015).

A third system, discovered in Vibrio cholerae, called CqsS (“Cholerae Quorum Sensing Sensor”), controls in parallel with systems 1 and 2 QS target genes in *Vibrio* sp. Analysis of the complete genome of *V*. *cholerae* revealed the presence of a new synthase that composes this system (HENKE; BASSLER, 2004), CqsA (“Cholerae Quorum Sensing Autoinducer”), which synthesizes CAI-1 (“Cholerae Autoinducer 1”) from S-adenosylmethionine and decanoyl-CoA (PAPENFORT; BASSLER, 2016). CqsA does not show homology to LuxM or to any other autoinducer synthase and LuxN is not able to recognize any CAI-1 (HENKE; BASSLER, 2004). Its worth’s to highlight other recently discovered QS autoinducers such as indole and AI-3 molecules related to interspecific and interkingdom communication respectively and reviewed elsewhere (WU; LUO, 2021), evidencing the huge variability of QS systems and effectors to be still disclosed.

In the literature it has been reported that in the traditional fermentation of Chinese rice wine, performed manually, greater amounts of QS regulatory genes were detected, possibly due to the greater number of contaminating species in manual fermentations compared to the process carried out industrially (HONG et al., 2016). For that product, in the fermentations considered of low quality, there was a prevalence of self-inducing genes for transport and synthesis, suggesting that the species present in that type of process had genes related to QS, while in fermentations considered of good quality, there was a predominance of self-inducing genes for regulation and inhibition of QS, suggesting that species in good quality fermentations were perhaps more independent of QS (HONG et al., 2016). By the other hand, QS may display a functional role in food quality by helping beneficial bacteria to tolerate adverse conditions in fermentations. In the literature it has been reported and reviewed, the importance of QS for biofilm formation, for tolerance to acid and heat, as well as for the regulation of bacteriocin production and bacterial competence (JOHANSEN; JESPERSEN, 2017).

Some authors defend the mixed cultures may favour the dominance of some bacteria, specifically LAB, as the multiple AI-2 signalling in the same environment and time could benefit bacteria with robust interspecific QS systems to replace less robust taxa (PARK et al., 2016), which would result in the standardization of the process with formation of desired flavours (RUL; MONNET, 2015).

To the best of our knowledge, there is a gap in the literature to relate how and at which extent QS pathways may contribute to the dominance of some microorganisms in food fermentations, especially in cocoa fermentation. As the LAB are predominant in this process, even when AAB are in high levels, our hypothesis is QS may contribute for LAB dominance during the fermentation, especially for *Lactiplantibacillus plantarum*, a dominant species identified in several fermentations (CAMU et al., 2007). Therefore, studies encompassing the influence of QS or it's inhibition in cocoa fermentation are needed to fill this gap and to select candidate starter strains with the ability to dominate, standardize and replace undesirable microorganisms in fermentation.

Objective

2. Objective

This thesis aims to unravel the influence of quorum sensing on cocoa fermentation dynamics and scrutinize bacterial genomes of strains recovered from spontaneous cocoa fermentation to characterize inter and intraspecific quorum sensing determinants as well as metabolic repertoire of adaptation in cocoa fermentation environment. Finally, it is also aiming at to monitor the *luxS* gene expression in laboratory scale fermentations to infer whether quorum sensing plays or not a role in cocoa fermentation quality.

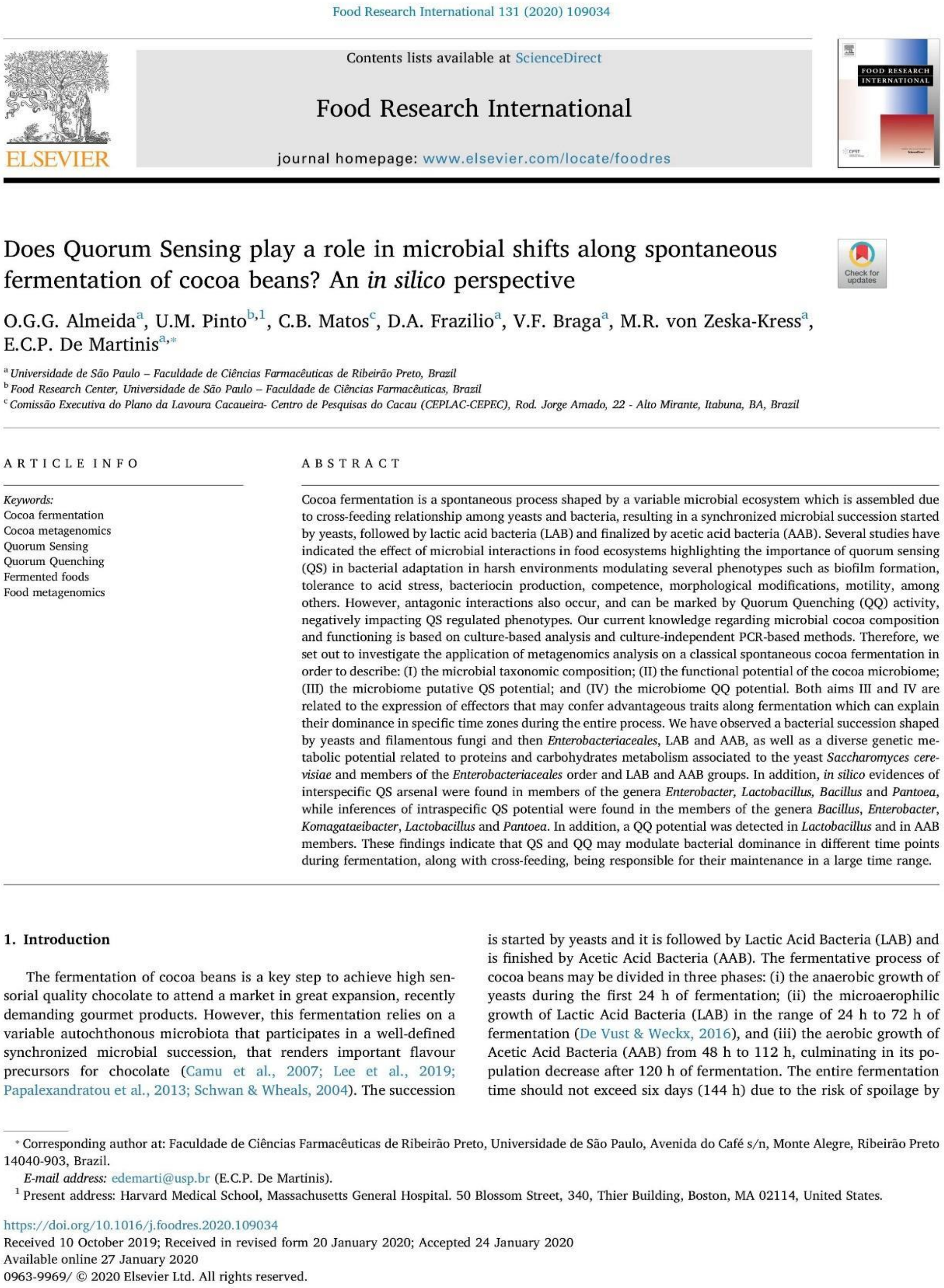
Chapter 1

3. Chapter 1 - Does Quorum Sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An *in silico* perspective

Specific objective

- To evaluate an entire spontaneous cocoa fermentation performed in field conditions, carried out in six days (144h) in Itabuna city of Bahia state of Brazil through metagenomics to determine microbial function and composition and monitor *luxS* derivate reads along fermentation. Besides, to identify cocoa-related bacterial members harbouring the *luxS* gene.

The following article was published in the Journal *Food Research International*, a specialized journal in the field of Food Microbiology (doi: https://doi.org/10.1016/j.foodres.2020.109034). The authorization for reuse in Thesis and Dissertations is depicted in the Attachment A.



Food Research International 131 (2020) 109034

Contents lists available at ScienceDirect

Food Research International

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

Does Quorum Sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An *in silico* perspective

O.G.G. Almeida[a], U.M. Pinto[b,1], C.B. Matos[c], D.A. Frazilio[a], V.F. Braga[a], M.R. von Zeska-Kress[a], E.C.P. De Martinis[a,*]

[a] *Universidade de São Paulo – Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Brazil*
[b] *Food Research Center, Universidade de São Paulo – Faculdade de Ciências Farmacêuticas, Brazil*
[c] *Comissão Executiva do Plano da Lavoura Cacaueira- Centro de Pesquisas do Cacau (CEPLAC-CEPEC), Rod. Jorge Amado, 22 - Alto Mirante, Itabuna, BA, Brazil*

ARTICLE INFO

Keywords:
Cocoa fermentation
Cocoa metagenomics
Quorum Sensing
Quorum Quenching
Fermented foods
Food metagenomics

ABSTRACT

Cocoa fermentation is a spontaneous process shaped by a variable microbial ecosystem which is assembled due to cross-feeding relationship among yeasts and bacteria, resulting in a synchronized microbial succession started by yeasts, followed by lactic acid bacteria (LAB) and finalized by acetic acid bacteria (AAB). Several studies have indicated the effect of microbial interactions in food ecosystems highlighting the importance of quorum sensing (QS) in bacterial adaptation in harsh environments modulating several phenotypes such as biofilm formation, tolerance to acid stress, bacteriocin production, competence, morphological modifications, motility, among others. However, antagonic interactions also occur, and can be marked by Quorum Quenching (QQ) activity, negatively impacting QS regulated phenotypes. Our current knowledge regarding microbial cocoa composition and functioning is based on culture-based analysis and culture-independent PCR-based methods. Therefore, we set out to investigate the application of metagenomics analysis on a classical spontaneous cocoa fermentation in order to describe: (I) the microbial taxonomic composition; (II) the functional potential of the cocoa microbiome; (III) the microbiome putative QS potential; and (IV) the microbiome QQ potential. Both aims III and IV are related to the expression of effectors that may confer advantageous traits along fermentation which can explain their dominance in specific time zones during the entire process. We have observed a bacterial succession shaped by yeasts and filamentous fungi and then *Enterobacteriaceales*, LAB and AAB, as well as a diverse genetic metabolic potential related to proteins and carbohydrates metabolism associated to the yeast *Saccharomyces cerevisiae* and members of the *Enterobacteriaceales* order and LAB and AAB groups. In addition, *in silico* evidences of interspecific QS arsenal were found in members of the genera *Enterobacter, Lactobacillus, Bacillus* and *Pantoea*, while inferences of intraspecific QS potential were found in the members of the genera *Bacillus*, *Enterobacter*, *Komagataeibacter*, *Lactobacillus* and *Pantoea*. In addition, a QQ potential was detected in *Lactobacillus* and in AAB members. These findings indicate that QS and QQ may modulate bacterial dominance in different time points during fermentation, along with cross-feeding, being responsible for their maintenance in a large time range.

1. Introduction

The fermentation of cocoa beans is a key step to achieve high sensorial quality chocolate to attend a market in great expansion, recently demanding gourmet products. However, this fermentation relies on a variable autochthonous microbiota that participates in a well-defined synchronized microbial succession, that renders important flavour precursors for chocolate (Camu et al., 2007; Lee et al., 2019; Papalexandratou et al., 2013; Schwan & Wheals, 2004). The succession is started by yeasts and it is followed by Lactic Acid Bacteria (LAB) and is finished by Acetic Acid Bacteria (AAB). The fermentative process of cocoa beans may be divided in three phases: (i) the anaerobic growth of yeasts during the first 24 h of fermentation; (ii) the microaerophilic growth of Lactic Acid Bacteria (LAB) in the range of 24 h to 72 h of fermentation (De Vust & Weckx, 2016), and (iii) the aerobic growth of Acetic Acid Bacteria (AAB) from 48 h to 112 h, culminating in its population decrease after 120 h of fermentation. The entire fermentation time should not exceed six days (144 h) due to the risk of spoilage by

[*] Corresponding author at: Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Avenida do Café s/n, Monte Alegre, Ribeirão Preto 14040-903, Brazil.
E-mail address: edemarti@usp.br (E.C.P. De Martinis).
[1] Present address: Harvard Medical School, Massachusetts General Hospital. 50 Blossom Street, 340, Thier Building, Boston, MA 02114, United States.

https://doi.org/10.1016/j.foodres.2020.109034
Received 10 October 2019; Received in revised form 20 January 2020; Accepted 24 January 2020
Available online 27 January 2020

filamentous fungi (Schwan & Wheals, 2004).

Each organism contributes to flavour generation due to the capacity of producing several compounds released during its growth along the spontaneous process (Schwan & Wheals, 2004). The consumption of nutrient-rich cocoa pulp compounds, such as citric acid, vitamin C, glucose, fructose, amino acids and proteins results in several intermediate chemical compounds generated by the microbial metabolism (Lee et al., 2019).

As it is being reported in the literature, yeasts start the fermentation of cocoa beans by the enzymatic degradation of pectin which is a part of the mucilaginous cocoa pulp, while also consuming citric acid which leads to pH increase of the pulp (Papalexandratou et al., 2011). Meanwhile, the temperature also increases due to exothermic reactions that lead to ethanol synthesis. The yeasts' pectinase activity releases simple sugars from the pulp and an increased level of ethanol due to their fermentation, favouring the succession of LAB due to cross-feeding and the higher pH. Homo- and heterofermentative metabolism of LAB causes the formation of lactic acid, ethanol, mannitol, acetic acid and CO_2, among other compounds (De Vust & Weckx, 2016; Papalexandratou & Nielsen, 2016; Schwan & Wheals, 2004).

Finally, the AAB convert these intermediate metabolites to acetic acid, which is oxidized to CO_2 and H_2O. Moreover, during these metabolic conversion steps, the temperature inside the fermentation boxes increases up to 50 °C, which in addition to the harsh presence of ethanol and acetic acid, results in the death of the embryo in the beans' cotyledons (Lee et al., 2019; Schwan & Wheals, 2004). The result is the release of the stored components and pigments in the seeds that may integrate complex biochemical reactions catalysed by endogenous hydrolases. These reactions will support the development of the desired characteristics of taste, aroma and colour in the fermented seeds, which are the raw material for chocolate production (Camu et al., 2007; Papalexandratou & De Vuyst, 2011; Lee et al., 2019). For that reason, understanding the microbial dynamics during the fermentation process is crucial to increase it yield and quality.

Although there have been a few reports on the characterization of the microbial diversity in cocoa beans fermentation with the application of Next-Generation Sequencing (NGS) for metataxonomics (massive and parallel marker gene sequencing - 16S *rRNA*/ITS/18S *rRNA*) (Illeghems, De Vuyst, Papalexandratou, & Weckx, 2012; Mota-Gutierrez et al., 2018; Serra et al., 2019) and for metagenomics (Agyirifo et al., 2019; Illeghems, Weckx, & De Vuyst, 2015), no previous study has been performed encompassing the entire fermentation process, but only part of it. Furthermore, there is a lack of descriptive studies regarding cocoa microbiome functional potential and a lack of information related to microbial capacity to maintain their high loads during microbial succession. It has been indicated in the literature that cross-feeding is responsible for microbial shifts along fermentation (Lee et al., 2019; Schwan & Wheals, 2004). However, there is no information on the possible role that Quorum Sensing (QS) may play in microbial synchronization at specific fermentation stages, since the microorganisms usually present high populations and a typical composition pattern (fungi, LAB and AAB, respectively) from the beginning to the end of the process.

Quorum sensing is a population density dependent communication mechanism mediated by small diffusible molecules known as autoinducers (AIs). As bacterial cells multiply, they release AIs in their surrounds and when a threshold concentration is reached, which coincides with high cell density, these molecules activate the expression of a set of genes, part of the quorum sensing regulon, which usually includes the production of more AIs (Johansen & Jespersen, 2017). Quorum sensing can mediate communication within a bacterial species (intraspecific communication) and across different species (interspecific communication) (Johansen & Jespersen, 2017).

In Gram-negative bacteria, intraspecific QS is commonly mediated by AI molecules of the group N-acyl-homoserine-lactones (AHLs), commonly referred to as AI-1, while in the Gram-positive bacteria the signalling molecules are called autoinducer peptides (AIPs) synthetized via ribosome (Thoendel et al., 2011). There are also several other kinds of signalling molecules mediating QS in different organisms, as recently reviewed (Lima et al., 2020). Even though there is great diversity of signalling molecules, the communication mechanism usually involves the synthesis, export and sensing of the AI molecule, which in turn, activates the quorum sensing regulon via conserved transcription factors such as LuxR homologues or response regulators belonging to two-component regulatory systems (Monnet, Juillard, & Gardan, 2016). Interspecific QS communication depends upon the presence of the *luxS* gene which is present in more than 530 bacterial genomes and it is responsible for the synthesis of Autoinducer-2 molecule (AI-2) (Pereira, Thompson, & Xavier, 2012).

Regardless of the QS system, the AI contacts a cell surface protein or diffuses into the cytoplasm, contacting a transcription regulator, inducing downstream responses to control several bacterial traits such as biofilm formation, tolerance to acid stress, regulation of bacteriocin production, competence, morphological modifications, differential adhesion and motility (Almeida et al., 2018; Johansen & Jespersen, 2017), which are important traits concerning bacterial adaptation in complex matrices.

Quorum sensing communication may be disrupted in a process termed Quorum Quenching (QQ) (Dong et al., 2001). This can be accomplished by bacteria harbouring genes that code for acylases, lactonases and oxidoreductases. The mechanisms in which acylases generally disrupt QS in Gram-negative bacteria are related to an enzymatic activity that causes the cleavage of the AHL molecule into the lactone ring and a free fatty acid moiety that is unable to bind to the transcriptional regulators, while the lactonases cleave directly the lactone ring from the molecule (Chen, Gao, Chen, Gao, Chen, Yu, & Li, 2013; Grandclément, Tannières, Moréra, Dessaux, & Faure, 2016). Conversely, the oxidoreductases modify the AHL molecules by oxi-reduction reactions, making them unrecognizable by the receptor protein. These mechanisms prevent QS regulated gene expression of specific traits (Chen et al., 2013).

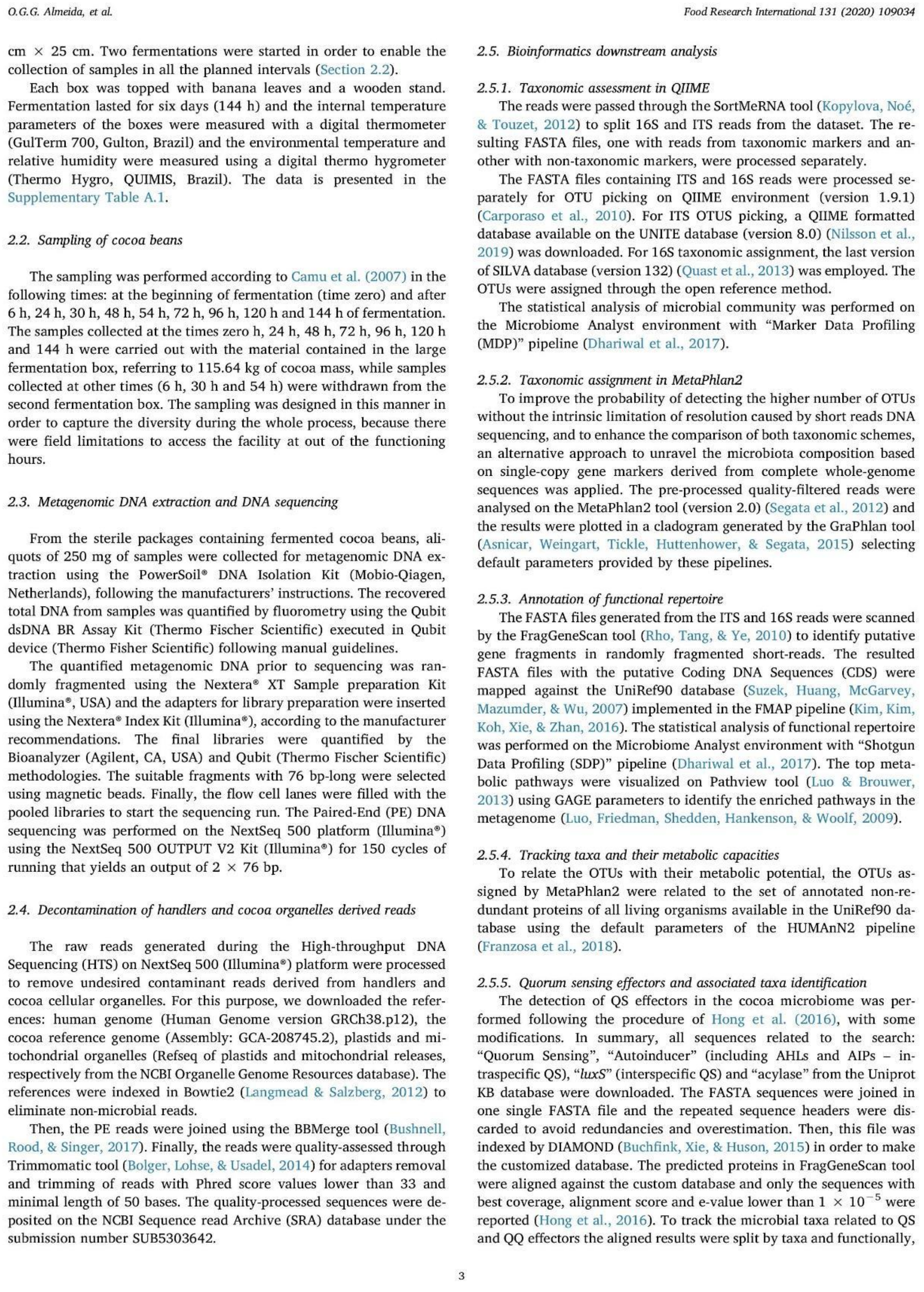
Taking advantage of NGS tools to assess complex microbial communities, the objective of this study was to unravel taxonomic and functional aspects of the microbial succession in cocoa fermentation. Specifically, we included an *in silico* search for Quorum Sensing (QS) effectors and agonists (Quenchers), since QS and QQ repertoire may be crucial for microbial adaptation to harsh conditions and we presume that in cocoa beans fermentation, the bacterial harbouring potential QS genes may successfully dominate the scenario for longer time ranges.

2. Materials and methods

2.1. Cocoa fermentation

The cocoa beans were fermented at the "Centro de Pesquisas do Cacau" (Cocoa Research Center), a department of "Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC)" of the Brazilian Federal Government, located in the Bahia state, Brazil. The fermentation took place during the cocoa harvest period (from 27 November 2017 to 04 December 2017).

A variable mixture of the *Criollo* and *Forastero* cocoa varieties was employed as the matrix for fermentation, as commonly practiced by cocoa farmers in the field conditions. The cocoa fruits were cut; the seeds were collected and deposited inside a wooden box with dimensions of 50 cm × 50 cm × 50 cm that also had small holes in its base (Supplementary Fig. A.1) in order to drain the cocoa honey and to aid in the drying of the seeds during the fermentation. After cutting the fruits, 115.64 kg of fresh cocoa mass was obtained and immediately transferred to the wooden fermentation box. On the second day, a second fermentation was started with 19.83 kg of cocoa mass from the same crop of cocoa from the Criollo and Forastero varieties, which was transferred to a second wooden box with dimensions of 25 cm × 25



cm × 25 cm. Two fermentations were started in order to enable the collection of samples in all the planned intervals (Section 2.2).

Each box was topped with banana leaves and a wooden stand. Fermentation lasted for six days (144 h) and the internal temperature parameters of the boxes were measured with a digital thermometer (GulTerm 700, Gulton, Brazil) and the environmental temperature and relative humidity were measured using a digital thermo hygrometer (Thermo Hygro, QUIMIS, Brazil). The data is presented in the Supplementary Table A.1.

2.2. Sampling of cocoa beans

The sampling was performed according to Camu et al. (2007) in the following times: at the beginning of fermentation (time zero) and after 6 h, 24 h, 30 h, 48 h, 54 h, 72 h, 96 h, 120 h and 144 h of fermentation. The samples collected at the times zero h, 24 h, 48 h, 72 h, 96 h, 120 h and 144 h were carried out with the material contained in the large fermentation box, referring to 115.64 kg of cocoa mass, while samples collected at other times (6 h, 30 h and 54 h) were withdrawn from the second fermentation box. The sampling was designed in this manner in order to capture the diversity during the whole process, because there were field limitations to access the facility at out of the functioning hours.

2.3. Metagenomic DNA extraction and DNA sequencing

From the sterile packages containing fermented cocoa beans, aliquots of 250 mg of samples were collected for metagenomic DNA extraction using the PowerSoil® DNA Isolation Kit (Mobio-Qiagen, Netherlands), following the manufacturers' instructions. The recovered total DNA from samples was quantified by fluorometry using the Qubit dsDNA BR Assay Kit (Thermo Fischer Scientific) executed in Qubit device (Thermo Fisher Scientific) following manual guidelines.

The quantified metagenomic DNA prior to sequencing was randomly fragmented using the Nextera® XT Sample preparation Kit (Illumina®, USA) and the adapters for library preparation were inserted using the Nextera® Index Kit (Illumina®), according to the manufacturer recommendations. The final libraries were quantified by the Bioanalyzer (Agilent, CA, USA) and Qubit (Thermo Fischer Scientific) methodologies. The suitable fragments with 76 bp-long were selected using magnetic beads. Finally, the flow cell lanes were filled with the pooled libraries to start the sequencing run. The Paired-End (PE) DNA sequencing was performed on the NextSeq 500 platform (Illumina®) using the NextSeq 500 OUTPUT V2 Kit (Illumina®) for 150 cycles of running that yields an output of 2 × 76 bp.

2.4. Decontamination of handlers and cocoa organelles derived reads

The raw reads generated during the High-throughput DNA Sequencing (HTS) on NextSeq 500 (Illumina®) platform were processed to remove undesired contaminant reads derived from handlers and cocoa cellular organelles. For this purpose, we downloaded the references: human genome (Human Genome version GRCh38.p12), the cocoa reference genome (Assembly: GCA-208745.2), plastids and mitochondrial organelles (Refseq of plastids and mitochondrial releases, respectively from the NCBI Organelle Genome Resources database). The references were indexed in Bowtie2 (Langmead & Salzberg, 2012) to eliminate non-microbial reads.

Then, the PE reads were joined using the BBMerge tool (Bushnell, Rood, & Singer, 2017). Finally, the reads were quality-assessed through Trimmomatic tool (Bolger, Lohse, & Usadel, 2014) for adapters removal and trimming of reads with Phred score values lower than 33 and minimal length of 50 bases. The quality-processed sequences were deposited on the NCBI Sequence read Archive (SRA) database under the submission number SUB5303642.

2.5. Bioinformatics downstream analysis

2.5.1. Taxonomic assessment in QIIME

The reads were passed through the SortMeRNA tool (Kopylova, Noé, & Touzet, 2012) to split 16S and ITS reads from the dataset. The resulting FASTA files, one with reads from taxonomic markers and another with non-taxonomic markers, were processed separately.

The FASTA files containing ITS and 16S reads were processed separately for OTU picking on QIIME environment (version 1.9.1) (Carporaso et al., 2010). For ITS OTUS picking, a QIIME formatted database available on the UNITE database (version 8.0) (Nilsson et al., 2019) was downloaded. For 16S taxonomic assignment, the last version of SILVA database (version 132) (Quast et al., 2013) was employed. The OTUs were assigned through the open reference method.

The statistical analysis of microbial community was performed on the Microbiome Analyst environment with "Marker Data Profiling (MDP)" pipeline (Dhariwal et al., 2017).

2.5.2. Taxonomic assignment in MetaPhlan2

To improve the probability of detecting the higher number of OTUs without the intrinsic limitation of resolution caused by short reads DNA sequencing, and to enhance the comparison of both taxonomic schemes, an alternative approach to unravel the microbiota composition based on single-copy gene markers derived from complete whole-genome sequences was applied. The pre-processed quality-filtered reads were analysed on the MetaPhlan2 tool (version 2.0) (Segata et al., 2012) and the results were plotted in a cladogram generated by the GraPhlan tool (Asnicar, Weingart, Tickle, Huttenhower, & Segata, 2015) selecting default parameters provided by these pipelines.

2.5.3. Annotation of functional repertoire

The FASTA files generated from the ITS and 16S reads were scanned by the FragGeneScan tool (Rho, Tang, & Ye, 2010) to identify putative gene fragments in randomly fragmented short-reads. The resulted FASTA files with the putative Coding DNA Sequences (CDS) were mapped against the UniRef90 database (Suzek, Huang, McGarvey, Mazumder, & Wu, 2007) implemented in the FMAP pipeline (Kim, Kim, Koh, Xie, & Zhan, 2016). The statistical analysis of functional repertoire was performed on the Microbiome Analyst environment with "Shotgun Data Profiling (SDP)" pipeline (Dhariwal et al., 2017). The top metabolic pathways were visualized on Pathview tool (Luo & Brouwer, 2013) using GAGE parameters to identify the enriched pathways in the metagenome (Luo, Friedman, Shedden, Hankenson, & Woolf, 2009).

2.5.4. Tracking taxa and their metabolic capacities

To relate the OTUs with their metabolic potential, the OTUs assigned by MetaPhlan2 were related to the set of annotated non-redundant proteins of all living organisms available in the UniRef90 database using the default parameters of the HUMAnN2 pipeline (Franzosa et al., 2018).

2.5.5. Quorum sensing effectors and associated taxa identification

The detection of QS effectors in the cocoa microbiome was performed following the procedure of Hong et al. (2016), with some modifications. In summary, all sequences related to the search: "Quorum Sensing", "Autoinducer" (including AHLs and AIPs – intraspecific QS), "*luxS*" (interspecific QS) and "acylase" from the Uniprot KB database were downloaded. The FASTA sequences were joined in one single FASTA file and the repeated sequence headers were discarded to avoid redundancies and overestimation. Then, this file was indexed by DIAMOND (Buchfink, Xie, & Huson, 2015) in order to make the customized database. The predicted proteins in FragGeneScan tool were aligned against the custom database and only the sequences with best coverage, alignment score and e-value lower than 1×10^{-5} were reported (Hong et al., 2016). To track the microbial taxa related to QS and QQ effectors the aligned results were split by taxa and functionally, divided by five functional classes: “Transporter”, “Regulator”, “*luxS*”, “Acylase” and “lactonase”. The reads count for each of these classes were normalized by the Counts per Million of Reads (CPM).

2.5.6. Correlation analysis

To determine the positive and negative correlations among the OTUs, the OTU tables generated by QIIME (one for fungal reads annotation and another for bacterial reads annotation) were merged and only the OTUs classified at genus level were processed on Microbiome Analyst environment with “Marker Data Profiling (MDP)” pipeline (Dhariwal et al., 2017) to build a Spearman’s matrix correlation.

3. Results

The HTS generated an average output of 32,597,488 PE reads per sample with Q30 average value of 90.9% (For detailed information see Supplementary Table A.2).

3.1. Taxonomic composition and diversity of cocoa metagenome

The fungal microbiota was composed mainly by members belonging to the phylum *Ascomycota* (Fig. 1A). The first 30 h of fermentation was characterized by the prevalence of the genera *Paraphaeosphaeria*, *Oidiodendron* and *Hanseniaspora*, with dominance of *Hanseniaspora* genus (Fig. 1A). From 48 h to 144 h, other genera were detected, specially the genera *Saccharomyces*, *Pichia* and *Kazachstania*. Interestingly, the genera *Paraphaeosphaeria, Oidiodendron* and *Cystoagaricus* were detected mostly during the entire fermentation process.

Regarding the bacterial profile, the first 24 h of fermentation were mainly represented by members of the *Enterobacteriaceales* order (Fig. 1B), particularly the genus *Rosenbergiella*, which was dominant from the period of 6 h to 24 h. The raw material (freshly-harvested cocoa beans – time zero h) was characterized by the presence of an *Enterobacteriaceales* member, *Enterobacter* and one AAB member, belonging to the genus *Acetobacter*. At 30 h of fermentation a shift in the bacterial composition started to take place. Although the OTUs belonging to *Enterobacteriaceales* order were still prevalent, a discrete appearance of LAB and AAB members, represented by the genera *Lactobacillus* and *Gluconobacter*, respectively, was observed (Fig. 1B). Then, the range of 48– 72 h was almost fully dominated by *Lactobacillus* species, whereas from 96 h to 144 h the AAB group prevailed, mainly with members derived from the genera *Acetobacter*, *Komagataeibacter* and *Gluconobacter*. During the range of 96–144 h, the presence of *Aquabacterium* taxon in the fermentation was also observed. Another curious fact was the presence of AAB concomitantly with the occurrence of *Enterobacteriaceales* members (96 h–144 h). At the start and at

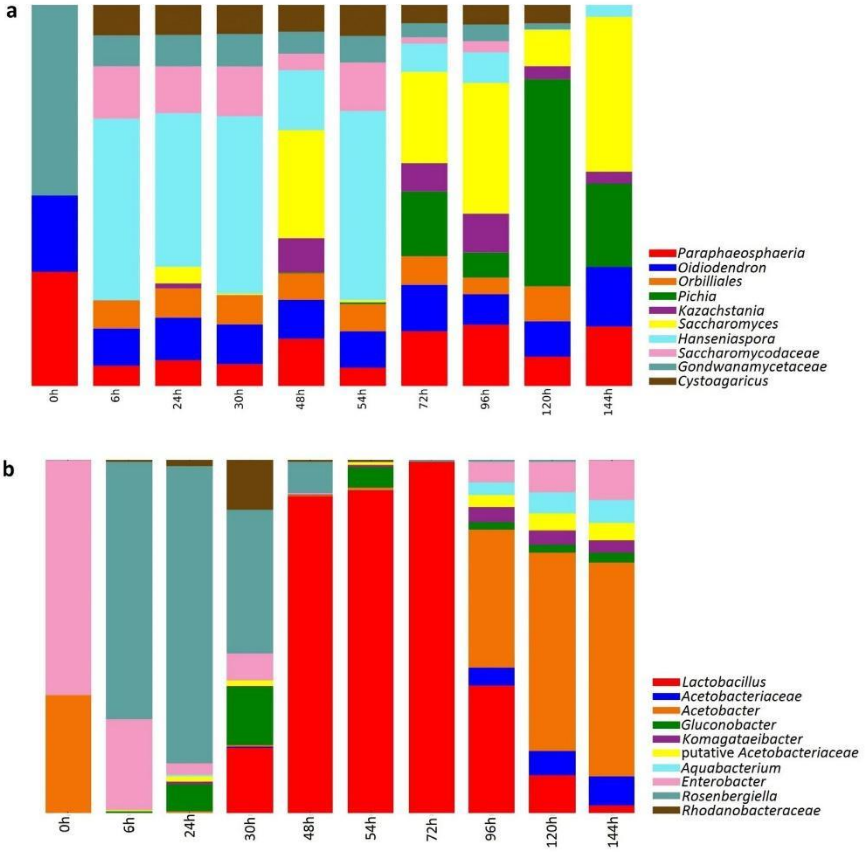


Fig. 1. Cocoa microbiota taxonomic composition with maximum taxonomical limit reached to the genus level. Fungal composition (a) and bacterial composition (b).

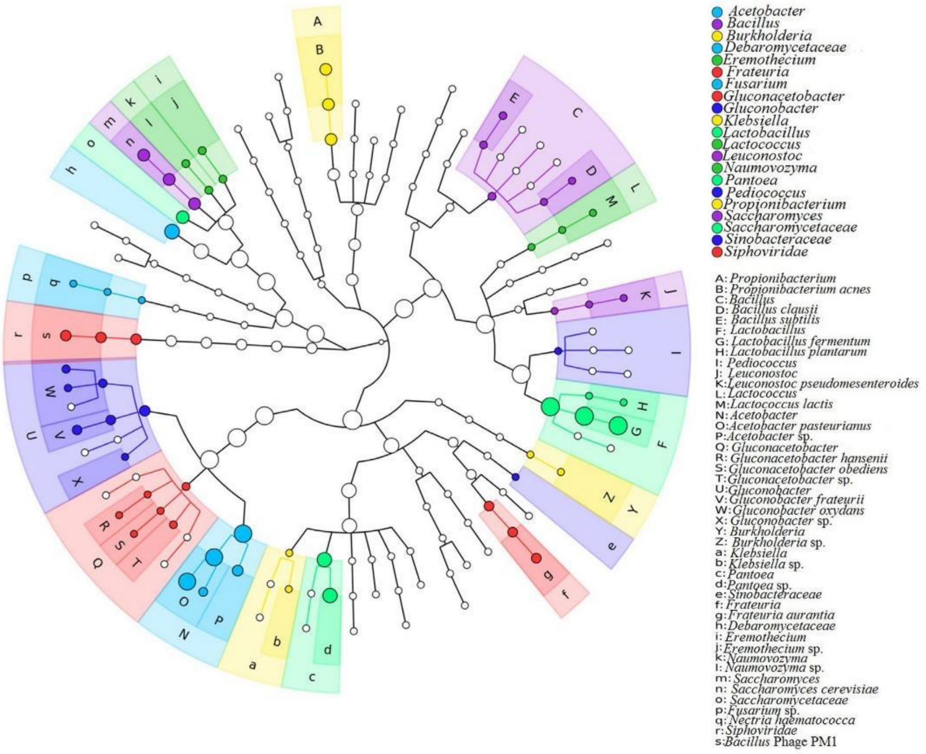


Fig. 2. Cladogram showing the species level reconstruction from specific marker genes dispersed on cocoa metagenome considering the entire microbiota. The Figure represents the OTUs with a minimal of 20% of relative abundance and a maximal of 150 representative marker genes. The external clades refer to the genus level, while the deeper clades mention to the species level taxonomy assignment. Each node represents the number of taxonomic levels that support the resolution (from kingdom to the species level) and their sizes are proportional to the number of specific marker genes that support the classification.

the end of fermentation, where AAB have been detected, the *Enterobacteriaceales* group was also found (Fig. 1B).

The taxonomic analysis, based on unique clade-specific gene markers of MetaPhlan2 database (Fig. 2), was useful to elucidate the complexity of microbiota at species level. By using this approach, we detected: (i) bacteria, represented by *Propionibacterium acnes*, *Bacillus subtilis*, *Bacillus clausii*, *Lactobacillus fermentum*, *Lactobacillus plantarum*, *Lactococcus lactis*, *Leuconostoc pseudomesenteroides*, *Acetobacter pasteurianus*, *Gluconacetobacter hansenii*, *Gluconacetobacter oboediens*, *Gluconobacter frateurii*, *Gluconobacter oxydans*, *Frateuria aurantia*; (ii) fungi, *Saccharomyces cerevisiae* and *Nectria haematococca (Fusarium solani); and* (iii) virus represented by the *Bacillus* bacteriophage PM1 (Fig. 2).

The microbial profile determined in this work was coherent, since the observed number of OTUs and the theoretical expected number of OTUs (Chao1) was paired in almost all sampling events (Fig. 3A). The α-diversity among samples showed that the cocoa microbiota is highly diverse during the whole process because almost all the samples presented an index value near of 1.0, except for the sample taken at 144 h of fermentation, which presented the lowest diversity (Fig. 3B). Considering also the results expressed by the rarefaction curve (Fig. 3C), our sampling was comprehensive enough to describe almost the entire diversity in most samples, meaning that the maximum number of OTUs have been detected by sequencing; with the exception of samples at zero h and 96 h which did not reach a plateau, (Fig. 3C).

3.2. Functional potential

The Fig. 4A shows the functional profile of the dispersed reads throughout fermentation indicating an enrichment for protein and carbohydrate metabolism, which was supported by the GAGE analysis (Supplementary Table A.3) and visualized using Pathview tool (Supplementary Figs. A.2 and A.3).

In terms of β-diversity (difference among the samples), the samples were slightly distinct, as indicated by the absence of cluster formation in the PCA graph (Fig. 4B). This result indicates that specific functional traits of the microbiota are variable with the fermentation conditions at each a given time.

To better represent the purpose of the metagenomics analysis in finding out the members of the microbiota and their functional roles, we tracked the taxonomy and functional roles of each OTU with a specific metabolic pathway (Fig. 5). In cocoa, the flavour generation generally comes from the activity of proteases and through carbohydrate metabolism (Muñoz, Cortina, Vaillant, & Parra, 2019). In fact, the taxa linked to several amino acid biosynthesis pathways were *Saccharomyces cerevisiae, Frateuria aurantia, Gluconacetobacter oxydans, Lactococcus lacticis, Bacillus subtilis, Enterobacter cloacae, Gluconacetobacter hansenii, Gluconacetobacter obediens, Bacillus clausii* and *Bacillus licheniformis* (Fig. 5: A → C, E → H). On the other hand, the carbohydrate metabolism was linked to the invertase activity from *S. cerevisiae, Lactobacillus plantarum, L. lacticis, E. cloacae, B. subtilis* and *B. clausii*

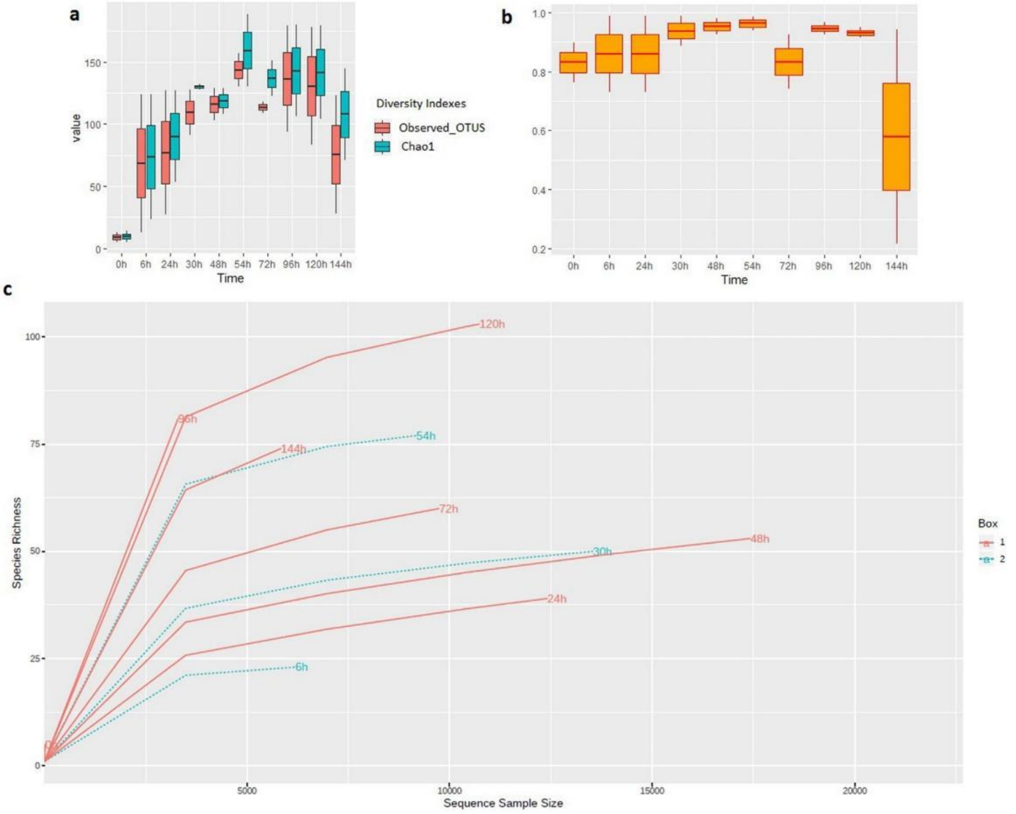


Fig. 3. Diversity indexes. Comparison among the α-diversity indexes represented by (a) observed OTUs and Chao1, (b) Simpson index and (c) rarefaction curve.

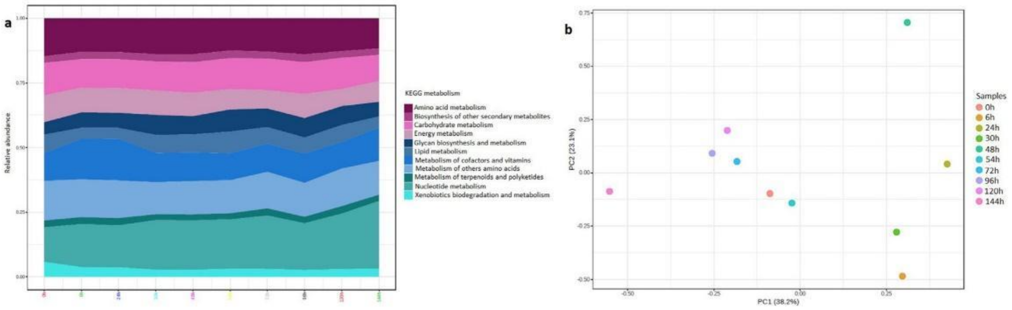


Fig. 4. Functional analysis based on short-reads alignment against UniRef90 database. (a) General stratigraphy of functional repertoire based on KEEG metabolic pathways; (b) PCA analysis showing the functional dissimilarity among the samples (β-diversity).

(Fig. 5D).

3.3. Quorum Sensing effectors and agonists' tracking

The Fig. 6A shows the oscillations of QS effectors during all fermentation steps. Two scenarios occurred in the fermentation: the first, characterized by the prevalence of transporter and regulator-associated reads, with a discrete detection of lactonase reads that were characterized by two peaks of prevalence at 6 h and 30 h, respectively; and a second one, in which the acylase and the AI-2 (*luxS*) genes prevailed in the range of 48–96 h of fermentation. Interestingly, most of these effectors could be related to the bacterial taxa present in the fermentation

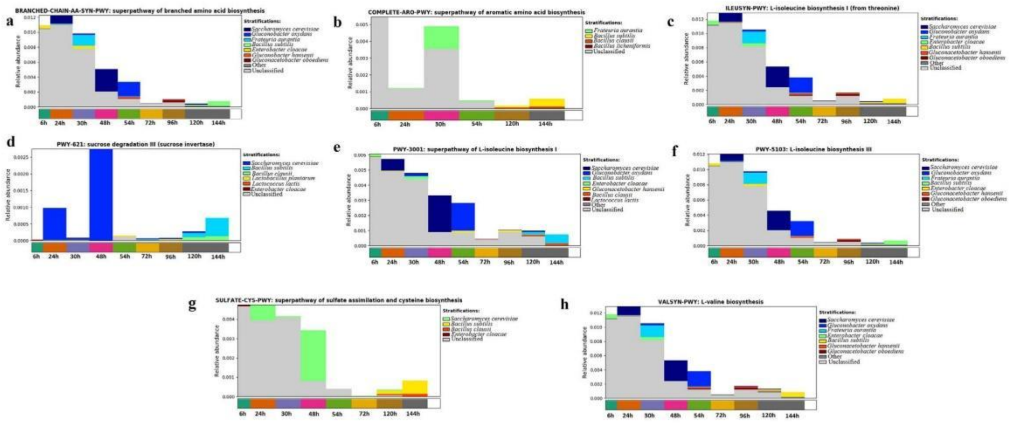


Fig. 5. Barplots showing the relationships between OTUs and their metabolic pathways. The barplots show only the OTUs and the samples in which the metabolic pathway was detected. From a to b the metabolic pathways are sorted in terms of the relative abundances of reads related to these pathways. The pathways related to amino acid metabolism are a, b, c, e, f, g and h, while the pathway d refers to carbohydrate metabolism.

(Fig. 5B), especially *Lactobacillus,* the genus that presented the higher abundance for interspecific bacterial communication (*luxS* reads), followed by *Pantoea, Bacillus* and *Enterobacter* genera which also presented *luxS* QS gene-derived reads. The AAB group, on the other hand, represented by *Komagataeibacter*, *Acetobacter* and *Gluconobacter* seemed to harbour genes for QQ, represented by the acylase category (Fig. 6B). The reads associated to lactonase (another QQ effector) were linked to *Pantoea vagans* and *Bacillus subtilis*. These reads were detected by QIIME and MetaPhlan2 analysis, even though they were not reported in the Fig. 1B (QIIME analysis) since these two taxa occurred in lower relative abundances.

As QS in fungi is poorly understood when compared to bacteria, it was hard to link a specific fungal taxon to a potential QS effector. So, to increase the probability of detecting whether fungi members could also harbour QS effectors, we searched for the fungal members detected by the taxonomic analysis (Fig. 1A) in our dataset of aligned reads against our customized database for QS detection. Surprisingly, we tracked the genera *Hanseniaspora*, *Saccharomyces* and *Pichia* as presenting putative QS and/or QQ effectors (Fig. 6C) still not yet characterized, possibly because all hits related to these microorganisms derived sequences against the custom database were related to putative proteins still not characterized deposited on the UniProt KB database by means of shared homology with known QS effectors from other species.

3.4. Correlation among microbial taxa

In general terms, the correlation among fungi and bacteria was variable and mostly represented by negative correlations (lower scores values, Fig. 7) were observed for fungi: *Cystoagaricus*, *Hanseniaspora*, *Oidiodendron* and *Paraphaeosphaeria* versus the genera *Pichia*, *Lactobacillus*, *Komagataeibacter*, *Gluconobacter*, *Aquabacterium* and *Acetobacter* (Fig. 7).

Conversely, strong positive associations were observed among *Cystoagaricus*, *Hanseniaspora* and *Rosenbergiella*; *Oidiodendron* and *Paraphaeosphaeria*; and finally, among *Acetobacter*, *Aquabacterium*, *Gluconobacter*, *Komagataeibacter*, *Lactobacillus* and *Pichia* (Fig. 7). The remaining correlations appeared to be neutral and indicate independency among the taxa co-present during the fermentation.

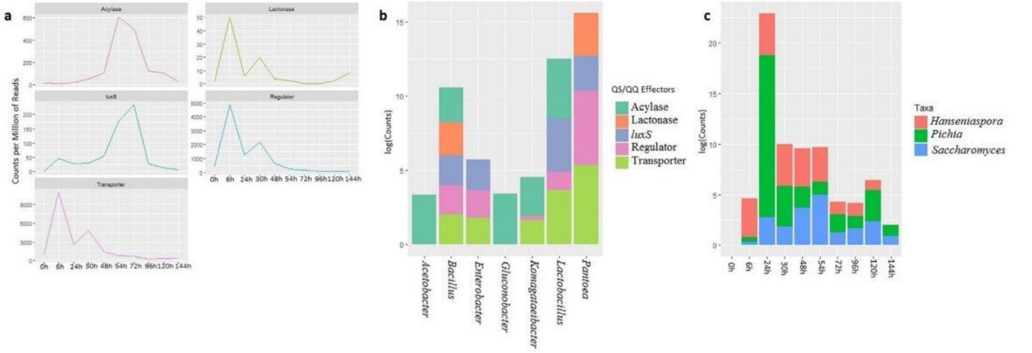


Fig. 6. Quorum sensing analysis based on short-reads alignment against a custom database with 224,713 dereplicated sequences from UniprotKB. (a) QS/QQ effectors, in Counts per Million of reads (CPM), in Bacteria and their oscillations during the fermentation process; (b) Stacked barplot of bacterial genera related to the annotated QS/QQ effectors (based on the logarithmic scaled read counts); (c) Stacked barplot of fungal taxa harbouring potential QS effectors (based on the logarithmic scaled read counts).

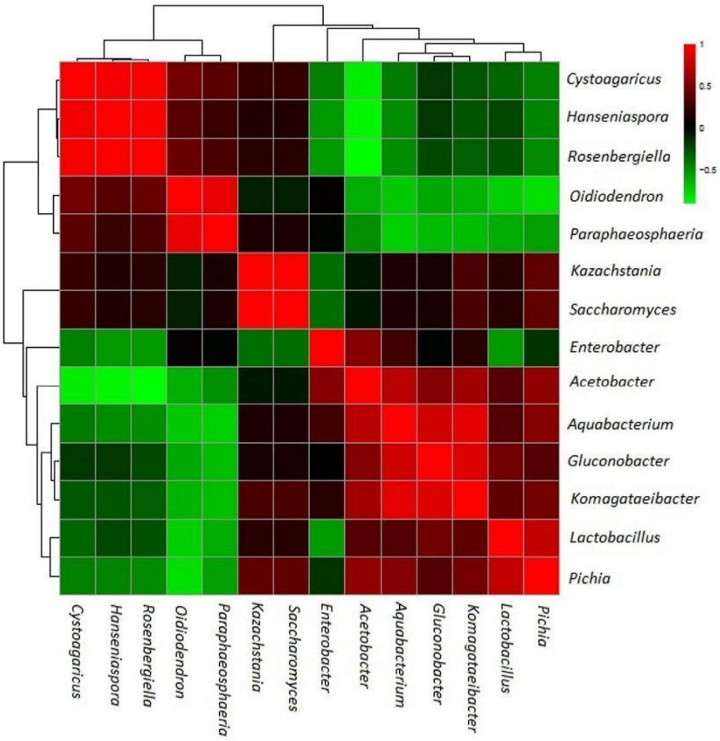


Fig. 7. Spearman correlation matrix. The Heatmap exhibits the negative, neutral and positive correlations among the microbial genera detected by taxonomic genes (16S *rRNA* and ITS) analysis on QIIME. The scale (upper corner right) represents the interactions, being the red colour a reference to a strength relationship between two taxa, while the green colour refers to negative correlation scores. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

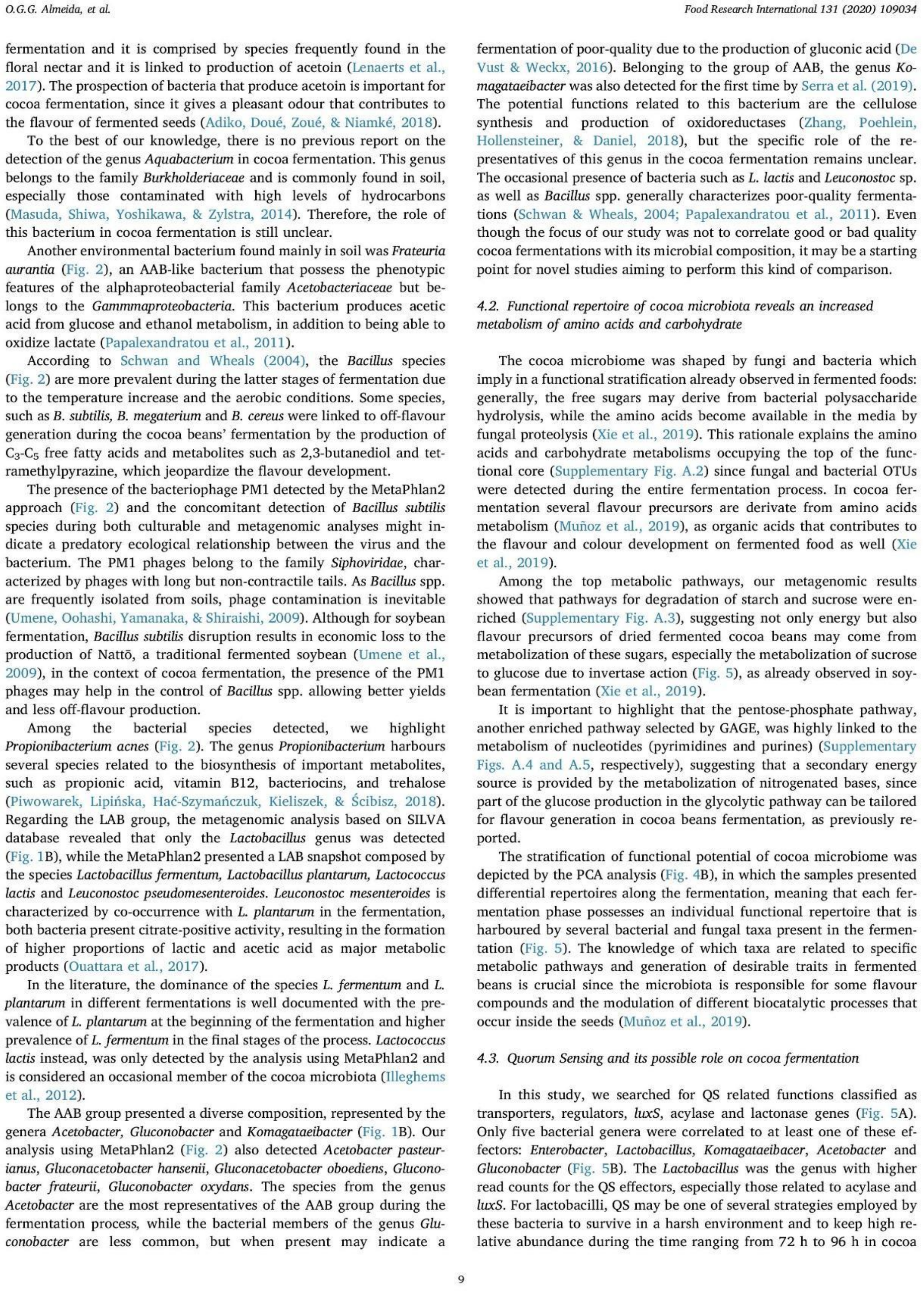
4.1. Metagenomics reveal that cocoa microbiota is highly variable and diverse

As reported in the methods, we decided to unravel the cocoa microbiota by two approaches: one based on QIIME formatted databases and another based on the MetaPhlan2, which uses specific marker genes other than only ITS or 16S *rRNA* (as used by our analysis on QIIME). In addition, these approaches are not only important to support a higher accuracy on the taxonomy, but also to increase the probability of detection of other bacterial and fungal species which may have lower taxonomic sequences deposited on the databases used in the QIIME analysis.

The metagenomics revealed fungal taxa that had been described by HTS analysis in several fermentation studies, such as *Hanseniaspora* sp., *Saccharomyces* sp. (Mota-Gutierrez et al., 2018; Serra et al., 2019), and *Kazachstania* sp. (Mota-Gutierrez et al., 2018). From fermentation studies conducted in Brazil and Mexico, the yeasts *Hanseniaspora* sp. and *Saccharomyces* sp. have been reported as the dominant taxa. The yeasts belonging to the genus *Hanseniaspora* are reported in several cocoa fermentations, with special attention to the species *H. opuntiae* that may produce monoterpenes such as nerolidol, linalool, geraniol and citronellol related to the development of floral flavour in the fermented cocoa beans (Serra et al., 2019). The yeasts belonging to the genus *Saccharomyces* are better represented by *S. cerevisiae*, which has been reported as an important producer of desirable sensorial components of the flavour in the fermented beans (Koffi, Samagaci, Goualie, & Niamke, 2017).

To the best of our knowledge, the yeast genus *Paraphaeosphaeria* and the filamentous fungi genera *Oidiodendron* and *Cystoagaricus* have not been previously detected by metataxonomics or metagenomics analysis for cocoa fermentation. The genus *Paraphaeosphaeria* is characterized as harbouring endophytic fungi species that may produce antifungal metabolites against some fruit pathogens, such as those belonging to the genus *Acremonium* (Shen et al., 2018). Regarding the genus *Oidiodendron*, it comprises cosmopolitan species commonly found in soils and several cellulose-rich substrates (as litter, pulp, paper and wood) with some species being ericoid mycorrhizal fungi (Adhikari et al., 2014). As both wood box and cocoa pulp are cellulose-rich substrates, a more in-depth study may track if the genus members detected during all fermentation phases are maintained in the fermentation due to the cocoa pulp metabolization or whether they intrinsically inhabit the cocoa box, which may explain the prevalence of this genus along the process.

According to our results, among the *Enterobacteriaceae* detected by metagenomics, the genera *Enterobacter* and *Rosenbergiella* stand out. The *Enterobacter* spp. was prevalent during the first stages of fermentation and may be related to the citric acid degradation resulting in the production of gluconic and lactic acids (Papalexandratou et al., 2013). The genus *Rosenbergiella* had not yet been described in any cocoa



fermentation and it is comprised by species frequently found in the floral nectar and it is linked to production of acetoin (Lenaerts et al., 2017). The prospection of bacteria that produce acetoin is important for cocoa fermentation, since it gives a pleasant odour that contributes to the flavour of fermented seeds (Adiko, Doué, Zoué, & Niamké, 2018).

To the best of our knowledge, there is no previous report on the detection of the genus *Aquabacterium* in cocoa fermentation. This genus belongs to the family *Burkholderiaceae* and is commonly found in soil, especially those contaminated with high levels of hydrocarbons (Masuda, Shiwa, Yoshikawa, & Zylstra, 2014). Therefore, the role of this bacterium in cocoa fermentation is still unclear.

Another environmental bacterium found mainly in soil was *Frateuria aurantia* (Fig. 2), an AAB-like bacterium that possess the phenotypic features of the alphaproteobacterial family *Acetobacteriaceae* but belongs to the *Gammmaproteobacteria*. This bacterium produces acetic acid from glucose and ethanol metabolism, in addition to being able to oxidize lactate (Papalexandratou et al., 2011).

According to Schwan and Wheals (2004), the *Bacillus* species (Fig. 2) are more prevalent during the latter stages of fermentation due to the temperature increase and the aerobic conditions. Some species, such as *B. subtilis, B. megaterium* and *B. cereus* were linked to off-flavour generation during the cocoa beans' fermentation by the production of C_3-C_5 free fatty acids and metabolites such as 2,3-butanediol and tetramethylpyrazine, which jeopardize the flavour development.

The presence of the bacteriophage PM1 detected by the MetaPhlan2 approach (Fig. 2) and the concomitant detection of *Bacillus subtilis* species during both culturable and metagenomic analyses might indicate a predatory ecological relationship between the virus and the bacterium. The PM1 phages belong to the family *Siphoviridae*, characterized by phages with long but non-contractile tails. As *Bacillus* spp. are frequently isolated from soils, phage contamination is inevitable (Umene, Oohashi, Yamanaka, & Shiraishi, 2009). Although for soybean fermentation, *Bacillus subtilis* disruption results in economic loss to the production of Nattō, a traditional fermented soybean (Umene et al., 2009), in the context of cocoa fermentation, the presence of the PM1 phages may help in the control of *Bacillus* spp. allowing better yields and less off-flavour production.

Among the bacterial species detected, we highlight *Propionibacterium acnes* (Fig. 2). The genus *Propionibacterium* harbours several species related to the biosynthesis of important metabolites, such as propionic acid, vitamin B12, bacteriocins, and trehalose (Piwowarek, Lipińska, Hać-Szymańczuk, Kieliszek, & Ścibisz, 2018). Regarding the LAB group, the metagenomic analysis based on SILVA database revealed that only the *Lactobacillus* genus was detected (Fig. 1B), while the MetaPhlan2 presented a LAB snapshot composed by the species *Lactobacillus fermentum, Lactobacillus plantarum, Lactococcus lactis* and *Leuconostoc pseudomesenteroides*. *Leuconostoc mesenteroides* is characterized by co-occurrence with *L. plantarum* in the fermentation, both bacteria present citrate-positive activity, resulting in the formation of higher proportions of lactic and acetic acid as major metabolic products (Ouattara et al., 2017).

In the literature, the dominance of the species *L. fermentum* and *L. plantarum* in different fermentations is well documented with the prevalence of *L. plantarum* at the beginning of the fermentation and higher prevalence of *L. fermentum* in the final stages of the process. *Lactococcus lactis* instead, was only detected by the analysis using MetaPhlan2 and is considered an occasional member of the cocoa microbiota (Illeghems et al., 2012).

The AAB group presented a diverse composition, represented by the genera *Acetobacter, Gluconobacter* and *Komagataeibacter* (Fig. 1B). Our analysis using MetaPhlan2 (Fig. 2) also detected *Acetobacter pasteurianus*, *Gluconacetobacter hansenii*, *Gluconacetobacter oboediens*, *Gluconobacter frateurii*, *Gluconobacter oxydans*. The species from the genus *Acetobacter* are the most representatives of the AAB group during the fermentation process, while the bacterial members of the genus *Gluconobacter* are less common, but when present may indicate a fermentation of poor-quality due to the production of gluconic acid (De Vust & Weckx, 2016). Belonging to the group of AAB, the genus *Komagataeibacter* was also detected for the first time by Serra et al. (2019). The potential functions related to this bacterium are the cellulose synthesis and production of oxidoreductases (Zhang, Poehlein, Hollensteiner, & Daniel, 2018), but the specific role of the representatives of this genus in the cocoa fermentation remains unclear. The occasional presence of bacteria such as *L. lactis* and *Leuconostoc* sp. as well as *Bacillus* spp. generally characterizes poor-quality fermentations (Schwan & Wheals, 2004; Papalexandratou et al., 2011). Even though the focus of our study was not to correlate good or bad quality cocoa fermentations with its microbial composition, it may be a starting point for novel studies aiming to perform this kind of comparison.

4.2. Functional repertoire of cocoa microbiota reveals an increased metabolism of amino acids and carbohydrate

The cocoa microbiome was shaped by fungi and bacteria which imply in a functional stratification already observed in fermented foods: generally, the free sugars may derive from bacterial polysaccharide hydrolysis, while the amino acids become available in the media by fungal proteolysis (Xie et al., 2019). This rationale explains the amino acids and carbohydrate metabolisms occupying the top of the functional core (Supplementary Fig. A.2) since fungal and bacterial OTUs were detected during the entire fermentation process. In cocoa fermentation several flavour precursors are derivate from amino acids metabolism (Muñoz et al., 2019), as organic acids that contributes to the flavour and colour development on fermented food as well (Xie et al., 2019).

Among the top metabolic pathways, our metagenomic results showed that pathways for degradation of starch and sucrose were enriched (Supplementary Fig. A.3), suggesting not only energy but also flavour precursors of dried fermented cocoa beans may come from metabolization of these sugars, especially the metabolization of sucrose to glucose due to invertase action (Fig. 5), as already observed in soybean fermentation (Xie et al., 2019).

It is important to highlight that the pentose-phosphate pathway, another enriched pathway selected by GAGE, was highly linked to the metabolism of nucleotides (pyrimidines and purines) (Supplementary Figs. A.4 and A.5, respectively), suggesting that a secondary energy source is provided by the metabolization of nitrogenated bases, since part of the glucose production in the glycolytic pathway can be tailored for flavour generation in cocoa beans fermentation, as previously reported.

The stratification of functional potential of cocoa microbiome was depicted by the PCA analysis (Fig. 4B), in which the samples presented differential repertoires along the fermentation, meaning that each fermentation phase possesses an individual functional repertoire that is harboured by several bacterial and fungal taxa present in the fermentation (Fig. 5). The knowledge of which taxa are related to specific metabolic pathways and generation of desirable traits in fermented beans is crucial since the microbiota is responsible for some flavour compounds and the modulation of different biocatalytic processes that occur inside the seeds (Muñoz et al., 2019).

4.3. Quorum Sensing and its possible role on cocoa fermentation

In this study, we searched for QS related functions classified as transporters, regulators, *luxS*, acylase and lactonase genes (Fig. 5A). Only five bacterial genera were correlated to at least one of these effectors: *Enterobacter*, *Lactobacillus*, *Komagataeibacer*, *Acetobacter* and *Gluconobacter* (Fig. 5B). The *Lactobacillus* was the genus with higher read counts for the QS effectors, especially those related to acylase and *luxS*. For lactobacilli, QS may be one of several strategies employed by these bacteria to survive in a harsh environment and to keep high relative abundance during the time ranging from 72 h to 96 h in cocoa fermentation (Fig. 1B). The role of QS in bacteria conferring adaptive fitness and maintaining cellular viability is still unknown, but some studies have observed that QS modulates an array of phenotypes, which influences microbial survival and resilience in extreme environments (Montgomery, Charlesworthy, LeBard, Visscher, & Burns, 2013).

The QS repertoire of *Lactobacillus* spp. was marked by carrying all detected QS effectors, with lower representation of "regulators" (Fig. 5B). On the other hand, AAB were only characterized by harbouring genes with acylase function (Fig. 5B). In both cases, the enrichment for acylases indicates that QS mediated by AHL (or AI-1) intraspecific QS mediated by many Gram-negative bacteria would be disrupted under these conditions. On Fig. 5A it is striking to see the enrichment for acylase genes from 48 h forward, as well as the increase in the presence of *luxS* gene which is involved in interspecies communication.

As a low number of reads were related to oxidoreductase class, no discussion about this effector could be performed. On the other hand, the lactonase activity was related to *Pantoea vagans* and *Bacillus subtilis*.

The increase in acylase activity likely influenced the decreasing levels of transporter and regulator QS-related genes after the first 48 h of fermentation (Fig. 5A). This result suggests that the majority of the QS-related genes detected during the initial phase was related to a population of Gram-negative bacteria, likely belonging to the members of the *Enterobacteriaceae* family, that disappeared after 48 h of fermentation (Fig. 1B). However, a direct link between the acylase activity and the disappearance of the *Enterobacteriaceae* family should be further confirmed by additional studies.

Following this rationale, the increase of acylase levels from 48 h to around 96 h (Fig. 5A) was likely related to *Lactobacillus* spp. dominance in this period (Fig. 1B) and may be interpreted as a response to displace other bacteria in the media through a QQ repertoire. The *luxS* gene levels related to *Lactobacillus* spp. also perform an adaptive function because the expression of this gene confers advantages in environmental changes that may lead to a stressful acid shock and/or to an oxidative stress (Johansen & Jespersen, 2017).

We noted that during the range of 96–144 h, in which the AAB group dominates the fermentation scenario, members of the *Enterobacteriaceae* family were also detected (Fig. 1B). It is possible that acylase genes related to AAB taxa may confer an adaptive response to limit QS in spoilage bacteria such as *Enterobacteriaceae*. Our results did not reveal any relationship between QS interspecific bacterial communication for AAB establishment in the fermentative scenario, but only QQ acylase genes that may be useful to AAB in order to dominate the last phases of fermentation, since after 96 h, the LAB group was replaced by AAB and *Enterobacteriaceae*. The possible reason for AAB appearance in a scenario dominated by *Lactobacillus* spp. is the cross-feeding since the LAB produces several metabolites such as lactic acid, ethanol and mannitol that may be metabolized by AAB, allowing their growth (Lee et al., 2019). The AAB maintenance during the end of the fermentation process may be further enhanced by their genetic QQ repertoire (acylase genes) which may inhibit several QS regulated phenotypes in other competing bacteria.

It is noteworthy that nutrient availability precedes the effect that QS might have on microbial establishment, because in order for a quorum of microbes to be established, bacteria need to proliferate first, which is depending upon the adequate balance of nutrients. In other words, cross-feeding does not depend on QS, while QS is dependent of cross-feeding. Microbial shifts should respond according to nutrient availability, and QS can extend the bacterial maintenance in a particular environment since it can modulate bacterial adaptation to capture more carbon and nitrogen sources (Boyle, Monaco, & Xavier, 2015).

Only the fungal taxa that could be matched to our customized database were reported in our QS analysis (Fig. 5C). To the best of our knowledge the literature reports that QS in fungi has been especially related to morphological transition to filamentous form or to yeast form in response to sesquiterpenes accumulation as it was observed for the effect of farnesol on the morphology of *Candida* spp. Aromatic alcohols such as farnesol, tyrosol, 1-phenylethanol and tryptophol are related to morphogenic changes and biofilm formation in fungi, since they act as autoinducers when the fungal growth reaches high cell densities (Padder, Prassad, & Shah, 2018). *Hanseniaspora* and *Saccharomyces* genera have been reported to have their gene expression regulated by QS (Avbelj, Zupan, & Raspor, 2016). There are scientific evidences of interkingdom communication between bacteria and fungi. It is a two-way road in which bacteria can modulate the fungal biofilm formation and fungi may reduce the bacterial growth (Barriuso, Hogan, Keshavarz, & Martínez, 2018). Curiously, matches against the custom database relating fungal reads against QS and/or QQ effectors were detected, but the effectors were still not characterized. Possibly because all hits related to these microorganisms derived sequences may share homology pieces with known QS/QQ effectors from other species already deposited on UniProt KB database, evidencing the need of more studies QS/QQ markers in fungi to allow further insights on yeasts succession and whether it affects microbial succession during the fermentation of cocoa beans.

4.4. Forces driving and shaping cocoa fermentation

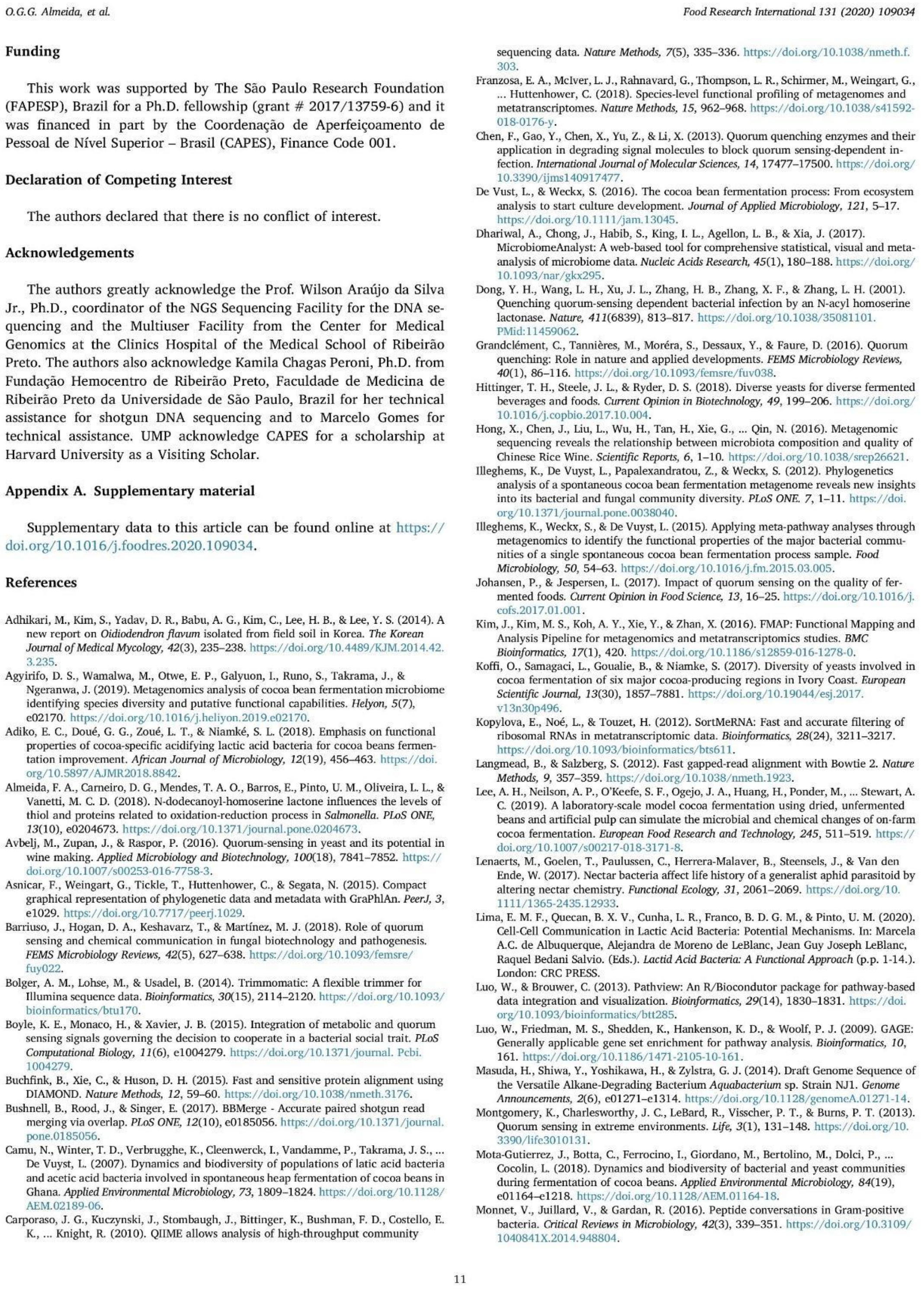
In light of the current results cocoa spontaneous fermentation may be carried out by the influence of cross-feeding and cross-talking among bacteria, and perhaps fungi. The evidences for cross-feeding reside in the well-established cooperation between *Lactobacillus* × *Kazachstania* and *Lactobacillus* × *Saccharomyces*. It is known that yeast cells from *Kazachstania* sp. and *Saccharomyces* sp. lack the gene that codes for the β-galactosidase enzyme. Therefore, these yeasts may consume free monosaccharides or lactic acid made available by the lactobacilli (Hittinger, Steele, & Ryder, 2018), which explains why the genera *Kazachstania* and *Saccharomyces* were detected only with the samples in which the genus *Lactobacillus* were also found, further decreasing their detection when *Lactobacillus* sp. decreased in relative abundance (Fig. 1A and B, 72–96 h). This observation was corroborated by the Spearman's correlation, in which *Lactobacillus* and these yeasts presented scores between zero and 0.5, suggesting a positive correlation among these taxa. At the same time, strong positive correlations among *Lactobacillus* and AAB were measured (Fig. 7), indicating an interdependence among these taxa. This may be explained by cross-feeding as the responsible for AAB emergence in the fermentation scenario.

5. Conclusions

One of the greatest contributions of this work was the identification of QS and QQ related effectors in bacteria that participate in fermentation of cocoa beans. Even though the fermentation experimental design was not ideal due to field limitations that imposed the need of two unbalanced fermentations, the results were coherent in terms of bacterial and fungal detection, allowing for the observation of microbial succession as described by several previous works. This knowledge may advance the field with the possibility of modulating the presence of desired and undesired microorganisms by selection of strains with QS and/or QQ potential to standardize the fermentation and to replace the non-desired microbiota.

CRediT authorship contribution statement

O.G.G. Almeida: Conceptualization, Formal analysis, Investigation, Methodology, Writing original draft, Writing - review & editing. **U.M. Pinto:** Conceptualization, Methodology, Writing - review & editing. **C.B. Matos:** Methodology, Supervision. **D.A. Frazilio:** Methodology. **V.F. Braga:** Methodology. **M.R. von Zeska-Kress:** Methology. **E.C.P. De Martinis:** Conceptualization, Formal analysis, Funding acquision, Methodology, Project administration, Resources, Supervision, Writing - review and editing.



Funding

This work was supported by The São Paulo Research Foundation (FAPESP), Brazil for a Ph.D. fellowship (grant # 2017/13759-6) and it was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgements

The authors greatly acknowledge the Prof. Wilson Araújo da Silva Jr., Ph.D., coordinator of the NGS Sequencing Facility for the DNA sequencing and the Multiuser Facility from the Center for Medical Genomics at the Clinics Hospital of the Medical School of Ribeirão Preto. The authors also acknowledge Kamila Chagas Peroni, Ph.D. from Fundação Hemocentro de Ribeirão Preto, Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, Brazil for her technical assistance for shotgun DNA sequencing and to Marcelo Gomes for technical assistance. UMP acknowledge CAPES for a scholarship at Harvard University as a Visiting Scholar.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2020.109034.

References

Adhikari, M., Kim, S., Yadav, D. R., Babu, A. G., Kim, C., Lee, H. B., & Lee, Y. S. (2014). A new report on *Oidiodendron flavum* isolated from field soil in Korea. *The Korean Journal of Medical Mycology, 42*(3), 235–238. https://doi.org/10.4489/KJM.2014.42.3.235.

Agyirifo, D. S., Wamalwa, M., Otwe, E. P., Galyuon, I., Runo, S., Takrama, J., & Ngeranwa, J. (2019). Metagenomics analysis of cocoa bean fermentation microbiome identifying species diversity and putative functional capabilities. *Helyon, 5*(7), e02170. https://doi.org/10.1016/j.heliyon.2019.e02170.

Adiko, E. C., Doué, G. G., Zoué, L. T., & Niamké, S. L. (2018). Emphasis on functional properties of cocoa-specific acidifying lactic acid bacteria for cocoa beans fermentation improvement. *African Journal of Microbiology, 12*(19), 456–463. https://doi.org/10.5897/AJMR2018.8842.

Almeida, F. A., Carneiro, D. G., Mendes, T. A. O., Barros, E., Pinto, U. M., Oliveira, L. L., & Vanetti, M. C. D. (2018). N-dodecanoyl-homoserine lactone influences the levels of thiol and proteins related to oxidation-reduction process in *Salmonella*. *PLoS ONE, 13*(10), e0204673. https://doi.org/10.1371/journal.pone.0204673.

Avbelj, M., Zupan, J., & Raspor, P. (2016). Quorum-sensing in yeast and its potential in wine making. *Applied Microbiology and Biotechnology, 100*(18), 7841–7852. https://doi.org/10.1007/s00253-016-7758-3.

Asnicar, F., Weingart, G., Tickle, T., Huttenhower, C., & Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ, 3*, e1029. https://doi.org/10.7717/peerj.1029.

Barriuso, J., Hogan, D. A., Keshavarz, T., & Martínez, M. J. (2018). Role of quorum sensing and chemical communication in fungal biotechnology and pathogenesis. *FEMS Microbiology Reviews, 42*(5), 627–638. https://doi.org/10.1093/femsre/fuy022.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Boyle, K. E., Monaco, H., & Xavier, J. B. (2015). Integration of metabolic and quorum sensing signals governing the decision to cooperate in a bacterial social trait. *PLoS Computational Biology, 11*(6), e1004279. https://doi.org/10.1371/journal. Pcbi.1004279.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*, 59–60. https://doi.org/10.1038/nmeth.3176.

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge - Accurate paired shotgun read merging via overlap. *PLoS ONE, 12*(10), e0185056. https://doi.org/10.1371/journal.pone.0185056.

Camu, N., Winter, T. D., Verbrugghe, K., Cleenwerck, I., Vandamme, P., Takrama, J. S., ... De Vuyst, L. (2007). Dynamics and biodiversity of populations of latic acid bacteria and acetic acid bacteria involved in spontaneous heap fermentation of cocoa beans in Ghana. *Applied Environmental Microbiology, 73*, 1809–1824. https://doi.org/10.1128/AEM.02189-06.

Carporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303.

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., ... Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods, 15*, 962–968. https://doi.org/10.1038/s41592-018-0176-y.

Chen, F., Gao, Y., Chen, X., Yu, Z., & Li, X. (2013). Quorum quenching enzymes and their application in degrading signal molecules to block quorum sensing-dependent infection. *International Journal of Molecular Sciences, 14*, 17477–17500. https://doi.org/10.3390/ijms140917477.

De Vust, L., & Weckx, S. (2016). The cocoa bean fermentation process: From ecosystem analysis to start culture development. *Journal of Applied Microbiology, 121*, 5–17. https://doi.org/10.1111/jam.13045.

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., & Xia, J. (2017). MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research, 45*(1), 180–188. https://doi.org/10.1093/nar/gkx295.

Dong, Y. H., Wang, L. H., Xu, J. L., Zhang, H. B., Zhang, X. F., & Zhang, L. H. (2001). Quenching quorum-sensing dependent bacterial infection by an N-acyl homoserine lactonase. *Nature, 411*(6839), 813–817. https://doi.org/10.1038/35081101. PMid:11459062.

Grandclément, C., Tannières, M., Moréra, S., Dessaux, Y., & Faure, D. (2016). Quorum quenching: Role in nature and applied developments. *FEMS Microbiology Reviews, 40*(1), 86–116. https://doi.org/10.1093/femsre/fuv038.

Hittinger, T. H., Steele, J. L., & Ryder, D. S. (2018). Diverse yeasts for diverse fermented beverages and foods. *Current Opinion in Biotechnology, 49*, 199–206. https://doi.org/10.1016/j.copbio.2017.10.004.

Hong, X., Chen, J., Liu, L., Wu, H., Tan, H., Xie, G., ... Qin, N. (2016). Metagenomic sequencing reveals the relationship between microbiota composition and quality of Chinese Rice Wine. *Scientific Reports, 6*, 1–10. https://doi.org/10.1038/srep26621.

Illeghems, K., De Vuyst, L., Papalexandratou, Z., & Weckx, S. (2012). Phylogenetics analysis of a spontaneous cocoa bean fermentation metagenome reveals new insights into its bacterial and fungal community diversity. *PLoS ONE. 7*, 1–11. https://doi.org/10.1371/journal.pone.0038040.

Illeghems, K., Weckx, S., & De Vuyst, L. (2015). Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. *Food Microbiology, 50*, 54–63. https://doi.org/10.1016/j.fm.2015.03.005.

Johansen, P., & Jespersen, L. (2017). Impact of quorum sensing on the quality of fermented foods. *Current Opinion in Food Science, 13*, 16–25. https://doi.org/10.1016/j.cofs.2017.01.001.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., & Zhan, X. (2016). FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics, 17*(1), 420. https://doi.org/10.1186/s12859-016-1278-0.

Koffi, O., Samagaci, L., Goualie, B., & Niamke, S. (2017). Diversity of yeasts involved in cocoa fermentation of six major cocoa-producing regions in Ivory Coast. *European Scientific Journal, 13*(30), 1857–7881. https://doi.org/10.19044/esj.2017.v13n30p496.

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics, 28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Lee, A. H., Neilson, A. P., O'Keefe, S. F., Ogejo, J. A., Huang, H., Ponder, M., ... Stewart, A. C. (2019). A laboratory-scale model cocoa fermentation using dried, unfermented beans and artificial pulp can simulate the microbial and chemical changes of on-farm cocoa fermentation. *European Food Research and Technology, 245*, 511–519. https://doi.org/10.1007/s00217-018-3171-8.

Lenaerts, M., Goelen, T., Paulussen, C., Herrera-Malaver, B., Steensels, J., & Van den Ende, W. (2017). Nectar bacteria affect life history of a generalist aphid parasitoid by altering nectar chemistry. *Functional Ecology, 31*, 2061–2069. https://doi.org/10.1111/1365-2435.12933.

Lima, E. M. F., Quecan, B. X. V., Cunha, L. R., Franco, B. D. G. M., & Pinto, U. M. (2020). Cell-Cell Communication in Lactic Acid Bacteria: Potential Mechanisms. In: Marcela A.C. de Albuquerque, Alejandra de Moreno de LeBlanc, Jean Guy Joseph LeBlanc, Raquel Bedani Salvio. (Eds.). *Lactid Acid Bacteria: A Functional Approach* (p.p. 1-14.). London: CRC PRESS.

Luo, W., & Brouwer, C. (2013). Pathview: An R/Biocondutor package for pathway-based data integration and visualization. *Bioinformatics, 29*(14), 1830–1831. https://doi.org/10.1093/bioinformatics/btt285.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., & Woolf, P. J. (2009). GAGE: Generally applicable gene set enrichment for pathway analysis. *Bioinformatics, 10*, 161. https://doi.org/10.1186/1471-2105-10-161.

Masuda, H., Shiwa, Y., Yoshikawa, H., & Zylstra, G. J. (2014). Draft Genome Sequence of the Versatile Alkane-Degrading Bacterium *Aquabacterium* sp. Strain NJ1. *Genome Announcements, 2*(6), e01271–e1314. https://doi.org/10.1128/genomeA.01271-14.

Montgomery, K., Charlesworthy, J. C., LeBard, R., Visscher, P. T., & Burns, P. T. (2013). Quorum sensing in extreme environments. *Life, 3*(1), 131–148. https://doi.org/10.3390/life3010131.

Mota-Gutierrez, J., Botta, C., Ferrocino, I., Giordano, M., Bertolino, M., Dolci, P., ... Cocolin, L. (2018). Dynamics and biodiversity of bacterial and yeast communities during fermentation of cocoa beans. *Applied Environmental Microbiology, 84*(19), e01164–e1218. https://doi.org/10.1128/AEM.01164-18.

Monnet, V., Juillard, V., & Gardan, R. (2016). Peptide conversations in Gram-positive bacteria. *Critical Reviews in Microbiology, 42*(3), 339–351. https://doi.org/10.3109/1040841X.2014.948804.

Muñoz, M. S., Cortina, J. R., Vaillant, F. E., & Parra, S. E. (2019). An overview of the physical and biochemical transformation of cocoa seeds to beans and to chocolate: Flavor formation. *Critical Reviews in Food Science and Nutrition, 21*, 1–21. https://doi.org/10.1080/10408398.2019.1581726.

Nilsson, R. H., Larsson, K. H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., ... Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research, 47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022.

Ouattara, H. D., Ouattara, H. G., Droux, M., Reverchon, S., Nasser, W., & Niamke, S. L. (2017). Lactic acid bacteria involved in cocoa beans fermentation from Ivory Coast: Species diversity and citrate lyase production. *International Journal of Food Microbiology, 256*, 11–19. https://doi.org/10.1016/j.ijfoodmicro.2017.05.008.

Padder, S. A., Prassad, R., & Shah, A. H. (2018). Quorum sensing: A less known mode of communication among fungi. *Microbiology Research, 210*, 51–58. https://doi.org/10.1016/j.micres.2018.03.007.

Papalexandratou, Z., Falony, G., Romanens, E., Jimenez, J. C., Amores, F., Daniel, H. M., & De Vuyst, L. (2011). Species diversity, community dynamics, and metabolite kynetics of the microbiota associated with traditional Ecuadorian spontaneous cocoa bean fermentations. *Applied and Environmental Microbiology, 77*(21), 7698–7714. https://doi.org/10.1128/AEM.05523-11.

Papalexandratou, Z., & De Vuyst, L. (2011). Assessment of the yeast species composition of cocoa bean fermentations in different cocoa-producing regions using denaturing gradient gel electrophoresis. *FEMS Yeast Research, 11*, 564–574. https://doi.org/10.1111/j.1567-1364.2011.00747.x.

Papalexandratou, Z., Lefber, T., Bahrim, B., Lee, O. S., Daniel, H. D., & De Vuyst, L. (2013). *Hanseniaspora opuntiae*, *Saccharomyces cerevisiae*, *Lactobacillus fermentum*, and *Acetobacter pasteurianus* predominate during well-performed Malaysian cocoa bean box fermentations, underlining the importance of these microbial species for a successful cocoa bean fermentation process. *Food Microbiology, 35*, 73–85. https://doi.org/10.1016/j.fm.2013.02.015.

Papalexandratou, Z., & Nielsen, D. S. (2016). It's getting' hot in here: breeding robust yeast starter cultures for cocoa fermentation. *Trends in Microbiology, 24*, 168–170. https://doi.org/10.1016/j.tim.2016.01.003.

Pereira, C. S., Thompson, J. A., & Xavier, K. B. (2012). AI-2-mediated signalling in bacteria. *FEMS Microbiology Reviews, 37*(2), 156–181. https://doi.org/10.1111/j.1574-6976.2012.00345.x.

Piwowarek, K., Lipińska, E., Hać-Szymańczuk, E., Kieliszek, M., & Ścibisz, I. (2018). *Propionibacterium* spp.—source of propionic acid, vitamin B12, and other metabolites important for the industry. *Applied Microbiology and Biotechnology, 102*(2), 515–538. https://doi.org/10.1007/s00253-017-8616-7.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research, 41*, D590–D596. https://doi.org/10.1093/nar/gks1219.

Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research, 38*(20), e191. https://doi.org/10.1093/nar/gkq747.

Schwan, R. F., & Wheals, A. E. (2004). The microbiology of cocoa fermentation and its role in chocolate quality. *Critical Reviews in Food Science and Nutrition, 44*, 1–17. https://doi.org/10.1080/10408690490464104.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods, 8*, 811–814. https://doi.org/10.1038/nmeth.2066.

Serra, J. M., Moura, F., Pereira, G. V. M., Soccol, C. R., Rogez, H., & Darnet, S. (2019). Determination of the microbial community in Amazonian cocoa bean fermentation by Illumina-based metagenomic sequencing. *LWT – Food Science and Technology, 106*, 229–239. https://doi.org/10.1016/j.lwt.2019.02.038.

Shen, Y., Nie, J., Li, Z., Wu, Y., Dong, Y., & Zhang, J. (2018). Differentiated surface fungal communities at point of harvest on apple fruits from rural and peri-urban orchards. *Scientific Reports, 8*(1), 2165. https://doi.org/10.1038/s41598-017-17436-5.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics, 23*(10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

Thoendel, M., Kavanaugh, J. S., Flack, C. E., & Horswill, A. R. (2011). Peptide signaling in the staphylococci. *Chemical Reviews, 111*(1), 117–151. https://doi.org/10.1021/cr100370n.

Umene, K., Oohashi, S., Yamanaka, F., & Shiraishi, A. (2009). Molecular characterization of the genome of *Bacillus subtilis* (natto) bacteriophage PM1, a phage associated with disruption of food production. *World Journal of Microbiology and Biotechnology, 25*(10), 1877. https://doi.org/10.1007/s11274-009-0086-3.

Xie, M., Wu, J., An, F., Yue, X., Tao, D., Wu, R., & Lee, Y. (2019). An integrated metagenomic/metaproteomic investigation of microbiota in dajiang-meju, a traditional fermented soybean product in Northeast China. *Food Research International, 115*, 414–424. https://doi.org/10.1016/j.foodres.2018.10.076.

Zhang, Q., Poehlein, A., Hollensteiner, J., & Daniel, R. (2018). Draft genome sequence of Komagataeibacter maltaceti LMG 1529T, a vinegar-producing acetic acid bacterium isolated from malt vinegar brewery acetifiers. *Genome Announcements, 6*(16), e00330–e418. https://doi.org/10.1128/genomeA.00330-18.

Supplementary material

Does Quorum Sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An *in silico* perspective

O.G.G. Almeida[a], U.M. Pinto[b,1], C.B. Matos[c], D.A. Frazilio[a], V.F. Braga[a], M.R. von Zeska-Kress[a], E.C.P. De Martinis[a,*]

[a]Universidade de São Paulo – Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Brazil

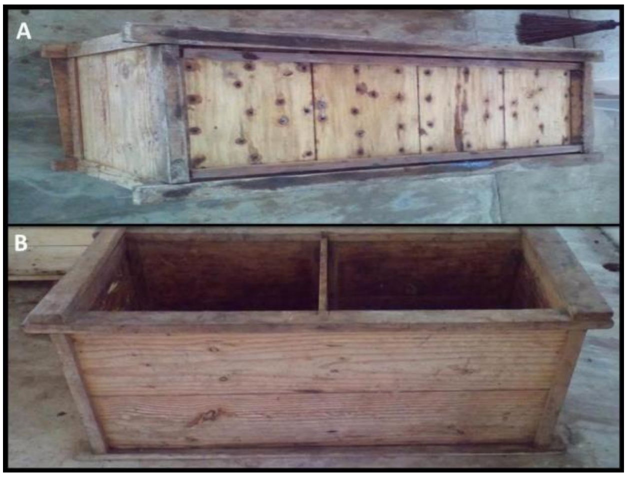
[b]Food Research Center, Universidade de São Paulo – Faculdade de Ciências Farmacêuticas, Brazil

[c]Comissão Executiva do Plano da Lavoura Cacaueira- Centro de Pesquisas do Cacau (CEPLAC-CEPEC), Rod. Jorge Amado, 22 - Alto Mirante, Itabuna, BA, Brazil

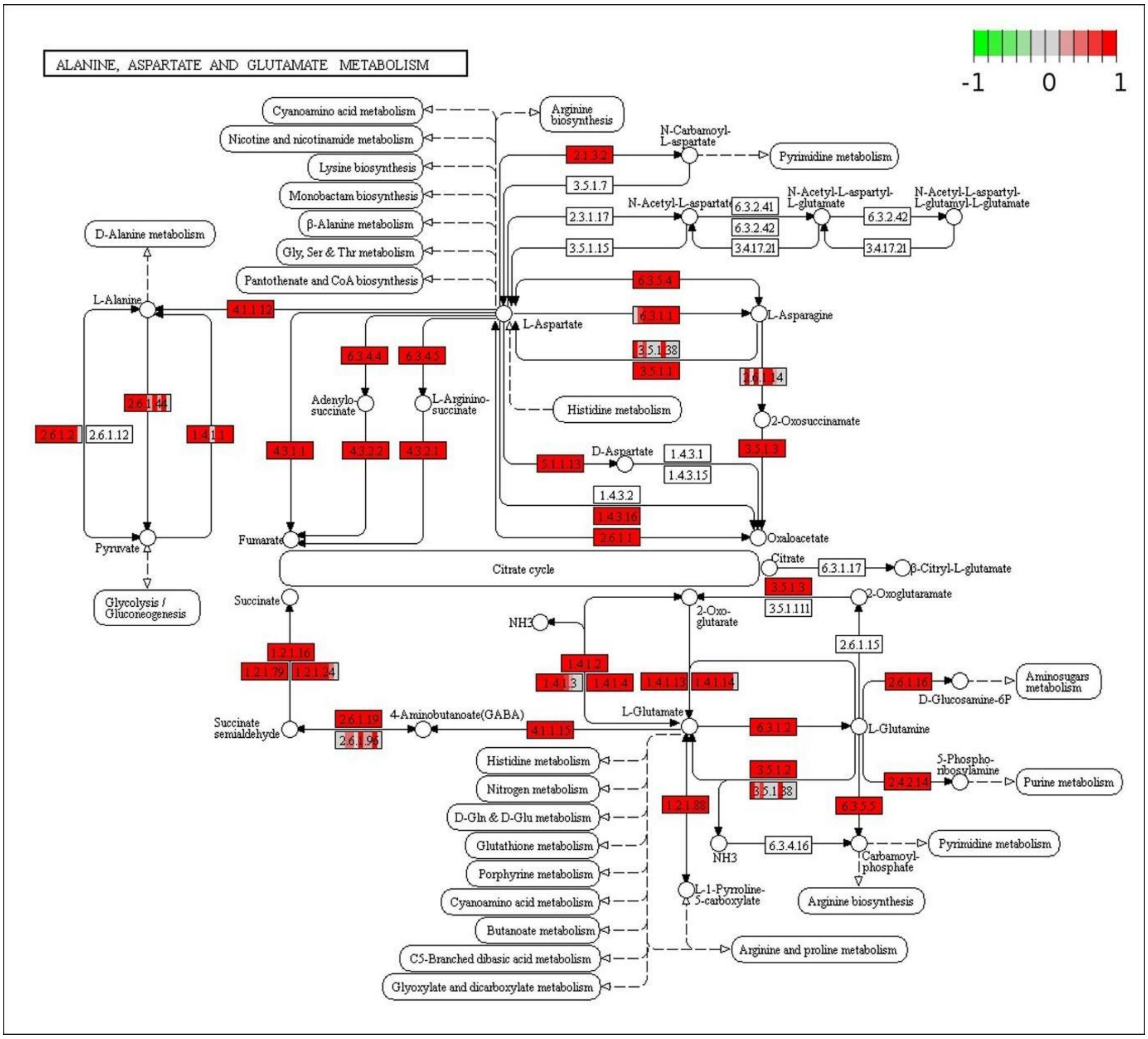
∗Corresponding author at: Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Avenida do Café s/n, Monte Alegre, Ribeirão Preto 14040-903, Brazil.

E-mail address: edemarti@usp.br (E.C.P. De Martinis).

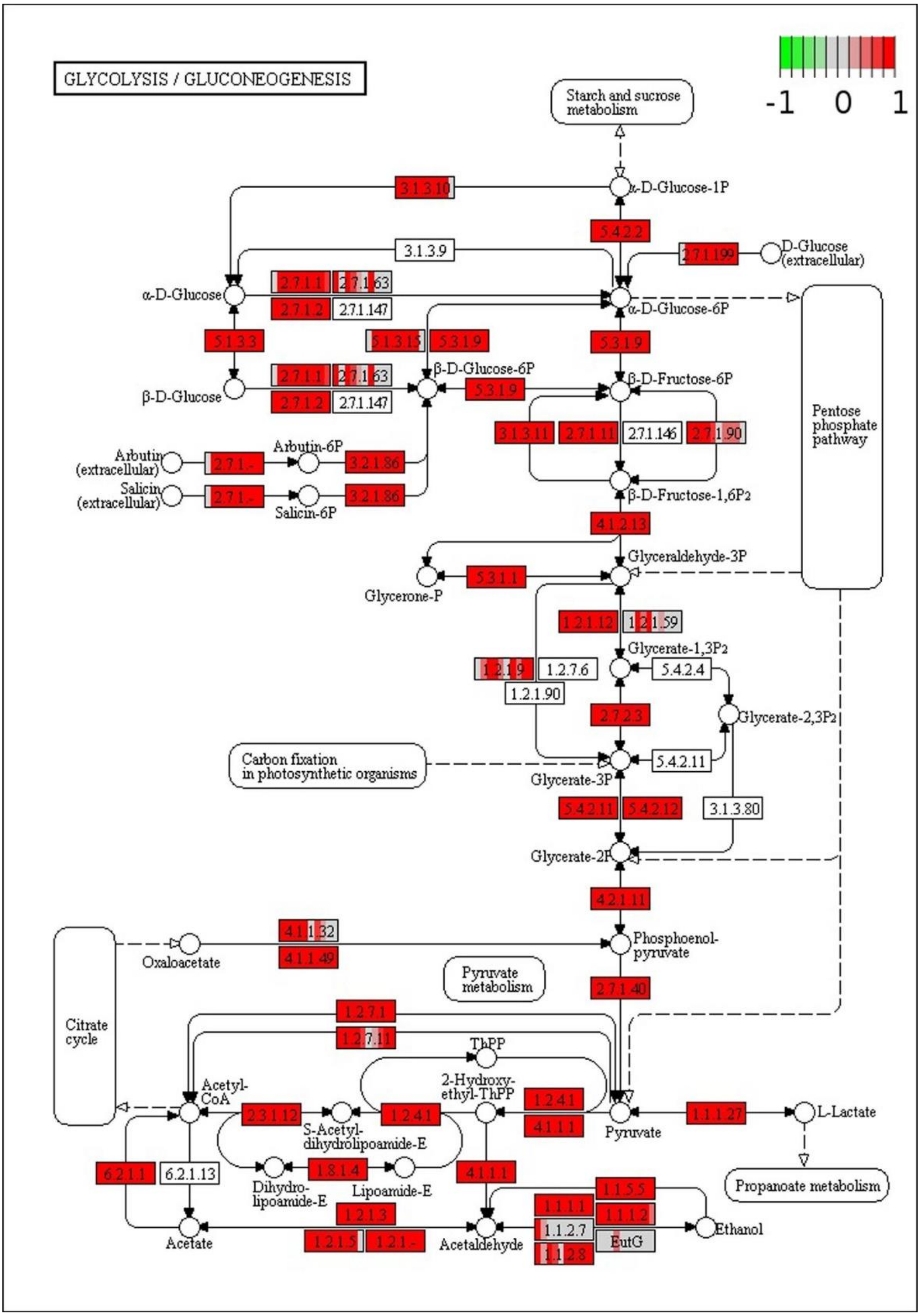
Supplementary Fig. A.1. Cocoa fermentation boxes. (A) The boxes are divided in the middle to allow cocoa mixing and to facilitate aeration and homogenization of the cocoa microbiota. (B) The bottom of the boxes is characterized by small holes that facilitate the cocoa honey leakage, which allows acceleration of the beans' drying process.



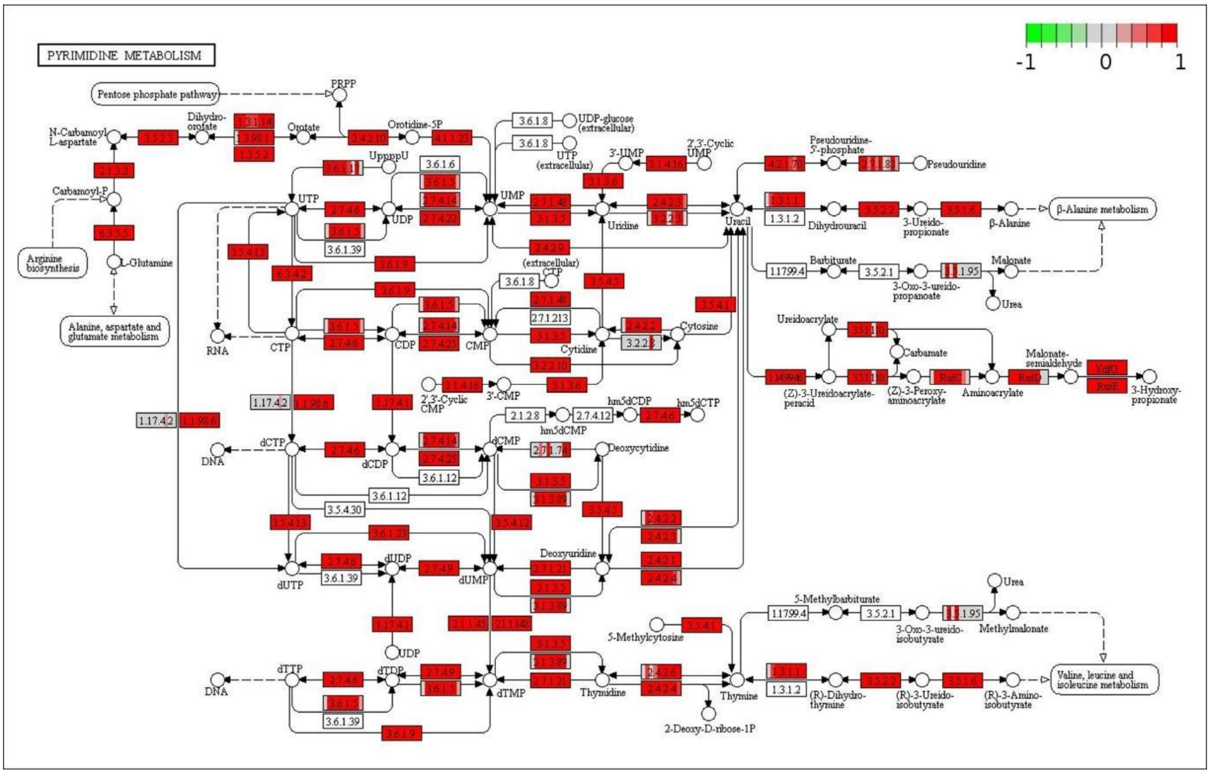
Supplementary Fig. A.2. Enriched alanine, aspartate, and glutamate amino acids metabolism. From yellow to red (−1 to 1) indicates an increased abundance of enzymatic repertoire that composes the entire pathway.



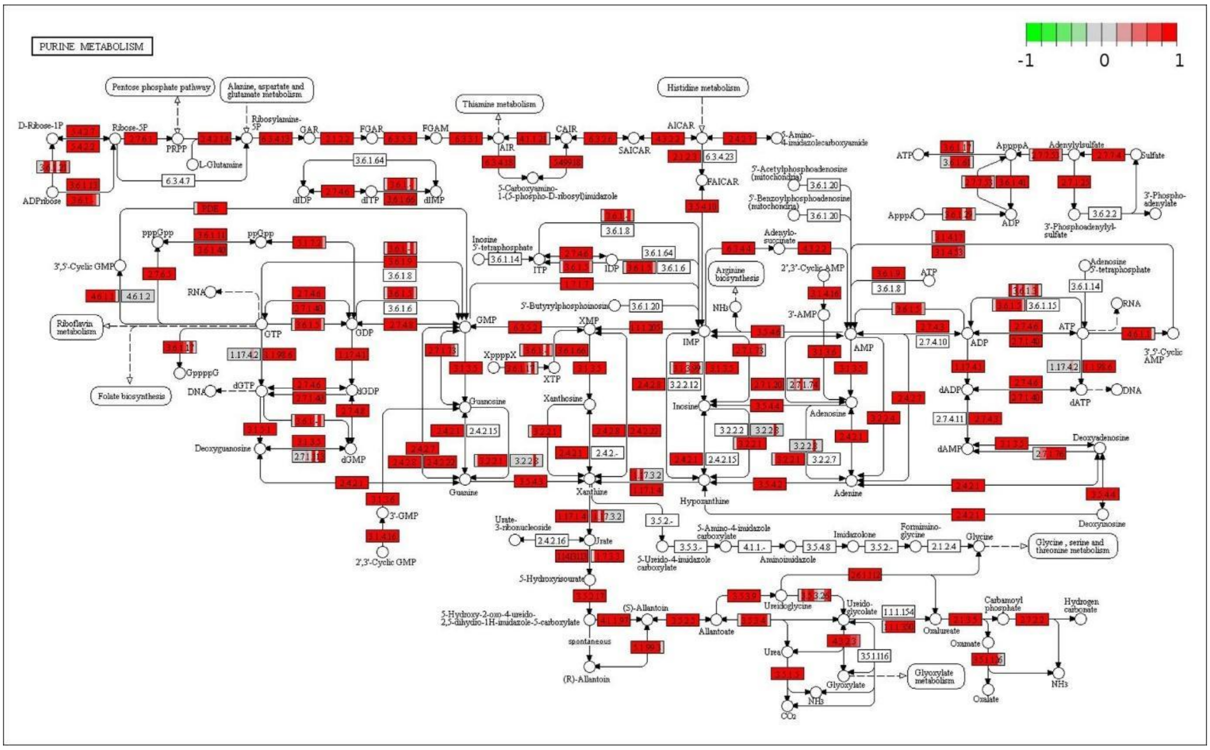
Supplementary Fig. A.3. Enriched glycolysis pathway related to metabolism of starch and sucrose carbohydrates. From yellow to red (−1 to 1) indicates an increased abundance of enzymatic repertoire that composes the entire pathway.



Supplementary Fig. A.4. Enriched pentose-phosphate pathway related to pyrimidine’s metabolism. From yellow to red (−1 to 1) indicates an increased abundance of enzymatic repertoire that composes the entire pathway.



Supplementary Fig. A.5. Enriched pentose-phosphate pathway related to purine metabolism. From yellow to red (−1 to 1) indicates an increased abundance of enzymatic repertoire that composes the entire pathway.



Supplementary Table A.1. Physical parameters of temperature (°C) and humidity of the fermentation environment.

Sampling time (h)	Air humidity (%)	Box temperature (°C)	Environmental temperature (°C)
Fermentation start	87	25.4	26.5
6	50	23.4	31.9
24	50	26.7	31.9
30	55	30	30.3
48	55	30.6	30.3
54	59	31.4	32.3
72	59	39.8	32,2
96	55	45.8	32.8
120	62	45	31.3
144	64	46.5	31.5

Supplementary Table A.2. Metagenomic DNA sequencing outputs.

Sample	Number of reads per sample	Number of base pairs per sample	Number of generated clusters	File size
Zero	20,857,138	1,585,142,488	10,428,569	1.02 GB
6h	29,581,476	2,248,192,176	14,790,738	1.45 GB
24h	35,645,210	2,709,035,960	17,822,605	1.74 GB
30h	37,720,186	2,866,734,136	18,860,093	1.84 GB
48h	31,166,106	2,368,624,056	15,583,053	1.52 GB
54h	28,497,444	2,165,805,744	14,248,722	1.41 GB
72h	28,860,872	2,193,426,272	14,430,436	1.41 GB
96h	35,523,556	2,699,790,256	17,761,778	1.76 GB
120h	47,113,616	3,580,634,816	23,556,808	2.31 GB
144h	31,009,274	2,356,704,824	15,504,637	1.52 GB

Supplementary Table A.3. GAGE metabolic pathways enrichment report. The table shows the statistics regarding each putative KEGG pathway measured in the cocoa microbiome.

It was made available online since it has a high size and cannot be fitted in this document. As it is publicly available, it can be assessed at this link for download and conference.

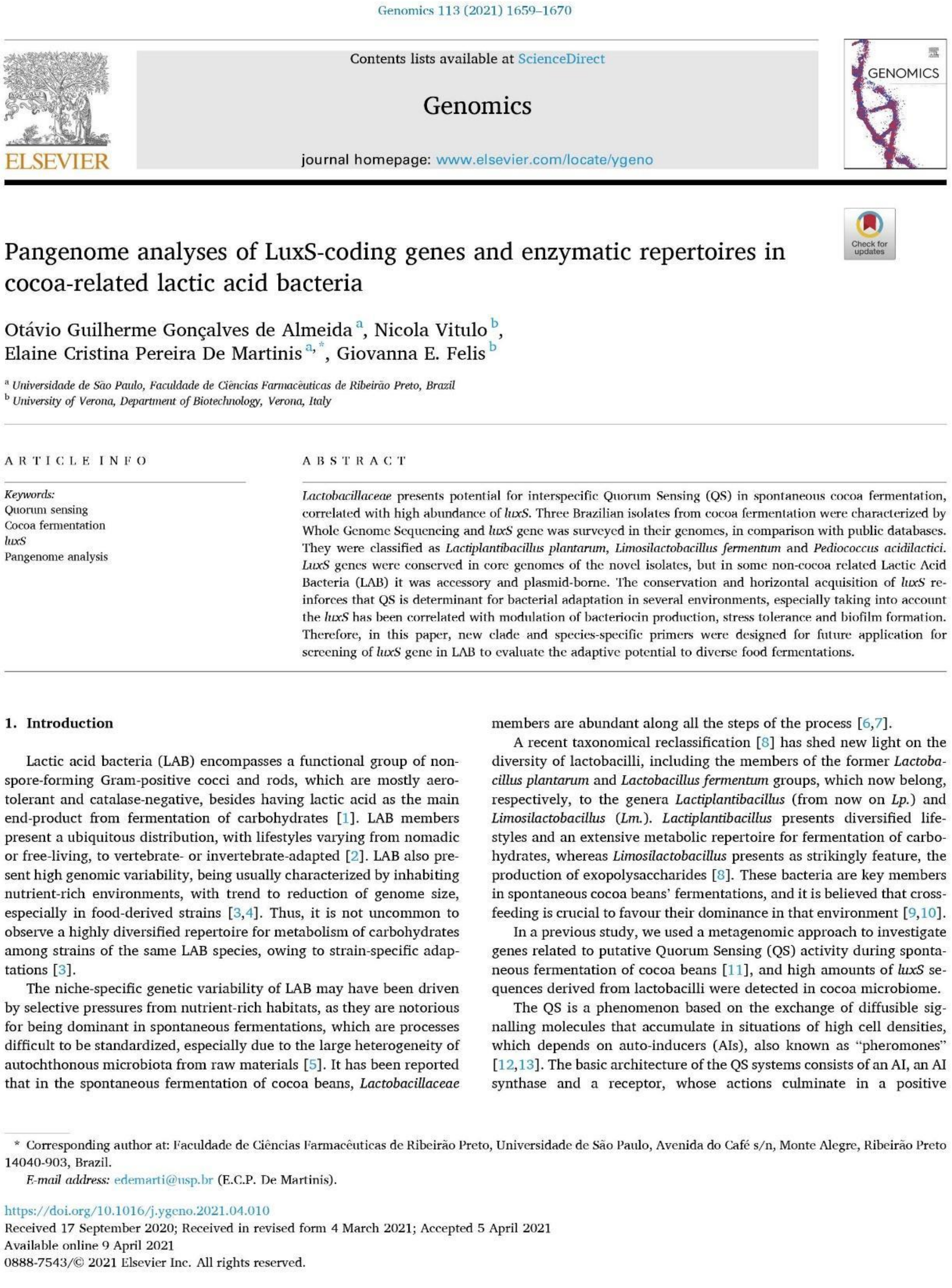
Chapter 2

4. Chapter 2 - Pangenome analyses of LuxS-coding genes and enzymatic repertoires in cocoa-related lactic acid bacteria

Specific objective

- To scrutinize the presence and distribution of *luxS* gene in lactic acid bacteria, focusing on the species *Lactiplantibacillus plantarum*, *Limosilactobacillus fermentum*, and *Pediococcus acidilactici* by comparative genome analysis. Moreover, this work aimed to identify conservative *luxS* gene sequences for primer design to allow a rapid-based PCR screening of *luxS* genes in these LAB species. Finally, it also aimed to compare metabolic pathways of bacterial metabolism of cocoa-related strains versus other LAB strains from different niches.

The article was published in the Journal *Genomics* (doi: https://doi.org/10.1016/j.ygeno.2021.04.010) a respected Journal in the field of comparative genomic analysis. The authorization for reuse in Thesis and Dissertations is depicted in the Attachment C.



Genomics 113 (2021) 1659–1670

Contents lists available at ScienceDirect

Genomics

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

Pangenome analyses of LuxS-coding genes and enzymatic repertoires in cocoa-related lactic acid bacteria

Otávio Guilherme Gonçalves de Almeida [a], Nicola Vitulo [b], Elaine Cristina Pereira De Martinis [a,*], Giovanna E. Felis [b]

[a] *Universidade de São Paulo, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Brazil*
[b] *University of Verona, Department of Biotechnology, Verona, Italy*

ARTICLE INFO

Keywords:
Quorum sensing
Cocoa fermentation
luxS
Pangenome analysis

ABSTRACT

Lactobacillaceae presents potential for interspecific Quorum Sensing (QS) in spontaneous cocoa fermentation, correlated with high abundance of *luxS*. Three Brazilian isolates from cocoa fermentation were characterized by Whole Genome Sequencing and *luxS* gene was surveyed in their genomes, in comparison with public databases. They were classified as *Lactiplantibacillus plantarum*, *Limosilactobacillus fermentum* and *Pediococcus acidilactici*. *LuxS* genes were conserved in core genomes of the novel isolates, but in some non-cocoa related Lactic Acid Bacteria (LAB) it was accessory and plasmid-borne. The conservation and horizontal acquisition of *luxS* reinforces that QS is determinant for bacterial adaptation in several environments, especially taking into account the *luxS* has been correlated with modulation of bacteriocin production, stress tolerance and biofilm formation. Therefore, in this paper, new clade and species-specific primers were designed for future application for screening of *luxS* gene in LAB to evaluate the adaptive potential to diverse food fermentations.

1. Introduction

Lactic acid bacteria (LAB) encompasses a functional group of non-spore-forming Gram-positive cocci and rods, which are mostly aerotolerant and catalase-negative, besides having lactic acid as the main end-product from fermentation of carbohydrates [1]. LAB members present a ubiquitous distribution, with lifestyles varying from nomadic or free-living, to vertebrate- or invertebrate-adapted [2]. LAB also present high genomic variability, being usually characterized by inhabiting nutrient-rich environments, with trend to reduction of genome size, especially in food-derived strains [3,4]. Thus, it is not uncommon to observe a highly diversified repertoire for metabolism of carbohydrates among strains of the same LAB species, owing to strain-specific adaptations [3].

The niche-specific genetic variability of LAB may have been driven by selective pressures from nutrient-rich habitats, as they are notorious for being dominant in spontaneous fermentations, which are processes difficult to be standardized, especially due to the large heterogeneity of autochthonous microbiota from raw materials [5]. It has been reported that in the spontaneous fermentation of cocoa beans, *Lactobacillaceae* members are abundant along all the steps of the process [6,7].

A recent taxonomical reclassification [8] has shed new light on the diversity of lactobacilli, including the members of the former *Lactobacillus plantarum* and *Lactobacillus fermentum* groups, which now belong, respectively, to the genera *Lactiplantibacillus* (from now on *Lp.*) and *Limosilactobacillus* (*Lm.*). *Lactiplantibacillus* presents diversified lifestyles and an extensive metabolic repertoire for fermentation of carbohydrates, whereas *Limosilactobacillus* presents as strikingly feature, the production of exopolysaccharides [8]. These bacteria are key members in spontaneous cocoa beans' fermentations, and it is believed that cross-feeding is crucial to favour their dominance in that environment [9,10].

In a previous study, we used a metagenomic approach to investigate genes related to putative Quorum Sensing (QS) activity during spontaneous fermentation of cocoa beans [11], and high amounts of *luxS* sequences derived from lactobacilli were detected in cocoa microbiome.

The QS is a phenomenon based on the exchange of diffusible signalling molecules that accumulate in situations of high cell densities, which depends on auto-inducers (AIs), also known as "pheromones" [12,13]. The basic architecture of the QS systems consists of an AI, an AI synthase and a receptor, whose actions culminate in a positive

* Corresponding author at: Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Avenida do Café s/n, Monte Alegre, Ribeirão Preto 14040-903, Brazil.
E-mail address: edemarti@usp.br (E.C.P. De Martinis).

https://doi.org/10.1016/j.ygeno.2021.04.010
Received 17 September 2020; Received in revised form 4 March 2021; Accepted 5 April 2021
Available online 9 April 2021

regulation of target genes [14].

An “universal” QS system shared by many Gram-negative and Gram-positive bacteria (AI-2) is mediated by the *luxS* gene, involved in the conversion of S-ribosylhomocysteine (SRH) in 4,5-dihidroxi-2,3-pentanodione (DPD), which is an unstable molecule that goes through random rearrangement and generates compounds with AI-2-like activity [15]. It has been reported more than 100 bacterial genomes harbour this gene [13], and its detection may suggest a metabolic potential for interspecific QS communication [16].

Some studies focused on the QS importance for LAB in bacteriocin production [17,18] and for characterization of several LAB isolates regarding AI-2 production [19], but there is no clear definition for the role of QS regulation in food fermentations.

To the best of our knowledge, no study has mapped the distribution of *luxS* gene in LAB populations or has described its possible role, based on the investigation of core and accessory genomes. This subject is very important to be investigated to provide insights on QS in LAB, since it potentially influences bacterial metabolism and expression of phenotypic traits [20].

In this study we pursuit four aims: (i) to determine the Whole-Genome Sequencing (WGS) of three selected LAB strains originated from a Brazilian cocoa fermentation, belonging to *Lp. plantarum, Lm. fermentum* and *Pediococcus acidilactici* species; (ii) to investigate the presence and location of *luxS* genes, in comparison with other relevant genome sequences previously deposited in databases; (iii) to compare the differential functional potential of cocoa adapted-strains; and (iv) to select the core *luxS* homologs to design clade and species-specific primers as a tool for further studies to monitor *luxS* activity during fermentations.

2. Material and methods

2.1. Bacterial strains

The strains *Lp. plantarum* Lb2, *Lm. fermentum* Lb1, *P. acidilactici* P1, previously isolated from a Brazilian spontaneous cocoa fermentation [21,22], were selected from our culture collection to be analysed by WGS. These strains were cultured in de Man, Rogosa & Sharpe (MRS) broth (Sigma-Aldrich/Merck, Darmstadt, Germany) at 30 °C for up to 24 h, streaked on MRS agar plates (Sigma-Aldrich/Merck) and incubated at 30 °C for up to 48 h, to confirm purity. Finally, a single colony was selected and inoculated in MRS broth to be used for genomic DNA extraction.

2.2. Genomic DNA extraction for PCR-based methods

Cultures were centrifuged at 14,000 rpm at 4 °C for 5 min. The pellet was saved and pre-treated for disruption the bacterial cell wall with 300 μl of lysozyme solution in Tris-EDTA buffer (TE) (10 mg/ml, Sigma-Aldrich/Merck), incubated at 37 °C for 60 min. The suspension was centrifuged at 14,000 rpm at 4 °C to remove cellular debris and the bacterial genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit reference A1125 (Promega, Madison, WI, USA) following the manufacturer's guidelines. DNA extracts were quality-assessed and quantified using Nanodrop® (Thermo Fisher Scientific, USA). The *Lactiplantibacillus plantarum* presented A_{260}/A_{280} and A_{260}/A_{230} ratios of 1.96 and 1.67, respectively; for *Limosilactobacillus fermentum* the ratio A_{260}/A_{280} was 2.00 and the ratio A_{260}/A_{230} was 1.92. Finally, *Pediococcus acidilactici* presented ratios of A_{260}/A_{280} equal to 1.96, and A_{260}/A_{230} equal to 1.52. The quantified DNA of each species was diluted at 20 ng/μl for PCR reactions.

2.3. Genomic DNA extraction for WGS

DNA was extracted as described above, except for the replacement of TE buffer by sterilized purified water to prepare the lysozyme solution, in order to get rid of EDTA and avoid potential interference with sequencing. Rehydration of genomic DNA was done according to the guidelines of the Wizard® Genomic DNA Purification Kit reference A1125 (Promega), but sterilized Dnase-free water was used, instead of rehydration buffer solution. The DNA extracts were quantified using the Qubit™ dsDNA BR assay Kit (Invitrogen/Thermo Fisher Scientific) following manufacturer's guidelines and sent to the IGA Technology Services facility (Udine, Italy), which processed the genomic DNA for WGS sequencing following Illumina® (Illumina®, USA) guidelines to yield a 300 bp paired-end (PE) library with, at least, 100× of sequencing coverage (approximately two million of reads per sample) on a Miseq® (Illumina®) platform.

2.4. De novo genome assembly and annotation

The reads were quality assessed through Trimommatic tool [23] to remove adapters and low-quality sequences (Phred score lower than 33). The remaining PE reads were *de novo* assembled into contigs and scaffolds using the SPADES tool [24] choosing the default parameters provided by the pipeline. The assembled scaffolds were annotated on Prokka pipeline [25], using Prodigal [26] as *ab initio* gene prediction tool, whereas ncRNAs, rRNA and tRNA coding DNA sequences (CDS) were predicted using Infernal + Rfam, Barrnap and Aragorn, respectively available at the Prokka pipeline.

2.5. Quality assessment and genomes completeness

The genome assemblies were quality evaluated with Quast [27] selecting a minimum threshold for scaffold length of 200 bp and default parameters. The genome completeness was measured using BUSCO tool [28] by means of presence of universal single-copy orthologous genes derived from the “lactobacillales_odb9” lineage-specific database detected in the newly assembled cocoa genomes. In the next step, the genome assemblies were classified as Complete (C) when their lengths were in the range of the two standard deviations of the BUSCO group mean length; Duplicated (D) when there was more than one copy of the orthologue gene; Fragmented (F) when the gene was partially recovered, and Missing (M) whenever no single-copy gene was found in the dataset. These classifications were based on the “number of genes used” to depict the completeness resolution and to describe the confidence of the metrics. These preliminary analyses were conducted on the Galaxy platform [29]. The assembled scaffolds were deposited on the NCBI GenBank database with accession numbers: *Lp. plantarum* Lb2 JACBJO000000000, *Lm. fermentum* Lb1 JACBJN000000000, and *P. acidilactici* P1 JACBJP000000000.

2.6. Selection of LAB strains from databases for genomic comparison with cocoa-derived LAB strains

The experimental design of comparative genomic analysis for each taxon was based on the pipeline described by Wittouck et al. [30] to select a representative number of strains among those already deposited in the GenBank database, with high-quality assemblies, to guarantee the assessment of meaningful biological data. In summary, all genome assemblies of ‘*Lactobacillus plantarum*’, ‘*Lactobacillus fermentum*’ (basonyms) and *P. acidilactici* taxa were downloaded from NCBI on February 1st of 2020, using the *in-house* scripts provided by the pipeline. Additionally, in an attempt to enlarge the dataset, we downloaded from the database several genomes that were unclassified at species level (annotated only as *Lactobacillus* sp. and *Pediococcus* sp.). These genomes were blasted using the blastn option with the blast+ software (version 2.2.9) [31], using as reference a customized database created by downloading all *16S rRNA* gene sequences from the RDP database v. 11.5 [32], to contain all the sequences longer than 1200 nucleotides derived from bacterial cultures. For all assemblies, quality and GC content were evaluated with QUAST tool [27], selecting only the assemblies with N75 values equal higher than 10,000 bp, 9000 bp and 10,000 bp for *Lp. plantarum*, *Lm. fermentum* and *P. acidilactici* strains, respectively. Moreover, the assemblies presented less than 500 undetermined bases per 100,000 bp. The selected genomes were re-annotated with Prokka [25] using the default standards, and to confirm the phylogenetic relatedness of the strains, the Average Nucleotide Identity (ANI) was calculated using the pyani package [33], setting the ANI threshold value ≥95% for species level classification.

With regard to *Lp. plantarum* dataset downloaded from NCBI database, it included 511 genomes of '*Lactobacillus plantarum*' that were already known, plus 28 novel *Lp. plantarum* genomes that had been previously classified only at genus level. After quality filtering and GC normalization, 415 genomes of *Lp. plantarum* were obtained in total, with 406 assemblies from known *Lp. plantarum* strains, plus nine assemblies previously regarded solely as *Lactobacillus* sp. Out of the 415 assemblies, six of them were removed from the dataset because they represented identical genomes (GCA_000474695.1, GCA_002370965.1, GCA_002631775.1, GCA_003258615.1, GCA_004102845.1 and GCA_001704595.1).

The ANI values of *Lp. plantarum* assemblies were evaluated to confirm the assignment at species level (ANI value ≥95%). Considering this parameter, other assemblies of *Lp. plantarum* were also removed from the dataset (GCA_005404945, GCA_005405125, GCA_005405105, GCA_009720675 and the assembly GCA_005405065). Overall, only four *Lp. plantarum* strains primarily assigned as unclassified *Lactobacillus* at NCBI were included in this study, summing up 404 *Lp. plantarum* strains to be evaluated against the novel strain *Lp. plantarum* Lb2.

For the *Lm. fermentum* group, a total of 71 '*Lactobacillus fermentum*' genomes were download and one *Lactobacillus* sp. was father identified by Blasting against the customized *16S rRNA* gene database. All the assemblies were quality-filtered and evaluate for the ANI values (considering the threshold ≥95%). In this way, there were 63 genomes (62 *Lm. fermentum* plus one *Lactobacillus* sp.) and the novel strain *Lm. fermentum* Lb1.

Concerning *P. acidilactici* group, 150 assemblies were downloaded from the NCBI database, and no genome from *Pediococcus* sp. was retrieved. After the quality-filtering step, 144 assemblies of *P. acidilactici* were left, and they were evaluated by ANI metric (setting the threshold ≥95%), revealing the strain GCA_009809575.1 should also be removed from the dataset due to erroneous identification. Thus, 143 quality-filtered assemblies were considered for pangenome analysis in comparison with *P. acidilactici* strain P1. All genome assemblies used in this study are listed in Table S1.

2.7. Pangenome analysis of cocoa strains and LAB genomes derived from several food matrices

The analyses of LAB pangenomes were performed with Roary pipeline [34] on the Galaxy platform [29] using the default parameter of 95% of minimum percentage of identity to cluster homologous genes based on similarity of nucleotide sequences. The core and accessory genomes were determined using the MAFFT aligner [35] to blast all-*vs*-all genes, clustering them in a set of homologous groups, setting 99% as the percentage of strains that should harbour a given cluster to be assigned to the core genome. From the "presence_absence.Rtab" output, the *luxS* distribution among the strains was determined as presence/absence (0/1) and the gene clusters were classified as core or accessory.

Besides, to compare the differences between the functional potential in the core of each LAB taxa against the set of niche-specific genes of each cocoa strain, an additional pangenome analysis was performed, using the Spine and Agent [36] pipelines, which identify respectively the homologous genes among the strains (core genome) and the accessory genes. The core and accessory genes were annotated and represented in COG (Cluster of Orthologues Genes) categories using the eggNOG-mapper v.2 [37] against the eggNOG v. 5.0 database [38]. For the determination of carbohydrases gene families, the script run_dbcan.py version 2.0 was used to scan the core genomes and the cocoa accessory genomes, against the Carbohydrates-Active Enzymes (CAZY) database, with the default options available at the pipeline. In the other hand, to search for putative protein families related to the metabolism of amino acids, the "Family Protein Sequences" was downloaded from the MEROPS database [39] and a similarity search was performed using blastx searches (e-value 1×10^{-5}).

2.8. Determination of horizontally-transferred genes (HGTs) in cocoa strains

Considering that in pangenomes with open structures, it is very likely the extensive occurrence of HGT-based events, the Alien Hunter tool [40] was used to investigate putative HGTs events in the genomes of newly sequenced cocoa LAB, as well as in the LAB genomes from the public database. The "Alien Hunter" performs independent scans of the genomes to detect regions of "abnormal composition", taking into account the whole genome. These "alien" regions are generally characterized by differential G + C content, frequency of dinucleotides and also codon bias. After the identification of these regions, the algorithm gives as output several scores for each predicted atypical region which may be a result of HGT [41]. By the use of the "extractseq" tool available with the EMBOSS package (version 6.6.0) [42], the sequences belonging to each putative HGT-related region were extracted from each cocoa LAB genome and the fragmented gene prediction was performed using Prodigal [26]. Finally, to confirm if the putative HGT sequences contained *luxS* genes, the predicted genes were annotated and represented in COG (Cluster of Orthologues Genes) categories using the eggNOG-mapper v.2 [37] against the eggNOG v. 5.0 database [38].

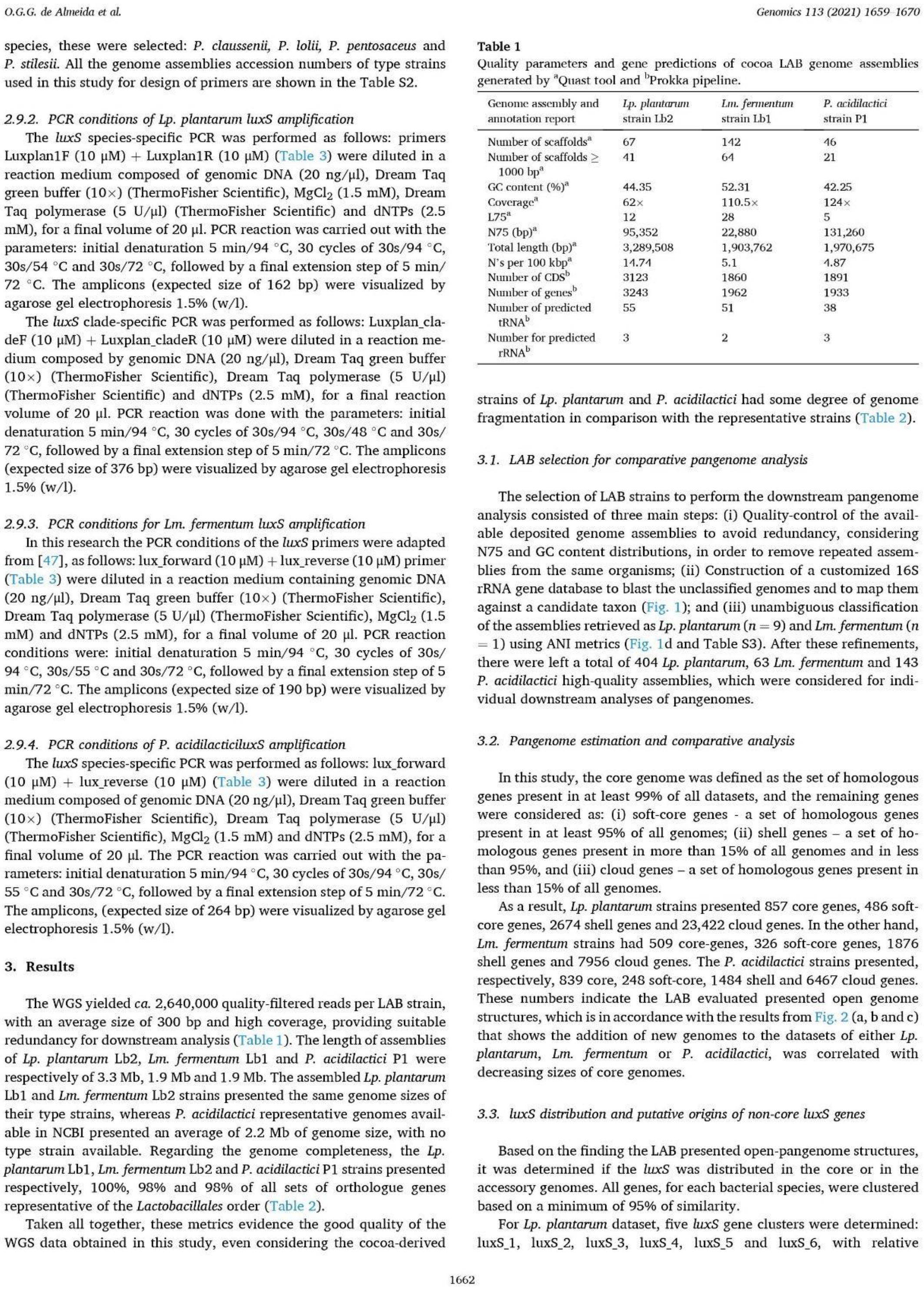
2.9. Putative origins of non-core luxS gene clusters

Considering that each gene was clustered based on a threshold of a minimum 95% of similarity, for the non-core luxS_2 clusters of *Lp. plantarum* and *Lm. fermentum*, one representative sequence was chosen and blasted against the entire non-redundant NCBI database and the matched homologous sequences were retrieved, joined in a single fasta file and analysed in MEGA-X version 10.0.5 [43], to build a neighbor-joining (NJ) tree with the parameters of 1000 bootstraps and partial deletion (threshold of 95%) to unravel the possible HGT origin of these *luxS* sequences.

2.9.1. Primer design to assess the core luxS genes

To design genus-specific and species-specific primers for *luxS* gene, the genomes of all type strains publicly available at the time of this study, were downloaded from the NCBI assembly database and their *luxS* gene sequences were retrieved using Unix command line. For the *P. acidilactici* group, the List of Prokaryotic names with Standing in Nomenclature database [44,45] was consulted to find the type strains for *Pediococcus* spp., in order to select *luxS* sequences for designing species-specific primers for the *luxS*. By the use of Unix command line, all *luxS* genes were recovered from the type strains, they were joined in a unique FASTA file and aligned with CLUSTAL OMEGA [46]. Based on the Multiple Sequence Alignments (MSLA), the types *luxS* sequences were selected to design degenerated primers. For that, the identity percentages among the sequences were determined and the presenting over 90% similarity were selected, considering that they would potentially allow for the differentiation of closely-related species.

For *Lp. plantarum* and genus-related species, these were selected: *Lp. daoliensis, Lp. daowaiensis, Lp. dongliensis, Lp. fabifermentans, Lp. herbarum, Lp. pentosus, Lp. plantarum* subsp. *argentoratensis, Lp. xiangfangensis, Lp. mudanjiangensis, Lp. songbeiensis, Lp. nangangensis, Lp. pingfangensis, Lp. modestisalitolerans, Lp. plajomi* and *Lp. paraplantarum*. Among *Lm. fermentum* and genus-related species, these were selected: *Lm. panis, Lm. reuteri, Lm. equigenerosi, Lm. gastricus, Lm. ingluviei, Lm. mucosae, Lm. pontis* and *Lm. gorillae*. Finally, among *P. acidilactici* and genus-related



species, these were selected: *P. claussenii, P. lolii, P. pentosaceus* and *P. stilesii*. All the genome assemblies accession numbers of type strains used in this study for design of primers are shown in the Table S2.

2.9.2. PCR conditions of *Lp. plantarum luxS* amplification

The *luxS* species-specific PCR was performed as follows: primers Luxplan1F (10 μM) + Luxplan1R (10 μM) (Table 3) were diluted in a reaction medium composed of genomic DNA (20 ng/μl), Dream Taq green buffer (10×) (ThermoFisher Scientific), $MgCl_2$ (1.5 mM), Dream Taq polymerase (5 U/μl) (ThermoFisher Scientific) and dNTPs (2.5 mM), for a final volume of 20 μl. PCR reaction was carried out with the parameters: initial denaturation 5 min/94 °C, 30 cycles of 30s/94 °C, 30s/54 °C and 30s/72 °C, followed by a final extension step of 5 min/ 72 °C. The amplicons (expected size of 162 bp) were visualized by agarose gel electrophoresis 1.5% (w/l).

The *luxS* clade-specific PCR was performed as follows: Luxplan_cladeF (10 μM) + Luxplan_cladeR (10 μM) were diluted in a reaction medium composed by genomic DNA (20 ng/μl), Dream Taq green buffer (10×) (ThermoFisher Scientific), Dream Taq polymerase (5 U/μl) (ThermoFisher Scientific) and dNTPs (2.5 mM), for a final reaction volume of 20 μl. PCR reaction was done with the parameters: initial denaturation 5 min/94 °C, 30 cycles of 30s/94 °C, 30s/48 °C and 30s/ 72 °C, followed by a final extension step of 5 min/72 °C. The amplicons (expected size of 376 bp) were visualized by agarose gel electrophoresis 1.5% (w/l).

2.9.3. PCR conditions for *Lm. fermentum luxS* amplification

In this research the PCR conditions of the *luxS* primers were adapted from [47], as follows: lux_forward (10 μM) + lux_reverse (10 μM) primer (Table 3) were diluted in a reaction medium containing genomic DNA (20 ng/μl), Dream Taq green buffer (10×) (ThermoFisher Scientific), Dream Taq polymerase (5 U/μl) (ThermoFisher Scientific), $MgCl_2$ (1.5 mM) and dNTPs (2.5 mM), for a final volume of 20 μl. PCR reaction conditions were: initial denaturation 5 min/94 °C, 30 cycles of 30s/ 94 °C, 30s/55 °C and 30s/72 °C, followed by a final extension step of 5 min/72 °C. The amplicons (expected size of 190 bp) were visualized by agarose gel electrophoresis 1.5% (w/l).

2.9.4. PCR conditions of *P. acidilacticiluxS* amplification

The *luxS* species-specific PCR was performed as follows: lux_forward (10 μM) + lux_reverse (10 μM) (Table 3) were diluted in a reaction medium composed of genomic DNA (20 ng/μl), Dream Taq green buffer (10×) (ThermoFisher Scientific), Dream Taq polymerase (5 U/μl) (ThermoFisher Scientific), $MgCl_2$ (1.5 mM) and dNTPs (2.5 mM), for a final volume of 20 μl. The PCR reaction was carried out with the parameters: initial denaturation 5 min/94 °C, 30 cycles of 30s/94 °C, 30s/ 55 °C and 30s/72 °C, followed by a final extension step of 5 min/72 °C. The amplicons, (expected size of 264 bp) were visualized by agarose gel electrophoresis 1.5% (w/l).

3. Results

The WGS yielded *ca.* 2,640,000 quality-filtered reads per LAB strain, with an average size of 300 bp and high coverage, providing suitable redundancy for downstream analysis (Table 1). The length of assemblies of *Lp. plantarum* Lb2, *Lm. fermentum* Lb1 and *P. acidilactici* P1 were respectively of 3.3 Mb, 1.9 Mb and 1.9 Mb. The assembled *Lp. plantarum* Lb1 and *Lm. fermentum* Lb2 strains presented the same genome sizes of their type strains, whereas *P. acidilactici* representative genomes available in NCBI presented an average of 2.2 Mb of genome size, with no type strain available. Regarding the genome completeness, the *Lp. plantarum* Lb1, *Lm. fermentum* Lb2 and *P. acidilactici* P1 strains presented respectively, 100%, 98% and 98% of all sets of orthologue genes representative of the *Lactobacillales* order (Table 2).

Taken all together, these metrics evidence the good quality of the WGS data obtained in this study, even considering the cocoa-derived strains of *Lp. plantarum* and *P. acidilactici* had some degree of genome fragmentation in comparison with the representative strains (Table 2).

Table 1
Quality parameters and gene predictions of cocoa LAB genome assemblies generated by [a]Quast tool and [b]Prokka pipeline.

Genome assembly and annotation report	*Lp. plantarum* strain Lb2	*Lm. fermentum* strain Lb1	*P. acidilactici* strain P1
Number of scaffolds[a]	67	142	46
Number of scaffolds ≥ 1000 bp[a]	41	64	21
GC content (%)[a]	44.35	52.31	42.25
Coverage[a]	62×	110.5×	124×
L75[a]	12	28	5
N75 (bp)[a]	95,352	22,880	131,260
Total length (bp)[a]	3,289,508	1,903,762	1,970,675
N's per 100 kbp[a]	14.74	5.1	4.87
Number of CDS[b]	3123	1860	1891
Number of genes[b]	3243	1962	1933
Number of predicted tRNA[b]	55	51	38
Number for predicted rRNA[b]	3	2	3

3.1. LAB selection for comparative pangenome analysis

The selection of LAB strains to perform the downstream pangenome analysis consisted of three main steps: (i) Quality-control of the available deposited genome assemblies to avoid redundancy, considering N75 and GC content distributions, in order to remove repeated assemblies from the same organisms; (ii) Construction of a customized 16S rRNA gene database to blast the unclassified genomes and to map them against a candidate taxon (Fig. 1); and (iii) unambiguous classification of the assemblies retrieved as *Lp. plantarum* ($n = 9$) and *Lm. fermentum* ($n = 1$) using ANI metrics (Fig. 1d and Table S3). After these refinements, there were left a total of 404 *Lp. plantarum*, 63 *Lm. fermentum* and 143 *P. acidilactici* high-quality assemblies, which were considered for individual downstream analyses of pangenomes.

3.2. Pangenome estimation and comparative analysis

In this study, the core genome was defined as the set of homologous genes present in at least 99% of all datasets, and the remaining genes were considered as: (i) soft-core genes - a set of homologous genes present in at least 95% of all genomes; (ii) shell genes – a set of homologous genes present in more than 15% of all genomes and in less than 95%, and (iii) cloud genes – a set of homologous genes present in less than 15% of all genomes.

As a result, *Lp. plantarum* strains presented 857 core genes, 486 soft-core genes, 2674 shell genes and 23,422 cloud genes. In the other hand, *Lm. fermentum* strains had 509 core-genes, 326 soft-core genes, 1876 shell genes and 7956 cloud genes. The *P. acidilactici* strains presented, respectively, 839 core, 248 soft-core, 1484 shell and 6467 cloud genes. These numbers indicate the LAB evaluated presented open genome structures, which is in accordance with the results from Fig. 2 (a, b and c) that shows the addition of new genomes to the datasets of either *Lp. plantarum*, *Lm. fermentum* or *P. acidilactici*, was correlated with decreasing sizes of core genomes.

3.3. *luxS* distribution and putative origins of non-core *luxS* genes

Based on the finding the LAB presented open-pangenome structures, it was determined if the *luxS* was distributed in the core or in the accessory genomes. All genes, for each bacterial species, were clustered based on a minimum of 95% of similarity.

For *Lp. plantarum* dataset, five *luxS* gene clusters were determined: luxS_1, luxS_2, luxS_3, luxS_4, luxS_5 and luxS_6, with relative

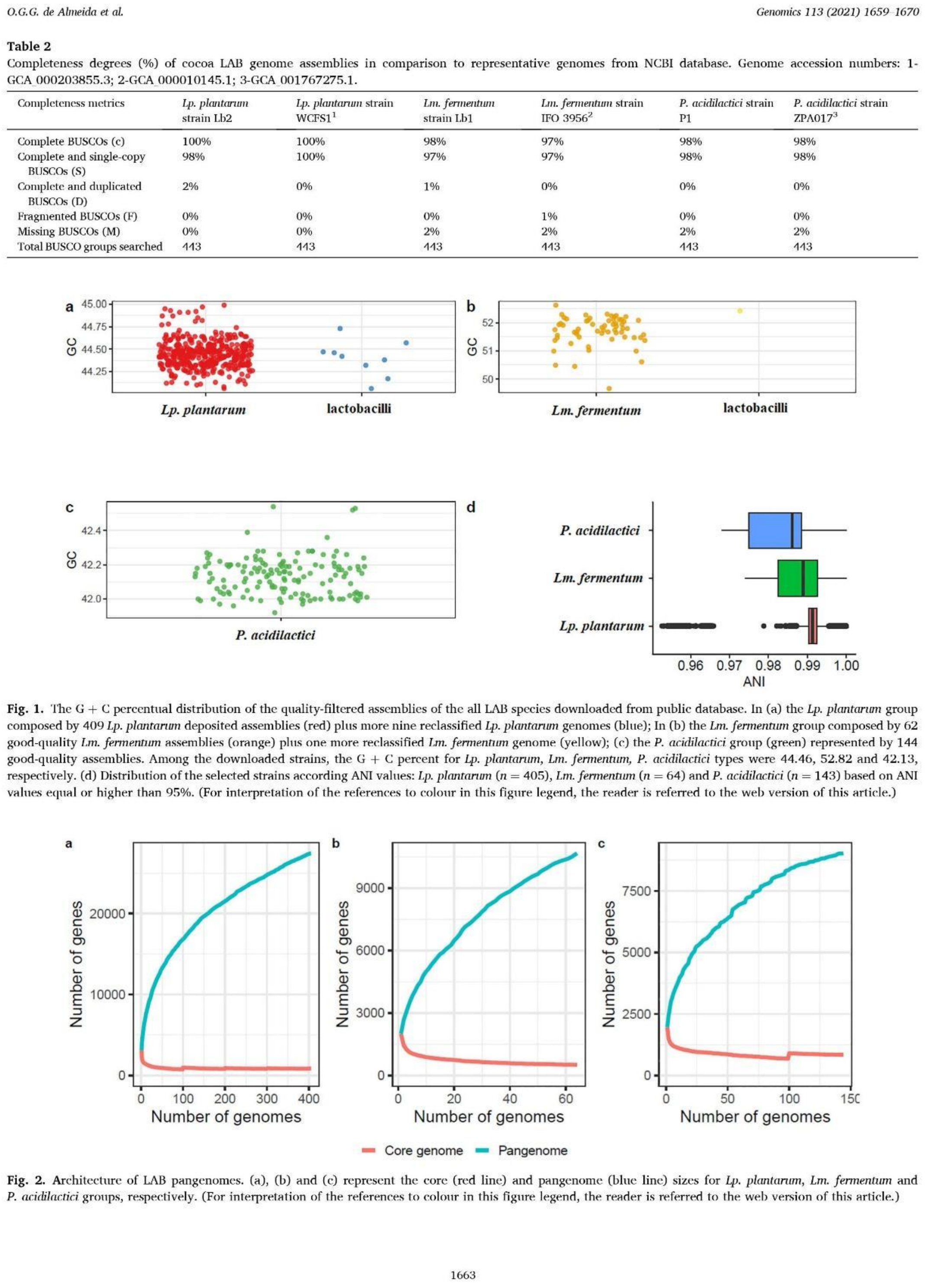


Table 2
Completeness degrees (%) of cocoa LAB genome assemblies in comparison to representative genomes from NCBI database. Genome accession numbers: 1-GCA_000203855.3; 2-GCA_000010145.1; 3-GCA_001767275.1.

Completeness metrics	*Lp. plantarum* strain Lb2	*Lp. plantarum* strain WCFS1[1]	*Lm. fermentum* strain Lb1	*Lm. fermentum* strain IFO 3956[2]	*P. acidilactici* strain P1	*P. acidilactici* strain ZPA017[3]
Complete BUSCOs (c)	100%	100%	98%	97%	98%	98%
Complete and single-copy BUSCOs (S)	98%	100%	97%	97%	98%	98%
Complete and duplicated BUSCOs (D)	2%	0%	1%	0%	0%	0%
Fragmented BUSCOs (F)	0%	0%	0%	1%	0%	0%
Missing BUSCOs (M)	0%	0%	2%	2%	2%	2%
Total BUSCO groups searched	443	443	443	443	443	443

Fig. 1. The G + C percentual distribution of the quality-filtered assemblies of the all LAB species downloaded from public database. In (a) the *Lp. plantarum* group composed by 409 *Lp. plantarum* deposited assemblies (red) plus more nine reclassified *Lp. plantarum* genomes (blue); In (b) the *Lm. fermentum* group composed by 62 good-quality *Lm. fermentum* assemblies (orange) plus one more reclassified *Lm. fermentum* genome (yellow); (c) the *P. acidilactici* group (green) represented by 144 good-quality assemblies. Among the downloaded strains, the G + C percent for *Lp. plantarum, Lm. fermentum, P. acidilactici* types were 44.46, 52.82 and 42.13, respectively. (d) Distribution of the selected strains according ANI values: *Lp. plantarum* (n = 405), *Lm. fermentum* (n = 64) and *P. acidilactici* (n = 143) based on ANI values equal or higher than 95%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. Architecture of LAB pangenomes. (a), (b) and (c) represent the core (red line) and pangenome (blue line) sizes for *Lp. plantarum, Lm. fermentum* and *P. acidilactici* groups, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

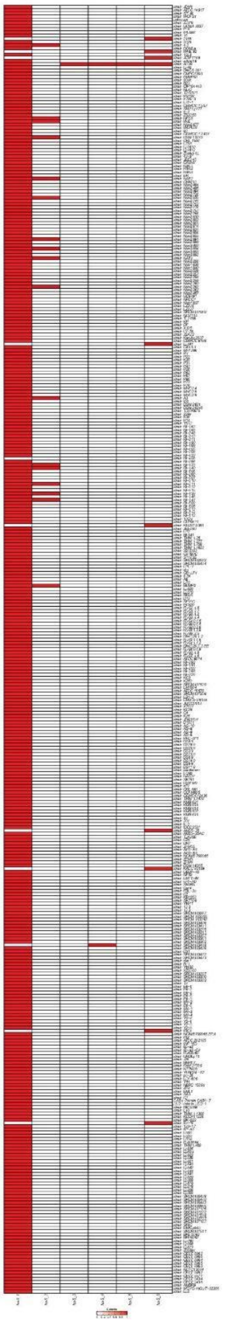


Fig. 3. Presence/absence heatmap of *luxS* genes in *Lp. plantarum* strains and its clusters. The "unknown" strain means a species without strain name deposited in the database.

abundances respectively of 97.28%, 4.20%, 0.25%, 0.49%, 0.25% and 2.5% (Fig. 3).

As shown for *Lm. fermentum* species in Fig. 4, only two homologous clusters were detected, with luxS_1 located in the core genome, and luxS_2 in the cloud genome (relative abundance of 9.38%). On the other hand, *P. acidilactici* had only one luxScluster, with genes belonging to the core genome for 100% of strains.

Interestingly, luxS_2 genes of *Lp. plantarum* and *Lm. fermentum* strains were clustered in the same branch, with plasmidial *luxS* sequences depicted by the Neighbour-Joining (NJ) gene phylogenies (Fig. 5a and b, respectively). Besides, it is likely only the *luxS* of *Lm. fermentum* from the luxS_2 cluster had been horizontally transferred by transduction. This is hypothesized due to the observation that the *luxS* gene was flanked by transposase IS30 sequences, by comparison with the genomes of the representative strains *Lp. plantarum* 4–3 and *Lm. fermentum* YL-11 (Fig. 5c).

Finally, taking into consideration the average global sequence alignment among the sequences of luxS_3, luxS_4, luxS_5 and luxS_6 gene clusters in *Lp. plantarum* was 99.63% (Fig. 5d), it is possible to affirm they represent a unique array of the same gene, which were individually clustered because they are paralogs [34], implicating they do not belong to accessory genome of *Lp. plantarum*.

3.4. Differences among the functional repertoires of cocoa LAB strains and putative origins of some orthologues

Based on the hypothesis that LAB genomes evolved by decreasing in size and fitting to metabolize specific sources of carbohydrates [3], it was of great interest to evaluate the accessory genomes of the cocoa-related LAB strains. Our findings revealed an enrichement in accessory genes associated with the metabolism of amino acids and carbohydrates, in comparison with the core genome (Fig. S1). Moreover, there were at least 50% of unknown COG genes in the LAB genomes, with the highest variation observed for the accessory genomes, except for *P. acidilactici* (Fig.S1).

These findings are of key importance to understand the microbiome of cocoa spontaneous fermentation, which is known to present large functional potential for the metabolism of amino acids and carbohydrates [11]. All gene families from these categories were studied with regard to their abundances and their proportions in the core as well as in the accessory genomes of LAB cocoa strains (Fig. S2).

The gene families associated with carbohydrate metabolism were classified in Glycoside Hydrolases (GH), Glycosyl Transferases (GT), Carbohydrate Esterases (CE) and Auxiliary Activities (AA), according to the CAZY database. On the other hand, the gene families related to amino acid metabolism were classified with basis on their catalytic activity, as follows: asparagine (A), cysteine (C), isoleucine (I), metalloproteins (M), serine (S), threonine (T) and unknown (U).

The Fig. S2 shows all annotated carbohydrases (left side, items a, c and e) and peptidases (right side, items b, d and f) gene families, for each cocoa LAB strain. Regarding *Lp. plantarum*, the core carbohydrases' repertoire was mainly represented by the genes associated with glycogen-binding proteins; glycosyl transferases of the families 2, 5, 26, 28, 35 and 51 (GT2, GT5, GT26, GT28, GT35 and GT51); glycosyl hydrolases of the families 1, 2, 13, 25, 65, 70, 73 (GH1, GH2, GH13, GH25, GH65, GH70 and GH73); carbohydrate esterases of the families 1 and 9 (CE1 and CE9); vanillyl-alcohol oxydase; starch-binding genes, and copper-dependent lytic polysaccharide monooxygenases (LPMOs) (Fig. S2a). In opposite, the majority of accessory genomes presented the highest abundances for glycosyl hydrolases of the family 13 (variant 31-GH13_31) and for glycosyl transferases of the families 2 and 4 (GT2 and GT4, respectively), with low diversity profile for glycosil hydrolases, glycosyl transferases and carbohydrate esterases (GHs, GTs and CEs).

As it regards to *Lp. plantarum* peptidases, it was observed the enrichment of metalloproteinases, and selenoproteinases, both in the core genome of *Lp. plantarum* populations and in the strain-specific

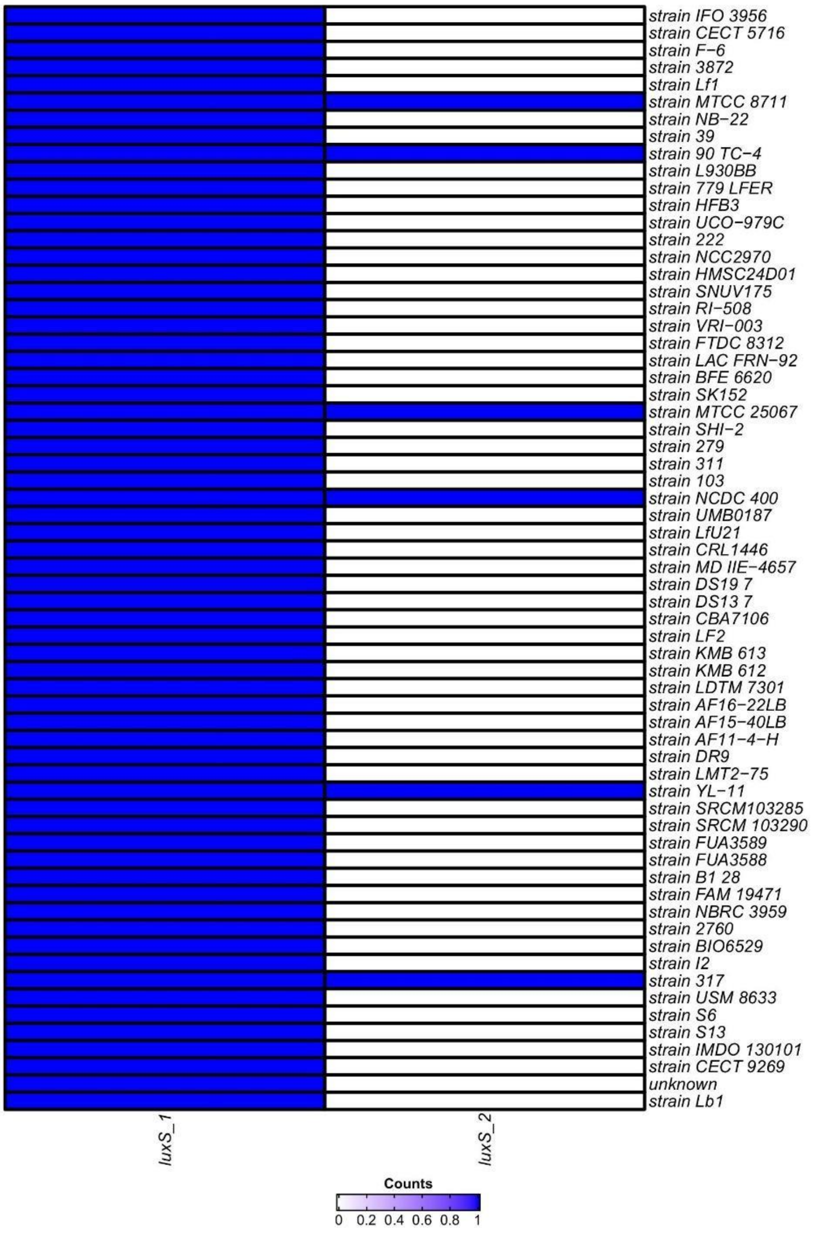


Fig. 4. Presence/absence heatmap of *luxS* genes in *Lm. fermentum* strains and its clusters. The "unknown" strain means a species without strain name deposited in the database.

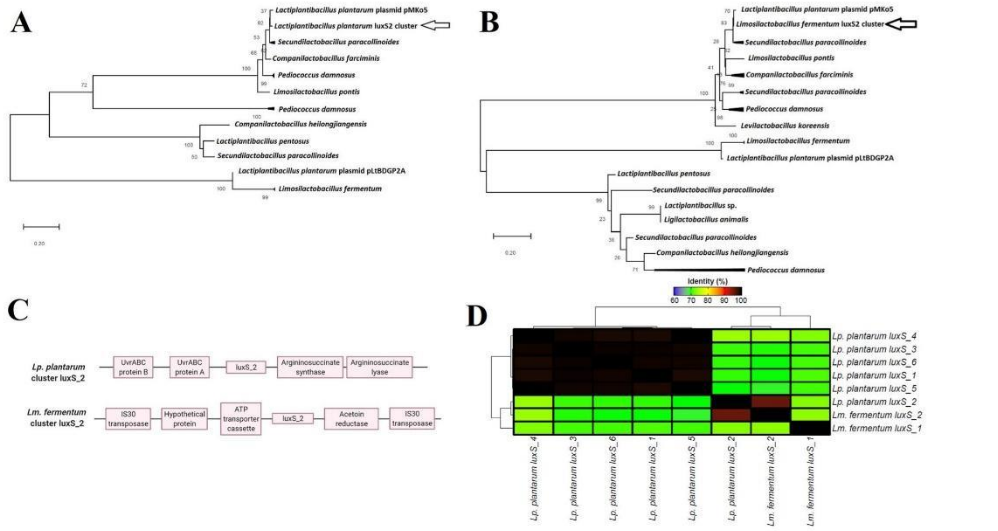


Fig. 5. *luxS* genes clusters determined by ROARY pipeline. (a) and (b) Neighbour-Joining phylogenetic tree of *luxS* genes of the clusters luxS_2 in *Lm. fermentum* and *Lp. plantarum*, respectively, determined with base on1,000 bootstrap replicates; (c) Organisation of *luxS* genes in *Lp. plantarum* strain 4-3 and *Lm. fermentum* strain YL-11; and (d) Global blast similarities of all *luxS* genes for each *Lp. plantarum* and *Lm. fermentum* clusters showing, with the exception of the luxS_2 cluster, the remaining *Lp. plantarum* gene clusters are composed by the same genes.

genes of cocoa-related LAB (Fig.S2b). The core genome presented the highest levels of serinopeptidases from the families 09c, 09×, 11, 12 and 33 (S09c, S09×, S11, S12 and S33), whereas the accessory genomes of all cocoa-related *Lp. plantarum* strains were enriched in metalloproteinases of the family 79 (S79), isoleucinase of the family 51 (I51), serinase of the families 09× (S09x) and 12 (S12), besides cysteinase genes of the families 82a (C82a), 26 (C26) and 44 (C44).

The *Lm. fermentum* group presented a core profile with marked abundance of carbohydrate esterases of family 10 (CE10) and glucosyl hydrolases of the family 73 (GH73), whereas in the accessory genomes, these genes were less abundant (Fig. S2c). However, there was an overall increase of CE10 genes in the cocoa-related *Lm. fermentum* strains, taking into consideration the CE10 core genes were complemented by accessory genes of the same family. The core genome also presented glycosyl hidrolases (GH) of the families 2 and 32, besides glycosyl transferases (GT) of family 51. It is also worth to note the enrichment of genes-coding for chitin-binding proteins, plus GH1, GT2 and GT14 carbohydrases in the accessory genomes of all *Lm. fermentum* cocoa-related strains (Fig. S2c).

Lm. fermentum strains also showed a rich repertoire of metalloproteinases and serinopeptidases genes (Fig. S2d). Among the metalloproteinases and serinopeptidases, the core was exclusively composed by the following gene families: M10a, M16b, M20, M24b, M41, M48a, M50b, S12, S15 and S16. Regarding the cocoa-related LAB genomes, it was observed a differential repertoire of serinopeptidases represented by the families of S01c and S33 (Fig. S2d).

Due to the absence of previous reports on *P. acidilactici* strains isolated from cocoa beans' fermentation, it was possible only to compare the novel strain *P. acidilactici* P1, with the core genome for the species (Fig. S2e).In the accessory genome of *P. acidilactici* P1, there was an enrichment of GH1 and GT2 families of carbohydrases in comparison with the core (Fig. S2e). Moreover, the accessory genome of the *P. acidilactici* P1 presented a singular repertoire of glycosyl hydrolases and glycosyl transferases, especially GH3, GH18, GH38, GH43, GH67, GH123, GH126, and GT26. With regard to peptidases, the cocoa-related strain *P. acidilactici* P1 presented a unique pool of gene families, as shown in Fig. S2f, namely cysteinase 56, metalloproteinase 41 and serinopeptidase 49c.

It is generally accepted that accessory genomes are more likely to receive foreign genes by HGT [50]. For that reason, we also predicted putative regions with HGT signatures in all cocoa-related LAB genomes (Fig. S3) and classified them according to COG categories (Fig. S4). The majority of all putative HGT annotated functions were unknown, but it was possible to observe high counts of genes associated to the metabolism of carbohydrates and amino acids, besides those associated with DNA transcription, replication and repair (Fig. S4). These functions were prevalent in *Lp. plantarum* 80, *Lp. plantarum* Lb2 and *P. acidilactici* P1, but the two *Lm. fermentum* strains presented a differential composition of COGs. *Lm. fermentum* Lb2 was differentiated from *Lm. fermentum* 222 due to the possible acquisition of functions related to post-translational modification, protein turnover, chaperones, lipid transport and metabolism (Fig. S4). Overall, all LAB strains presented more abundant putative HGT signatures in genes related to carbohydrate metabolism.

Furthermore, in this study, the number of putative HGT-genes classified in COGs related to the category of amino acid metabolism (Fig. S5) were annotated in the highest proportions for *Lp. plantarum* Lb2, *Lp. plantarum* 80, *Lp. plantarum* RI-515, *Lp. plantarum* RI-505, *P. acidilactici* P1, *Lm. fermentum* 222, *Lm. fermentum* Lb1 and *Lp. plantarum* RI-513.

3.5. Primer design

From the multiple sequence alignments (MLSA) analyses (Fig. S5), the sequences presenting over 90% similarity were selected to guide primer designing based on the genomes of *Lp. daowaiensis, Lp. mudanjiangensis, Lp. dongliensis, Lp. songbeiensis, Lp. fabifermernts, Lp. nangangensis, Lp. pingfangensis, Lp. daoliensis, Lp. plajomi, Lp. modestisalitolerans, Lp. herbarum, Lp. xiangfangensis, Lp. pentosus, Lp. plantarum* and L. *paraplantarum* (Fig. S6). Two primers were designed for *Lp. plantarum* species, one species-specific (forward: Luxplan_1F + reverse: Luxplan_1R) and another one clade-specific (forward: Luxplan_cladeF +

Table 3
Primers used in this study and band sizes according to the primer's combination. (*) approximately determined by this study.

Primer designation	Sequence 5′3′	Amplicon size (bp)	Tm (°C)	Source
Luxplan1F	ACGTGATCGTATGGATGG	162	53.7	This study
Luxplan1R	CCCTTGTACGTCTTCCCA		56	
Luxplan_cladeF	AGAAAGTTTTACATTAGATCA	376	48.1	
Luxplan_cladeR	TCRTCYTTGTRTTCCCACA		53.4	
lux forward	AAACGGTGAGCAGTCAATCA	190*	55.3	[47]
lux reverse	AAGGTCCTAAGGGTGACAAGA		57.9	
LuxPedio_cladeF	ATGGCAAAAGTAGAAAGCTTTG	264	54.7	This study
LuxPedio_cladeR	CCCCAARCGATTAAATGGAAAC		57.5	

reverse Luxplan_cladeR) (Table 3). Both primers were designed considering the alignment of the *luxS* gene WP_003641031.1, once the other *luxS* gene (WP_022638718.1) present in the *Lp. plantarum* strain DSM 13273 did not have identity with the conserved *luxS* gene (WP_003641031.1), which shared homologues with the remaining species from the taxa of *Lp. plantarum.*

The species-specific primers presented good specificity to amplify the *luxS* regions of *Lp. plantarum* strain but the closely related species *Lp. paraplantarum* was also amplified (expected amplicons of 162 bp). In the other hand, there was no amplification of *luxS* genes from *Lm. fermentum* and *P. acidilactici* (Fig. S7). Regarding the clade specific primers, all *luxS* genes of *Lp. plantarum* species were amplified, rendering amplicons with 376 bp with specificity for bacterial species from *Lp. plantarum* clade (Fig. S8).

Regarding *Limosilactobacillus*-specific primers, they presented a high degree of non-specificity, annealing in variable genome regions (data not shown). For this reason, the species-specific primers drawn by Gu et al. [47] were employed (Table 3). This pair of primers was developed to be used in a real time PCR approach, but here they were tested following a conventional PCR amplification protocol. These primers proved to be efficient to amplify *luxS* in *Lm. fermentum* strains, yielding 190 bp amplicons (Fig. S9). Nevertheless, they amplified non-specific regions of *Lp. plantarum* genome and a specific region of 900 bp of *P. acidilactici.* Blast analyses confirmed the primers specificity for this unrecognized region in *P. acidilactici* genomes (data not shown). Even though, this combination of primers can be used for *luxS* detection in *Lm. fermentum* strains, since the target gene was successfully amplified (amplicons with 190 bp).

The primer design for *P. acidilactici* species was based on the sequences from *P. claussenii, P. acidilactici, P. lolii, P. pentosaceus* and *P. stilesii* (Fig. S10). The combination of forward LuxPedio_cladeF and reverse LuxPedio_cladeR clade-specific primers (Table 3) amplified only *P. acidilacticiluxS* genes rendering amplicons with 164 bp as expected, with specificity for *P. acidilactici* genomes (Fig. S11).

4. Discussion

To the best of our knowledge, as of February 2020, only 11 *Lp. plantarum* and two *Lm. fermentum* strains were publicly available fromcocoa fermentations, with no record about the deposit of *P. acidilactici* from that environment. This research depicts a cocoa-related *P. acidilactici* genome, increasing the number of LAB genomes with high-quality and completeness available in public repositories, which can contribute to unravel LAB niche-specific adaptations. The novel assemblies presented inter-species genome size variations similar to those of several lactobacilli, in the range of 2.9 to 3.5 Mb [49]. Considering lactobacilli generally tend to have reduced genome sizes likely due to niche-specific adaptation, it is not surprising the nomadic *Lp. plantarum* [48] presents a larger genome. Literature reports are in accordance with the present study, indicating the occurrence of extensive HGT events in *Lp. plantarum* strains contributed for enlarging their genomes [50]. Moreover, phylogenetic analysis of LAB core genes have revealed similar patterns of genome sizes (~1.9 Mb) for *Lm. fermentum* and *P. acidilactici* [51].

The newly sequenced genomes of LAB cocoa strains were added to a collection of LAB genomes from public databases to run an accurate pangenome analyses, in order to unveil the potential for functional adaptation to cocoa fermentation environment, by finding genes encoding for LuxS, carbohydrases and peptidases.

However, there is no consensus about thresholds for pangenome estimations, which depend on the number of genomes analysed, the ability of different species to incorporate exogenous DNA and the influence of environmental pressures molding the genome [55]. With all these factors associated, pangenomes may converge to an open structure (large size and low number of core genes) or to a closed structure (small size and high number of core genes), respectively reflecting the ability of the organism to add or not add new gene clusters [52,53,54].

The measured open pangenomes in this study were in agreement with previous reports in the literature regarding the high variability of the LAB core genomes, corroborated by the inverse correlation between the core size and the addition of new genomes to the dataset [50,51,56]. Due to the heterogeneous composition of LAB genomes, it was also investigated the localization of the LuxS-coding genes, whether in the core or in the accessory genomes (with inference of HGT events), and the putative enzymatic repertoire specific for cocoa-related bacterial strains.

4.1. Distribution of luxS genes in LAB and putative origins of non-core luxS gene clusters

Pangenome analyses determined the presence of *luxS* gene in the soft-core of *Lp. plantarum*, and in the core genomes of *Lm. fermentum* and *P. acidilactici* species, which possibly indicate the importance of this gene for LAB adaptation to diverse niches [13]. This was also corroborated by the presence of additional *luxS* genes in some accessory genomes of *Lp. plantarum* and *Lm. fermentum.*

The results from this study indicated that although the genes from the luxS_2 cluster were horizontally transferred from *Lp. plantarum* to some strains of *Lm. fermentum,* the remaining four clusters of *Lp. plantarum* (luxS_3, luxS_4, luxS_5 and luxS_6) were copies of the core *luxS* genes (cluster luxS_1), with a mean of 99.63% of global sequence similarity, in comparison with the core sequences. These sequences were split from the core sequences due to technical considerations of the Roary pipeline [34], in which the sequences of paralogues are separated according to synteny. This implies that if the novel gene copy is located in a different portion of genome, with a diverse constitution with regard to the majority of copies from the first cluster, it forms another cluster [34]. In this research, high copy numbers for *luxS* gene were found in *Lp. plantarum* genome.

It can be hypothesized the presence of multiple gene copies in a bacterial cell contributes for increased growth rates, in the case of genes encoding for proteins involved in the transport of nutrients. In fact, it has been shown *Lp. plantarum* can quickly uptake various substrates, which is a general capability of LAB that correlates with fast growth [59]. In the other hand, the ability for fast growth may contribute for maintaining multiple gene copies in bacterial cells, with basis on the binomium "replication" *versus* "cellular division" [57]. This theory considers that under a normal pace of bacterial growth, there is a coordinated replication that starts at a defined point of the DNA strand, and continues symmetrically until converging at the opposed ends of the chromosome, with cell division taking place in the D-period, immediately after DNA replication [58]. This process allows only for small variations in the multiplicity of a given gene (one or two copies), but to the contrary, if DNA replication occurs too fast, the processes of "replication" and "cell division" are not synchronized, leading to a multifork DNA replication, with the synthesis of multiple copies of the same gene [57].

The presence of multiple *luxS* gene copies may be relevant for bacterial metabolism because the LuxS protein is crucial for the metabolism of cysteine and methionine. These sulphur-containing amino acids are linked to the Activated Methyl Cycle that provides methyl groups to several intermediate reactions, ending up with the conversion of SRH to homocysteine, a reactional step catalysed by LuxS. This also generates the DPD, a precursor of AI-2 like QS molecules [60,61].

According to the literature, *luxS* gene is crucial for interspecies bacterial communication, and a variability of homologous genes have already been described in over 530 prokaryotic genomes [62]. In food fermentation, the ecological role for QS is still under investigation, with reports on its influence on bacteriocin production, stress resistance and biofilm formation [63,64]. In a study using a proteomic approach and deletion-mutants [65], it was observed there was a *luxS*-dependent upregulation of the metabolism of carbohydrates, amino acids and fatty acids, accompanied by the increased expression the of the two-components system (TCS). As it regards to bacteriocin biosynthesis, which is an energy-consuming process, the QS-regulated TCS system could be important to mantain the balance between bacterial defense and cell viability [65].

On the other hand, it has been reported *luxS*-deleted lactobacilli displayed reduced ability of adhesion to Caco-2 cells, and the presence of *LuxS* was correlated with higher tolerance to stressful conditions related to the gastrointestinal environment [66]. It has also been demonstrated overexpression of *luxS* promotes stress resistance and biofilm formation in LAB [67].

Adhesion is a fundamental step for biofilm formation, but the production of exopolyssacharides (EPS) is necessary for the irreversible bacterial attachment to surfaces [68]. A study with *Lp. plantarum* reveled EPS production was directly correlated with the addition of AI-2 in culture medium, facilitating biofilm formation, and it has also been reported AI-2 may influence the morphology of bacterial colonies [69]. Although the mechanism for modulation of biofilm architecture by AI-2 is still unknown [69], it has been demonstrated mutation of *luxS* interfere with biofilm thickness [70].

Taken altogether, the results from this study and literature reports suggest the presence of multiple *luxS* gene copies, acquired or not by HGT events, may confer adaptive advantages to LAB in different food matrices and environments [71–73].

4.2. Pangenome structure and functional repertoire of LAB

This study revealed an enrichment in cocoa-related LAB strains for gene families related to sugar and amino acid metabolism, especially glycosyl hydrolases and glycosyl transferases, which is in agreement with proteomic evidences from other studies [74] that reported high abundance of glycosyl hydrolases linked the metabolism of galactose, glucose and mannose, as well as glycosyl transferases important for metabolic pathways involving methionine and glutathione. The higher abundance of glycosyl hydrolases and glycosyl transferases in cocoa-related LAB strains is coherent with the literature, as they intermediate the formation of organic acids, especially lactic and acetic, which are later removed by roasting of fermented cocoa beans [75].

Regarding the metabolism of amino acids, proteomic studies performed in several parts of the world [76,77] evidenced that during spontaneous cocoa fermentations there were increased levels of amino acids, both hydrophobic (alanine, leucine, isoleucine, proline, valine, phenylalanine and tyrosine) and hydrophilic (serine, glycine, histidine, threonine and lysine). In cocoa fermentation, oligopeptides are originated by the cleavage of longer peptides at the early stages, which later undergo intense proteolytic activity rendering free aminoacids [78]. These longer peptides observed at initial fermentation steps are originated from storage proteins (albumin and vicilin) from the cotyledons of cocoa seeds, which are metabolized by a combined action of aspartic endoproteases and serine carboxyexopeptidases [79]. Additionally, it has been reported high levels of metalloproteinases [80] and serinopeptidases gene families in the accessory cocoa-related LAB genomes, plus other peptidases encoded by genes from core, which reinforces the potential of LAB for adaptation to this environment rich in free amino acids, which shall be an alternative carbon source for LAB growth.

From a biotechnological point of view on cocoa fermentation, the presence of LAB strains with an extra repertoire of genes related to the metabolism of carbohydrates and amino acids may be advantageous for the release of reducing sugars and oligopeptides, which will become substrates for Maillard reaction during the roast phase of cocoa production, that is crucial for the development of typical cocoa colour and flavor [79]. However, the potential for serine decarboxylation may represent a health risk to consumers due to the production of biogenic amines during fermentation [75].

Other characteristic observed in this study on LAB cocoa genomes, is the presence of extensive HGT signatures, which suggests the incorporation of carbohydrates and amino acids metabolizing-genes in the accessory genome, in an open pangenome structure compatible with the high plasticity of LAB genomes and their ability to adapt to several environments [81,82].

The occurrence of horizontal acquisitions of genes related to carbohydrate metabolism is well recognized in the literature [50,55], since lactobacilli genomes were reduced along evolution and strictly-adapted to the utilization of a limited array of sugars [83]. In counterpart, considering the availability of short-amino acids in the later stages of fermentation, the occurrence of amino acid metabolizing-genes with HGT signatures in some cocoa LAB strains suggest a gain to better thrive in amino acids rich-environments.

5. Conclusions

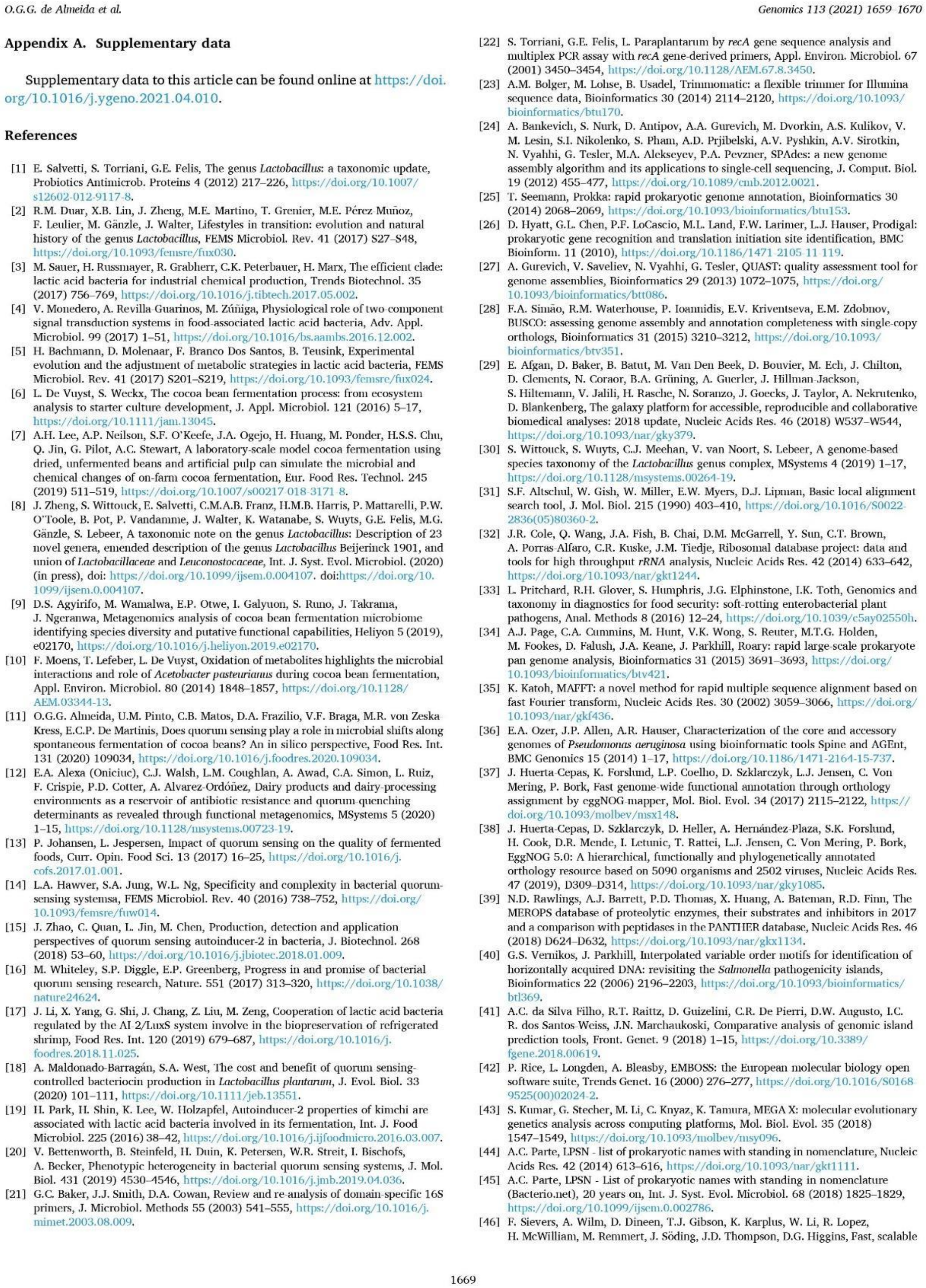
The distribution of *luxS* genes was studied in representative datasets of selected LAB species, with focus on core *luxS* genes for unambiguous differentiation of closely-related homologous sequences. The primers designed for detection of *luxS* genes and prospection of QS represent important tools to further gather evidences on physiological mechanisms for bacterial adaptation to diverse niches. As many metabolic pathways and phenotypic bacterial behaviours are governed by *luxS*, its detection and unequivocal identification through PCR-based methods is crucial for downstream characterization of LAB physiology.

Acknowledgments

Almeida, OGG is grateful to The São Paulo Research Foundation (FAPESP) for supporting the Ph.D. course (2017/13759-6; 2018/26719-5), Brazil. The authors are also grateful to FAPESP for financial support (2018/13564-3), Brazil. This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001. ECPDM is a fellow (PQ-2) of the National Council for Scientific and Technological Development (CNPq), Brazil. We also acknowledge Dr. Veronica Gatto for his technical assistance in molecular biology experiments and the "Centro Piattaforme Tecnologiche" of the University of Verona for the computing resources.

Conflict of interests

The authors declare that they have no conflict of interest.



Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.04.010.

References

[1] E. Salvetti, S. Torriani, G.E. Felis, The genus *Lactobacillus*: a taxonomic update, Probiotics Antimicrob. Proteins 4 (2012) 217–226, https://doi.org/10.1007/s12602-012-9117-8.

[2] R.M. Duar, X.B. Lin, J. Zheng, M.E. Martino, T. Grenier, M.E. Pérez-Muñoz, F. Leulier, M. Gänzle, J. Walter, Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*, FEMS Microbiol. Rev. 41 (2017) S27–S48, https://doi.org/10.1093/femsre/fux030.

[3] M. Sauer, H. Russmayer, R. Grabherr, C.K. Peterbauer, H. Marx, The efficient clade: lactic acid bacteria for industrial chemical production, Trends Biotechnol. 35 (2017) 756–769, https://doi.org/10.1016/j.tibtech.2017.05.002.

[4] V. Monedero, A. Revilla-Guarinos, M. Zúñiga, Physiological role of two-component signal transduction systems in food-associated lactic acid bacteria, Adv. Appl. Microbiol. 99 (2017) 1–51, https://doi.org/10.1016/bs.aambs.2016.12.002.

[5] H. Bachmann, D. Molenaar, F. Branco Dos Santos, B. Teusink, Experimental evolution and the adjustment of metabolic strategies in lactic acid bacteria, FEMS Microbiol. Rev. 41 (2017) S201–S219, https://doi.org/10.1093/femsre/fux024.

[6] L. De Vuyst, S. Weckx, The cocoa bean fermentation process: from ecosystem analysis to starter culture development, J. Appl. Microbiol. 121 (2016) 5–17, https://doi.org/10.1111/jam.13045.

[7] A.H. Lee, A.P. Neilson, S.F. O'Keefe, J.A. Ogejo, H. Huang, M. Ponder, H.S.S. Chu, Q. Jin, G. Pilot, A.C. Stewart, A laboratory-scale model cocoa fermentation using dried, unfermented beans and artificial pulp can simulate the microbial and chemical changes of on-farm cocoa fermentation, Eur. Food Res. Technol. 245 (2019) 511–519, https://doi.org/10.1007/s00217-018-3171-8.

[8] J. Zheng, S. Wittouck, E. Salvetti, C.M.A.B. Franz, H.M.B. Harris, P. Mattarelli, P.W. O'Toole, B. Pot, P. Vandamme, J. Walter, K. Watanabe, S. Wuyts, G.E. Felis, M.G. Gänzle, S. Lebeer, A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*, Int. J. Syst. Evol. Microbiol. (2020) (in press), doi: https://doi.org/10.1099/ijsem.0.004107. doi:https://doi.org/10.1099/ijsem.0.004107.

[9] D.S. Agyirifo, M. Wamalwa, E.P. Otwe, I. Galyuon, S. Runo, J. Takrama, J. Ngeranwa, Metagenomics analysis of cocoa bean fermentation microbiome identifying species diversity and putative functional capabilities, Heliyon 5 (2019), e02170, https://doi.org/10.1016/j.heliyon.2019.e02170.

[10] F. Moens, T. Lefeber, L. De Vuyst, Oxidation of metabolites highlights the microbial interactions and role of *Acetobacter pasteurianus* during cocoa bean fermentation, Appl. Environ. Microbiol. 80 (2014) 1848–1857, https://doi.org/10.1128/AEM.03344-13.

[11] O.G.G. Almeida, U.M. Pinto, C.B. Matos, D.A. Frazilio, V.F. Braga, M.R. von Zeska-Kress, E.C.P. De Martinis, Does quorum sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An in silico perspective, Food Res. Int. 131 (2020) 109034, https://doi.org/10.1016/j.foodres.2020.109034.

[12] E.A. Alexa (Oniciuc), C.J. Walsh, L.M. Coughlan, A. Awad, C.A. Simon, L. Ruiz, F. Crispie, P.D. Cotter, A. Alvarez-Ordóñez, Dairy products and dairy-processing environments as a reservoir of antibiotic resistance and quorum-quenching determinants as revealed through functional metagenomics, MSystems 5 (2020) 1–15, https://doi.org/10.1128/msystems.00723-19.

[13] P. Johansen, L. Jespersen, Impact of quorum sensing on the quality of fermented foods, Curr. Opin. Food Sci. 13 (2017) 16–25, https://doi.org/10.1016/j.cofs.2017.01.001.

[14] L.A. Hawver, S.A. Jung, W.L. Ng, Specificity and complexity in bacterial quorum-sensing systemsa, FEMS Microbiol. Rev. 40 (2016) 738–752, https://doi.org/10.1093/femsre/fuw014.

[15] J. Zhao, C. Quan, L. Jin, M. Chen, Production, detection and application perspectives of quorum sensing autoinducer-2 in bacteria, J. Biotechnol. 268 (2018) 53–60, https://doi.org/10.1016/j.jbiotec.2018.01.009.

[16] M. Whiteley, S.P. Diggle, E.P. Greenberg, Progress in and promise of bacterial quorum sensing research, Nature. 551 (2017) 313–320, https://doi.org/10.1038/nature24624.

[17] J. Li, X. Yang, G. Shi, J. Chang, Z. Liu, M. Zeng, Cooperation of lactic acid bacteria regulated by the AI-2/LuxS system involve in the biopreservation of refrigerated shrimp, Food Res. Int. 120 (2019) 679–687, https://doi.org/10.1016/j.foodres.2018.11.025.

[18] A. Maldonado-Barragán, S.A. West, The cost and benefit of quorum sensing-controlled bacteriocin production in *Lactobacillus plantarum*, J. Evol. Biol. 33 (2020) 101–111, https://doi.org/10.1111/jeb.13551.

[19] H. Park, H. Shin, K. Lee, W. Holzapfel, Autoinducer-2 properties of kimchi are associated with lactic acid bacteria involved in its fermentation, Int. J. Food Microbiol. 225 (2016) 38–42, https://doi.org/10.1016/j.ijfoodmicro.2016.03.007.

[20] V. Bettenworth, B. Steinfeld, H. Duin, K. Petersen, W.R. Streit, I. Bischofs, A. Becker, Phenotypic heterogeneity in bacterial quorum sensing systems, J. Mol. Biol. 431 (2019) 4530–4546, https://doi.org/10.1016/j.jmb.2019.04.036.

[21] G.C. Baker, J.J. Smith, D.A. Cowan, Review and re-analysis of domain-specific 16S primers, J. Microbiol. Methods 55 (2003) 541–555, https://doi.org/10.1016/j.mimet.2003.08.009.

[22] S. Torriani, G.E. Felis, L. Paraplantarum by *recA* gene sequence analysis and multiplex PCR assay with *recA* gene-derived primers, Appl. Environ. Microbiol. 67 (2001) 3450–3454, https://doi.org/10.1128/AEM.67.8.3450.

[23] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.

[24] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V. M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477, https://doi.org/10.1089/cmb.2012.0021.

[25] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069, https://doi.org/10.1093/bioinformatics/btu153.

[26] D. Hyatt, G.L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinform. 11 (2010), https://doi.org/10.1186/1471-2105-11-119.

[27] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, https://doi.org/10.1093/bioinformatics/btt086.

[28] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, https://doi.org/10.1093/bioinformatics/btv351.

[29] E. Afgan, D. Baker, B. Batut, M. Van Den Beek, D. Bouvier, M. Ech, J. Chilton, D. Clements, N. Coraor, B.A. Grüning, A. Guerler, J. Hillman-Jackson, S. Hiltemann, V. Jalili, H. Rasche, N. Soranzo, J. Goecks, J. Taylor, A. Nekrutenko, D. Blankenberg, The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic Acids Res. 46 (2018) W537–W544, https://doi.org/10.1093/nar/gky379.

[30] S. Wittouck, S. Wuyts, C.J. Meehan, V. van Noort, S. Lebeer, A genome-based species taxonomy of the *Lactobacillus* genus complex, MSystems 4 (2019) 1–17, https://doi.org/10.1128/msystems.00264-19.

[31] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.

[32] J.R. Cole, Q. Wang, J.A. Fish, B. Chai, D.M. McGarrell, Y. Sun, C.T. Brown, A. Porras-Alfaro, C.R. Kuske, J.M. Tiedje, Ribosomal database project: data and tools for high throughput *rRNA* analysis, Nucleic Acids Res. 42 (2014) 633–642, https://doi.org/10.1093/nar/gkt1244.

[33] L. Pritchard, R.H. Glover, S. Humphris, J.G. Elphinstone, I.K. Toth, Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens, Anal. Methods 8 (2016) 12–24, https://doi.org/10.1039/c5ay02550h.

[34] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, M.T.G. Holden, M. Fookes, D. Falush, J.A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics 31 (2015) 3691–3693, https://doi.org/10.1093/bioinformatics/btv421.

[35] K. Katoh, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066, https://doi.org/10.1093/nar/gkf436.

[36] E.A. Ozer, J.P. Allen, A.R. Hauser, Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt, BMC Genomics 15 (2014) 1–17, https://doi.org/10.1186/1471-2164-15-737.

[37] J. Huerta-Cepas, K. Forslund, L.P. Coelho, D. Szklarczyk, L.J. Jensen, C. Von Mering, P. Bork, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper, Mol. Biol. Evol. 34 (2017) 2115–2122, https://doi.org/10.1093/molbev/msx148.

[38] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S.K. Forslund, H. Cook, D.R. Mende, I. Letunic, T. Rattei, L.J. Jensen, C. Von Mering, P. Bork, EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Res. 47 (2019), D309–D314, https://doi.org/10.1093/nar/gky1085.

[39] N.D. Rawlings, A.J. Barrett, P.D. Thomas, X. Huang, A. Bateman, R.D. Finn, The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database, Nucleic Acids Res. 46 (2018) D624–D632, https://doi.org/10.1093/nar/gkx1134.

[40] G.S. Vernikos, J. Parkhill, Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands, Bioinformatics 22 (2006) 2196–2203, https://doi.org/10.1093/bioinformatics/btl369.

[41] A.C. da Silva Filho, R.T. Raittz, D. Guizelini, C.R. De Pierri, D.W. Augusto, I.C. R. dos Santos-Weiss, J.N. Marchaukoski, Comparative analysis of genomic island prediction tools, Front. Genet. 9 (2018) 1–15, https://doi.org/10.3389/fgene.2018.00619.

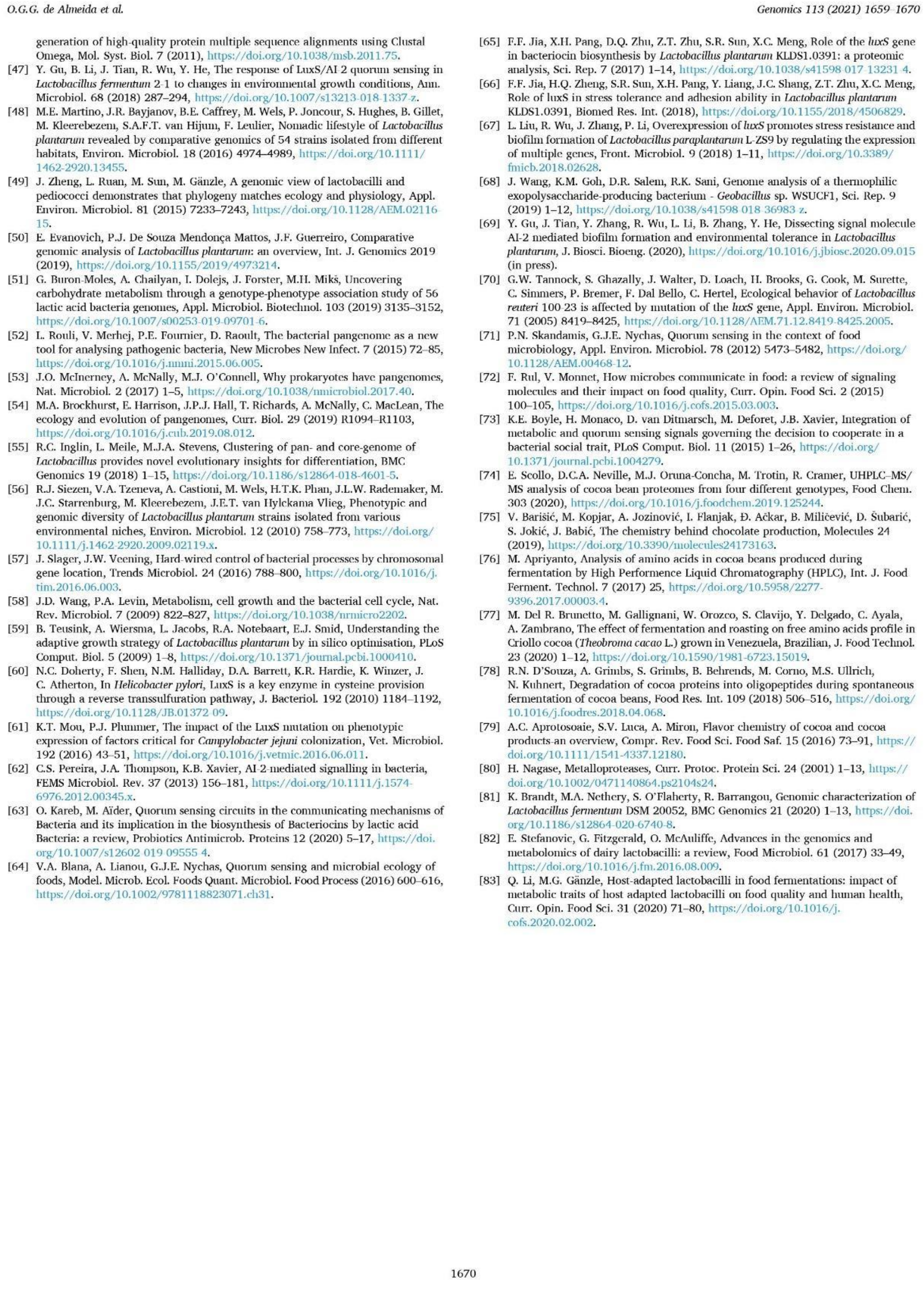
[42] P. Rice, L. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite, Trends Genet. 16 (2000) 276–277, https://doi.org/10.1016/S0168-9525(00)02024-2.

[43] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (2018) 1547–1549, https://doi.org/10.1093/molbev/msy096.

[44] A.C. Parte, LPSN - list of prokaryotic names with standing in nomenclature, Nucleic Acids Res. 42 (2014) 613–616, https://doi.org/10.1093/nar/gkt1111.

[45] A.C. Parte, LPSN - List of prokaryotic names with standing in nomenclature (Bacterio.net), 20 years on, Int. J. Syst. Evol. Microbiol. 68 (2018) 1825–1829, https://doi.org/10.1099/ijsem.0.002786.

[46] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable



generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011), https://doi.org/10.1038/msb.2011.75.

[47] Y. Gu, B. Li, J. Tian, R. Wu, Y. He, The response of LuxS/AI-2 quorum sensing in *Lactobacillus fermentum* 2-1 to changes in environmental growth conditions, Ann. Microbiol. 68 (2018) 287–294, https://doi.org/10.1007/s13213-018-1337-z.

[48] M.E. Martino, J.R. Bayjanov, B.E. Caffrey, M. Wels, P. Joncour, S. Hughes, B. Gillet, M. Kleerebezem, S.A.F.T. van Hijum, F. Leulier, Nomadic lifestyle of *Lactobacillus plantarum* revealed by comparative genomics of 54 strains isolated from different habitats, Environ. Microbiol. 18 (2016) 4974–4989, https://doi.org/10.1111/1462-2920.13455.

[49] J. Zheng, L. Ruan, M. Sun, M. Gänzle, A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology, Appl. Environ. Microbiol. 81 (2015) 7233–7243, https://doi.org/10.1128/AEM.02116-15.

[50] E. Evanovich, P.J. De Souza Mendonça Mattos, J.F. Guerreiro, Comparative genomic analysis of *Lactobacillus plantarum*: an overview, Int. J. Genomics 2019 (2019), https://doi.org/10.1155/2019/4973214.

[51] G. Buron-Moles, A. Chailyan, I. Dolejs, J. Forster, M.H. Mikš, Uncovering carbohydrate metabolism through a genotype-phenotype association study of 56 lactic acid bacteria genomes, Appl. Microbiol. Biotechnol. 103 (2019) 3135–3152, https://doi.org/10.1007/s00253-019-09701-6.

[52] L. Rouli, V. Merhej, P.E. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria, New Microbes New Infect. 7 (2015) 72–85, https://doi.org/10.1016/j.nmni.2015.06.005.

[53] J.O. McInerney, A. McNally, M.J. O'Connell, Why prokaryotes have pangenomes, Nat. Microbiol. 2 (2017) 1–5, https://doi.org/10.1038/nmicrobiol.2017.40.

[54] M.A. Brockhurst, E. Harrison, J.P.J. Hall, T. Richards, A. McNally, C. MacLean, The ecology and evolution of pangenomes, Curr. Biol. 29 (2019) R1094–R1103, https://doi.org/10.1016/j.cub.2019.08.012.

[55] R.C. Inglin, L. Meile, M.J.A. Stevens, Clustering of pan- and core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation, BMC Genomics 19 (2018) 1–15, https://doi.org/10.1186/s12864-018-4601-5.

[56] R.J. Siezen, V.A. Tzeneva, A. Castioni, M. Wels, H.T.K. Phan, J.L.W. Rademaker, M. J.C. Starrenburg, M. Kleerebezem, J.E.T. van Hylckama Vlieg, Phenotypic and genomic diversity of *Lactobacillus plantarum* strains isolated from various environmental niches, Environ. Microbiol. 12 (2010) 758–773, https://doi.org/10.1111/j.1462-2920.2009.02119.x.

[57] J. Slager, J.W. Veening, Hard-wired control of bacterial processes by chromosomal gene location, Trends Microbiol. 24 (2016) 788–800, https://doi.org/10.1016/j.tim.2016.06.003.

[58] J.D. Wang, P.A. Levin, Metabolism, cell growth and the bacterial cell cycle, Nat. Rev. Microbiol. 7 (2009) 822–827, https://doi.org/10.1038/nrmicro2202.

[59] B. Teusink, A. Wiersma, L. Jacobs, R.A. Notebaart, E.J. Smid, Understanding the adaptive growth strategy of *Lactobacillus plantarum* by in silico optimisation, PLoS Comput. Biol. 5 (2009) 1–8, https://doi.org/10.1371/journal.pcbi.1000410.

[60] N.C. Doherty, F. Shen, N.M. Halliday, D.A. Barrett, K.R. Hardie, K. Winzer, J. C. Atherton, In *Helicobacter pylori*, LuxS is a key enzyme in cysteine provision through a reverse transsulfuration pathway, J. Bacteriol. 192 (2010) 1184–1192, https://doi.org/10.1128/JB.01372-09.

[61] K.T. Mou, P.J. Plummer, The impact of the LuxS mutation on phenotypic expression of factors critical for *Campylobacter jejuni* colonization, Vet. Microbiol. 192 (2016) 43–51, https://doi.org/10.1016/j.vetmic.2016.06.011.

[62] C.S. Pereira, J.A. Thompson, K.B. Xavier, AI-2-mediated signalling in bacteria, FEMS Microbiol. Rev. 37 (2013) 156–181, https://doi.org/10.1111/j.1574-6976.2012.00345.x.

[63] O. Kareb, M. Aïder, Quorum sensing circuits in the communicating mechanisms of Bacteria and its implication in the biosynthesis of Bacteriocins by lactic acid Bacteria: a review, Probiotics Antimicrob. Proteins 12 (2020) 5–17, https://doi.org/10.1007/s12602-019-09555-4.

[64] V.A. Blana, A. Lianou, G.J.E. Nychas, Quorum sensing and microbial ecology of foods, Model. Microb. Ecol. Foods Quant. Microbiol. Food Process (2016) 600–616, https://doi.org/10.1002/9781118823071.ch31.

[65] F.F. Jia, X.H. Pang, D.Q. Zhu, Z.T. Zhu, S.R. Sun, X.C. Meng, Role of the *luxS* gene in bacteriocin biosynthesis by *Lactobacillus plantarum* KLDS1.0391: a proteomic analysis, Sci. Rep. 7 (2017) 1–14, https://doi.org/10.1038/s41598-017-13231-4.

[66] F.F. Jia, H.Q. Zheng, S.R. Sun, X.H. Pang, Y. Liang, J.C. Shang, Z.T. Zhu, X.C. Meng, Role of luxS in stress tolerance and adhesion ability in *Lactobacillus plantarum* KLDS1.0391, Biomed Res. Int. (2018), https://doi.org/10.1155/2018/4506829.

[67] L. Liu, R. Wu, J. Zhang, P. Li, Overexpression of *luxS* promotes stress resistance and biofilm formation of *Lactobacillus paraplantarum* L-ZS9 by regulating the expression of multiple genes, Front. Microbiol. 9 (2018) 1–11, https://doi.org/10.3389/fmicb.2018.02628.

[68] J. Wang, K.M. Goh, D.R. Salem, R.K. Sani, Genome analysis of a thermophilic exopolysaccharide-producing bacterium - *Geobacillus* sp. WSUCF1, Sci. Rep. 9 (2019) 1–12, https://doi.org/10.1038/s41598-018-36983-z.

[69] Y. Gu, J. Tian, Y. Zhang, R. Wu, L. Li, B. Zhang, Y. He, Dissecting signal molecule AI-2 mediated biofilm formation and environmental tolerance in *Lactobacillus plantarum*, J. Biosci. Bioeng. (2020), https://doi.org/10.1016/j.jbiosc.2020.09.015 (in press).

[70] G.W. Tannock, S. Ghazally, J. Walter, D. Loach, H. Brooks, G. Cook, M. Surette, C. Simmers, P. Bremer, F. Dal Bello, C. Hertel, Ecological behavior of *Lactobacillus reuteri* 100-23 is affected by mutation of the *luxS* gene, Appl. Environ. Microbiol. 71 (2005) 8419–8425, https://doi.org/10.1128/AEM.71.12.8419-8425.2005.

[71] P.N. Skandamis, G.J.E. Nychas, Quorum sensing in the context of food microbiology, Appl. Environ. Microbiol. 78 (2012) 5473–5482, https://doi.org/10.1128/AEM.00468-12.

[72] F. Rul, V. Monnet, How microbes communicate in food: a review of signaling molecules and their impact on food quality, Curr. Opin. Food Sci. 2 (2015) 100–105, https://doi.org/10.1016/j.cofs.2015.03.003.

[73] K.E. Boyle, H. Monaco, D. van Ditmarsch, M. Deforet, J.B. Xavier, Integration of metabolic and quorum sensing signals governing the decision to cooperate in a bacterial social trait, PLoS Comput. Biol. 11 (2015) 1–26, https://doi.org/10.1371/journal.pcbi.1004279.

[74] E. Scollo, D.C.A. Neville, M.J. Oruna-Concha, M. Trotin, R. Cramer, UHPLC–MS/MS analysis of cocoa bean proteomes from four different genotypes, Food Chem. 303 (2020), https://doi.org/10.1016/j.foodchem.2019.125244.

[75] V. Barišić, M. Kopjar, A. Jozinović, I. Flanjak, Đ. Ačkar, B. Miličević, D. Šubarić, S. Jokić, J. Babić, The chemistry behind chocolate production, Molecules 24 (2019), https://doi.org/10.3390/molecules24173163.

[76] M. Apriyanto, Analysis of amino acids in cocoa beans produced during fermentation by High Performence Liquid Chromatography (HPLC), Int. J. Food Ferment. Technol. 7 (2017) 25, https://doi.org/10.5958/2277-9396.2017.00003.4.

[77] M. Del R. Brunetto, M. Gallignani, W. Orozco, S. Clavijo, Y. Delgado, C. Ayala, A. Zambrano, The effect of fermentation and roasting on free amino acids profile in Criollo cocoa (*Theobroma cacao* L.) grown in Venezuela, Brazilian, J. Food Technol. 23 (2020) 1–12, https://doi.org/10.1590/1981-6723.15019.

[78] R.N. D'Souza, A. Grimbs, S. Grimbs, B. Behrends, M. Corno, M.S. Ullrich, N. Kuhnert, Degradation of cocoa proteins into oligopeptides during spontaneous fermentation of cocoa beans, Food Res. Int. 109 (2018) 506–516, https://doi.org/10.1016/j.foodres.2018.04.068.

[79] A.C. Aprotosoaie, S.V. Luca, A. Miron, Flavor chemistry of cocoa and cocoa products-an overview, Compr. Rev. Food Sci. Food Saf. 15 (2016) 73–91, https://doi.org/10.1111/1541-4337.12180.

[80] H. Nagase, Metalloproteases, Curr. Protoc. Protein Sci. 24 (2001) 1–13, https://doi.org/10.1002/0471140864.ps2104s24.

[81] K. Brandt, M.A. Nethery, S. O'Flaherty, R. Barrangou, Genomic characterization of *Lactobacillus fermentum* DSM 20052, BMC Genomics 21 (2020) 1–13, https://doi.org/10.1186/s12864-020-6740-8.

[82] E. Stefanovic, G. Fitzgerald, O. McAuliffe, Advances in the genomics and metabolomics of dairy lactobacilli: a review, Food Microbiol. 61 (2017) 33–49, https://doi.org/10.1016/j.fm.2016.08.009.

[83] Q. Li, M.G. Gänzle, Host-adapted lactobacilli in food fermentations: impact of metabolic traits of host adapted lactobacilli on food quality and human health, Curr. Opin. Food Sci. 31 (2020) 71–80, https://doi.org/10.1016/j.cofs.2020.02.002.

Supplementary material

Pangenome analyses of LuxS-coding genes and enzymatic repertoires in cocoa-related lactic acid bacteria

Otávio Guilherme Gonçalves de Almeida[a], Nicola Vitulo[b], Elaine Cristina Pereira De Martinis[a], Giovanna E. Felis[b]

Corresponding author at: Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Avenida do Café s/n, Monte Alegre, Ribeirão Preto 14040-903, Brazil. E-mail: edemarti@usp.br

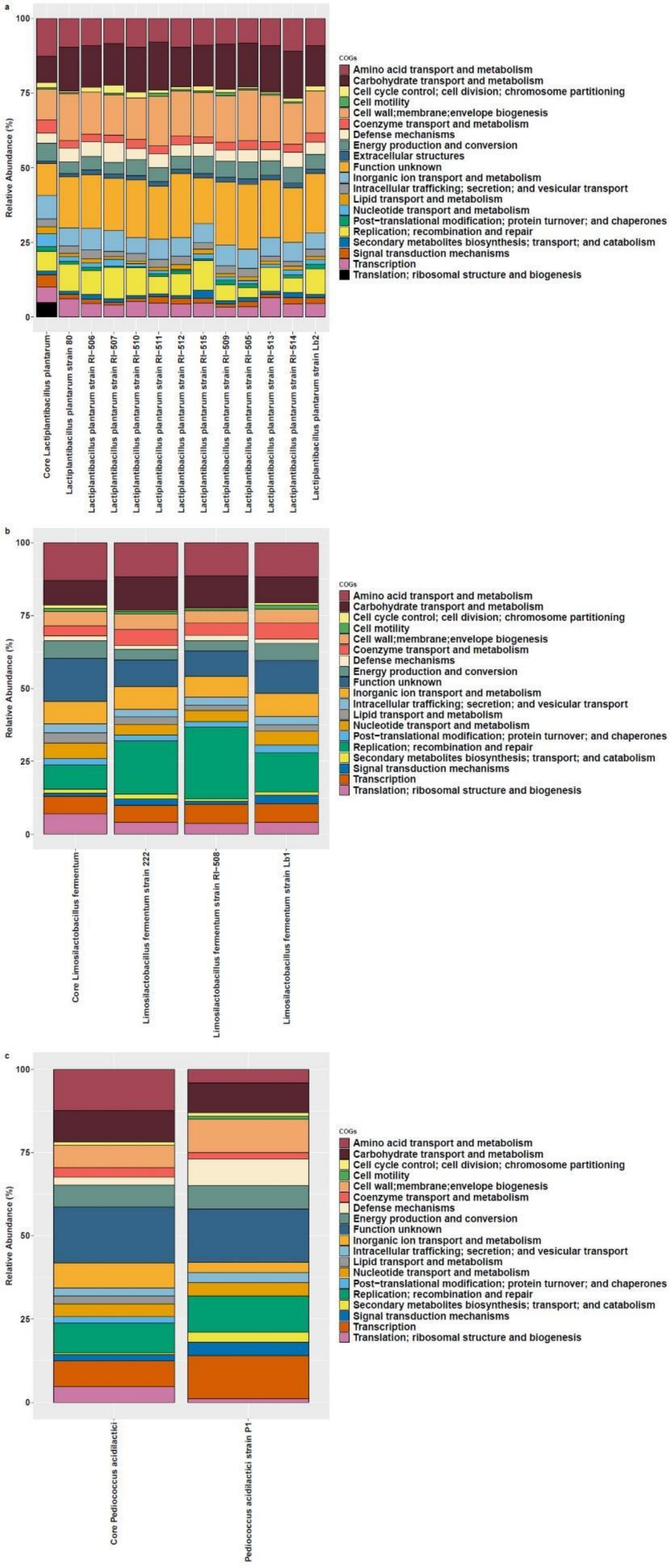


Fig. S1. Cluster of orthologous groups (COGs) of core genomes of *Lp. plantarum*, *Lm. fermentum* and *P. acidilactici* strains compared with the accessory genomes of cocoa derivate strains from the same taxa (items a, b and c, respectively).

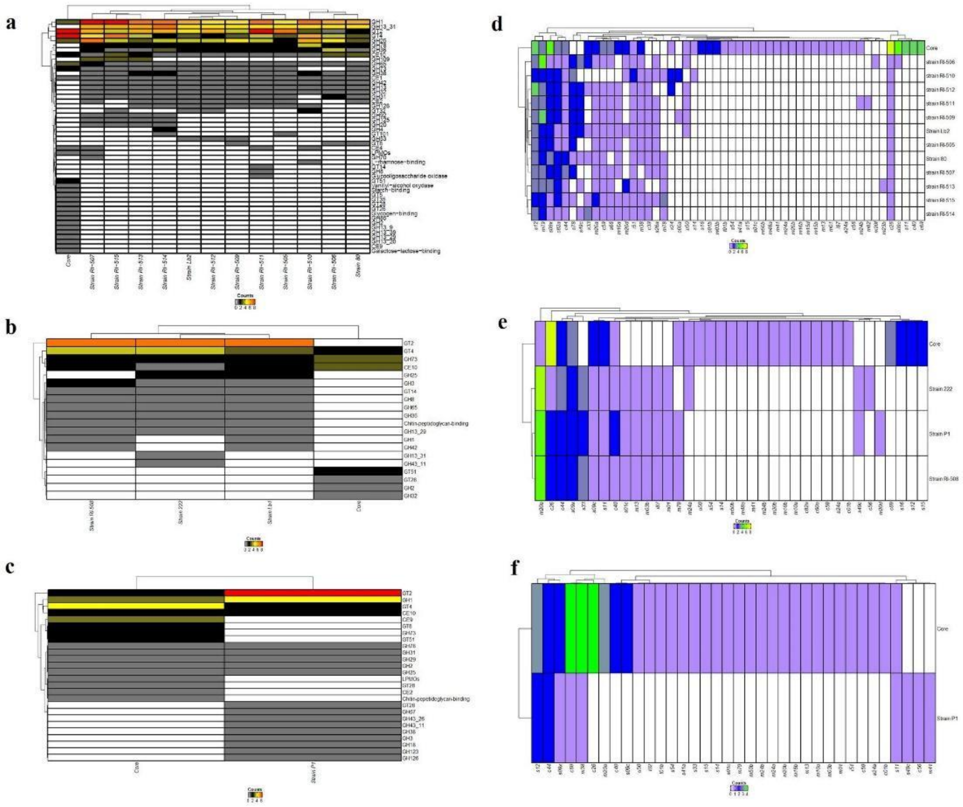


Fig. S2. Carbohydrases (a, c and e) and peptidases gene families (b, d and f) of *Lp. plantarum*, *Lm. fermentum* and *P. acidilactici* strains according to the CAZY and Merops databases, respectively.

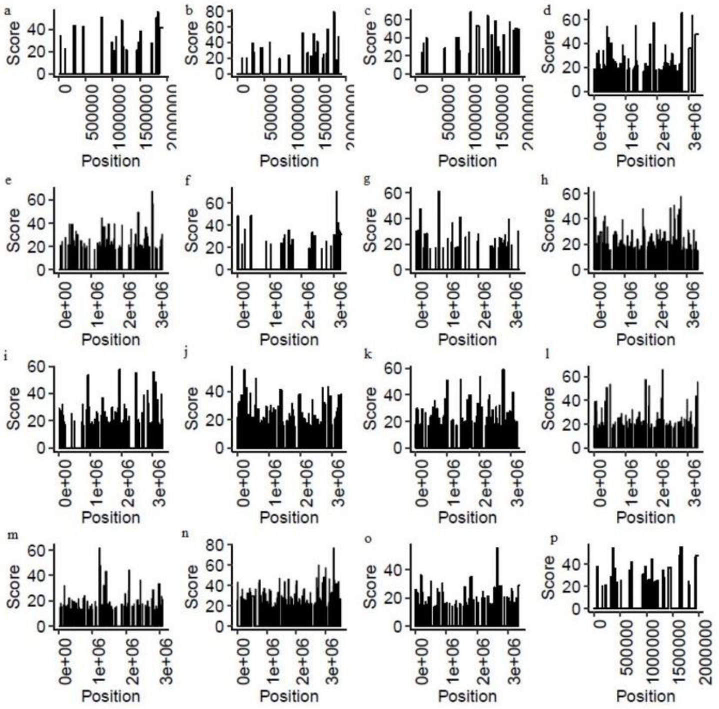


Fig. S3. Predicted HGT-derivate regions in cocoa LAB genomes. (a) *L. fermentum* strain Lb1; (b) *L. fermentum* strain RI-508; (c) *L. fermentum* strain 222; (d) *L. plantarum* strain Lb2; (e) *L. plantarum* strain RI-514; (f) *L. plantarum* strain 80; (g) *L. plantarum* strain RI-509; (h) *L. plantarum* strain RI-510; (i) *L. plantarum* strain RI-511; (j) *L. plantarum* strain RI-512; (k) *L. plantarum* strain RI-513; (l) *L. plantarum* strain RI-515; (m) *L. plantarum* strain RI-505; (n) *L. plantarum* strain RI-507; (o) *L. plantarum* strain Ri-506; (p) *P. acidilactici* strain P1.

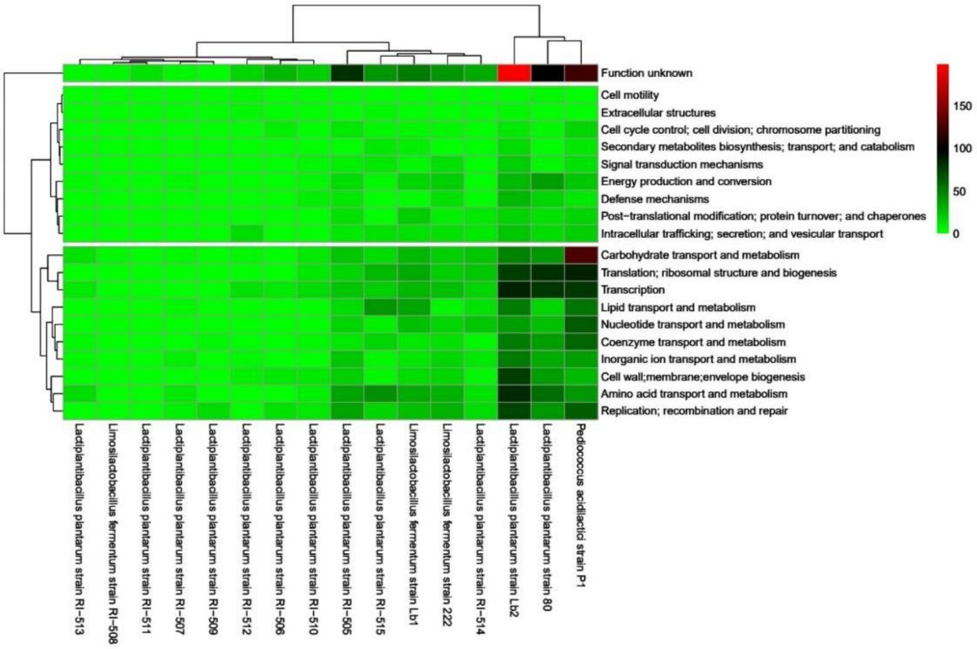


Fig. S4. Cluster of orthologous groups (COGs) of putative HGT regions detected in the cocoa-LAB genomes.

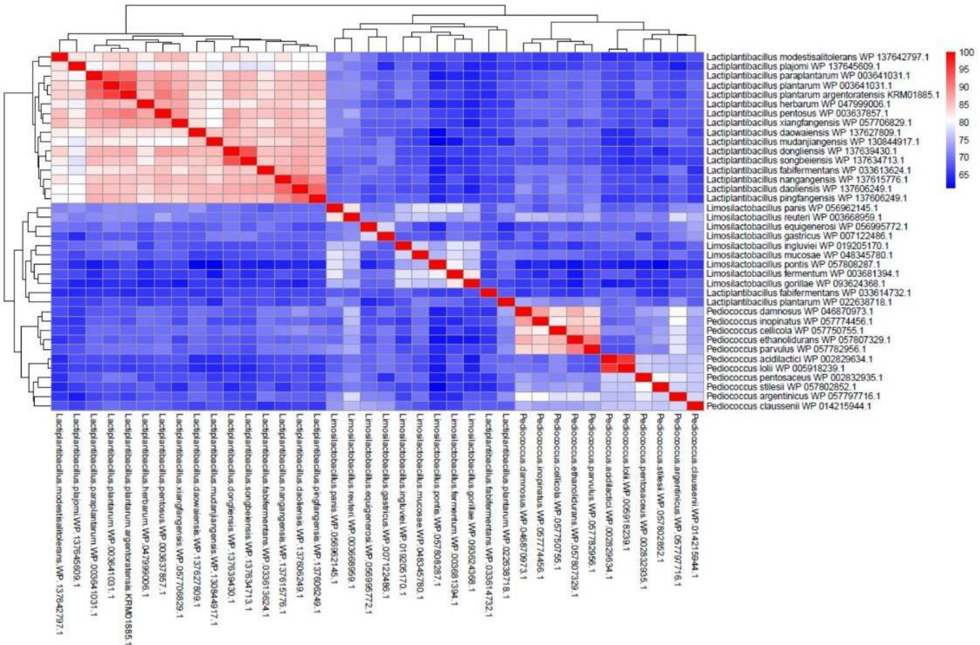
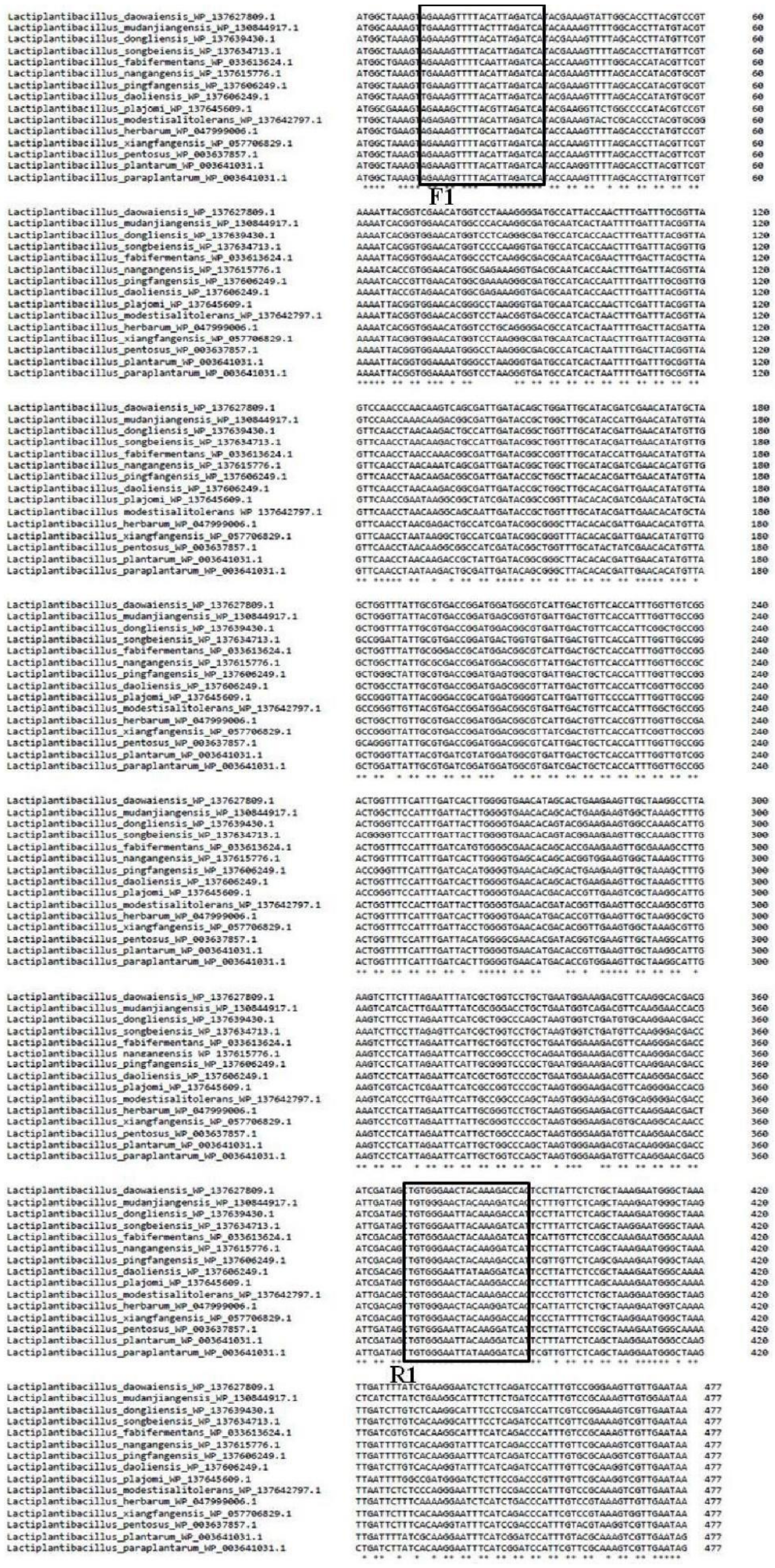


Fig. S5. Heatmap of overall *luxS* gene similarity among type strains from *L. plantarum, L. fermentum* and *P. acidilactici* clades.



```
Lactiplantibacillus_daowaiensis_WP_137627809.1             ATGGCTAAAGTAGAAAGTTTTACATTAGATCATACGAAAGTATTGGCACCTTACGTCCGT    60
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         ATGGCAAAAGTTGAAAGTTTTACTTTAGATCATACAAAAGTTTTGGCACCTTATGTACGT    60
Lactiplantibacillus_dongliensis_WP_137639430.1             ATGGCTAAAGTAGAAAGTTTTACATTAGATCATACGAAAGTTTTAGCACCTTATGTTCGT    60
Lactiplantibacillus_songbeiensis_WP_137634713.1            ATGGCTAAAGTAGAAAGTTTTACATTAGATCATACGAAAGTTTTAGCACCTTATGTTCGT    60
Lactiplantibacillus_fabifermentans_WP_033613624.1          ATGGCTGAAGTAGAAAGTTTTCAATTAGATCACACCAAAGTTTTGGCACCATACGTTCGT    60
Lactiplantibacillus_nangangensis_WP_137615776.1            ATGGCTAAAGTTGAAAGTTTTACATTAGATCATACGAAAGTTTTAGCACCATACGTGCGT    60
Lactiplantibacillus_pingfangensis_WP_137606249.1           ATGGCTAAAGTTGAAAGTTTTACATTAGATCATACGAAAGTTTTAGCACCATACGTGCGT    60
Lactiplantibacillus_daoliensis_WP_137606249.1              ATGGCTAAAGTTGAAAGTTTTACATTAGATCATACGAAAGTTTTAGCACCATATGTGCGT    60
Lactiplantibacillus_plajomi_WP_137645609.1                 ATGGCGAAAGTAGAAAGCTTTACGTTAGATCATACGAAGGTTCTGGCCCCATACGTCCGT    60
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     TTGGCTAAAGTAGAGAGTTTTACATTAGATCATACGAAAGTACTCGCACCCTACGTGCGG    60
Lactiplantibacillus_herbarum_WP_047999006.1                ATGGCTGAAGTAGAAAGTTTTGCATTAGATCATACCAAAGTTTTAGCACCCTATGTCCGT    60
Lactiplantibacillus_xiangfangensis_WP_057706829.1          ATGGCTAAAGTAGAAAGTTTTACGTTAGATCATACCAAAGTTTTAGCACCTTACGTTCGT    60
Lactiplantibacillus_pentosus_WP_003637857.1                ATGGCTAAAGTAGAAAGTTTTACATTAGATCATACCAAAGTTTTAGCACCTTACGTTCGT    60
Lactiplantibacillus_plantarum_WP_003641031.1               ATGGCTAAAGTAGAAAGTTTTACATTAGATCATACCAAGGTTTTAGCACCTTACGTTCGT    60
Lactiplantibacillus_paraplantarum_WP_003641031.1           ATGGCTAAAGTAGAAAGTTTTACATTAGATCATACCAAAGTTTTAGCACCTTATGTTCGT    60
                                                            ****  ****  **  ***  ******** ** ** **   * ** ** ** ** **

Lactiplantibacillus_daowaiensis_WP_137627809.1             AAAATTACGGTCGAACATGGTCCTAAAGGGGATGCCATTACCAACTTTGATTTGCGGTTA    120
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         AAAATCACGGTGGAACATGGCCCACAAGGCGATGCAATCACTAATTTTGATTTACGGTTA    120
Lactiplantibacillus_dongliensis_WP_137639430.1             AAAATCACGGTGGAACATGGTCCTCAGGGCGATGCCATCACCAACTTTGATTTACGGTTA    120
Lactiplantibacillus_songbeiensis_WP_137634713.1            AAAATCACGGTGGAACATGGTCCCCAAGGTGATGCCATCACCAACTTTGATTTACGGTTG    120
Lactiplantibacillus_fabifermentans_WP_033613624.1          AAAATTACGGTGGAACATGGCCCTCAAGGCGACGCAATCACGAACTTTGACTTACGCTTA    120
Lactiplantibacillus_nangangensis_WP_137615776.1            AAAATCACCGTGGAACATGGCGAGAAAGGTGACGCAATCACCAACTTTGATTTACGGTTA    120
Lactiplantibacillus_pingfangensis_WP_137606249.1           AAAATCACCGTTGAACATGGCGAAAAGGGCGATGCCATCACCAATTTTGATTTGCGGTTG    120
Lactiplantibacillus_daoliensis_WP_137606249.1              AAAATTACCGTAGAACATGGCGAGAAAGGTGACGCAATCACCAACTTTGATTTACGGTTA    120
Lactiplantibacillus_plajomi_WP_137645609.1                 AAAATTACGGTGGAACACGGGCCTAAGGGTGATGCAATCACCAACTTCGATTTACGGTTA    120
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     AAAATTACGGTGGAACACGGTCCTAACGGTGACGCCATCACTAACTTTGATTTACGGTTA    120
Lactiplantibacillus_herbarum_WP_047999006.1                AAAATCACGGTGGAACATGGTCCTGCAGGGGACGCCATCACTAATTTTGACTTACGATTA    120
Lactiplantibacillus_xiangfangensis_WP_057706829.1          AAAATTACGGTGGAACATGGTCCTAAGGGCGATGCAATCACTAACTTTGATTTACGGTTA    120
Lactiplantibacillus_pentosus_WP_003637857.1                AAAATTACGGTGGAAAATGGGCCTAAGGGCGACGCCATCACTAATTTTGACTTACGGTTA    120
Lactiplantibacillus_plantarum_WP_003641031.1               AAAATTACGGTGGAAAATGGGCCTAAGGGTGATGCCATCACTAATTTTGATTTGCGGTTA    120
Lactiplantibacillus_paraplantarum_WP_003641031.1           AAAATTACGGTGGAAAATGGTCCTAAGGGTGATGCCATCACTAATTTTGATTTGCGGTTA    120
                                                           ***** ** ** *** * **        ** ** ** ** ** ** ** ** ** ** **

Lactiplantibacillus_daowaiensis_WP_137627809.1             GTCCAACCCAACAAGTCAGCGATTGATACAGCTGGATTGCATACGATCGAACATATGCTA    180
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         GTCCAACCAAACAAGACGGCGATTGATACCGCTGGCTTGCATACCATTGAACATATGTTA    180
Lactiplantibacillus_dongliensis_WP_137639430.1             GTTCAACCTAACAAGACTGCCATTGATACGGCTGGTTTGCATACGATTGAACATATGTTG    180
Lactiplantibacillus_songbeiensis_WP_137634713.1            GTTCAACCTAACAAGACTGCCATTGATACGGCTGGTTTGCATACGATTGAACATATGTTG    180
Lactiplantibacillus_fabifermentans_WP_033613624.1          GTTCAACCTAACCAAACGGCGATTGATACGGCCGGTTTGCATACCATTGAACATATGTTA    180
Lactiplantibacillus_nangangensis_WP_137615776.1            GTTCAACCTAACAAATCAGCGATTGATACGGCTGGCTTGCATACGATCGAACACATGTTG    180
Lactiplantibacillus_pingfangensis_WP_137606249.1           GTTCAACCTAACAAGACGGCGATTGATACCGCTGGCTTACACACGATTGAACATATGTTA    180
Lactiplantibacillus_daoliensis_WP_137606249.1              GTTCAACCTAACAAGACGGCGATTGATACCGCTGGCTTGCACACGATTGAACATATGTTA    180
Lactiplantibacillus_plajomi_WP_137645609.1                 GTTCAACCGAATAAGGCGGCTATCGATACGGCCGGTTTACACACGATCGAACATATGCTA    180
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     GTTCAACCTAACAAGGCAGCAATTGATACCGCTGGTTTGCATACGATTGAACACATGCTA    180
Lactiplantibacillus_herbarum_WP_047999006.1                GTTCAACCTAACGAGACTGCCATCGATACGGCGGGCTTACACACGATTGAACACATGTTA    180
Lactiplantibacillus_xiangfangensis_WP_057706829.1          GTTCAACCTAATAAGGCTGCCATCGATACGGCGGGTTTACACACGATTGAACATATGTTG    180
Lactiplantibacillus_pentosus_WP_003637857.1                GTTCAACCTAACAAGGCGGCCATCGATACGGCTGGTTTGCATACTATCGAACACATGTTA    180
Lactiplantibacillus_plantarum_WP_003641031.1               GTTCAACCTAACAAGACCGCTATTGATACGGCGGGCTTACACACGATTGAACATATGTTA    180
Lactiplantibacillus_paraplantarum_WP_003641031.1           GTTCAACCTAATAAGACTGCGATTGATACAGCGGGCTTACACACGATTGAACACATGTTA    180
                                                           ** ***** **   *   * ** ** ***** ** ** ** ** ** ** ***** *** *

Lactiplantibacillus_daowaiensis_WP_137627809.1             GCTGGTTTATTGCGTGACCGGATGGATGGCGTCATTGACTGTTCACCATTTGGTTGTCGG    240
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         GCTGGGTTATTACGTGACCGGATGAGCGGTGTGATTGACTGTTCACCATTTGGTTGCCGG    240
Lactiplantibacillus_dongliensis_WP_137639430.1             GCTGGTTTATTGCGTGACCGGATGGACGGCGTGATTGACTGTTCACCATTCGGCTGCCGG    240
Lactiplantibacillus_songbeiensis_WP_137634713.1            GCCGGATTATTGCGTGACCGGATGACTGGTGTGATTGACTGTTCACCATTTGGTTGCCGG    240
Lactiplantibacillus_fabifermentans_WP_033613624.1          GCTGGTTTATTGCGGGACCGCATGGACGGCGTCATTGACTGCTCACCATTTGGTTGCCGG    240
Lactiplantibacillus_nangangensis_WP_137615776.1            GCTGGCTTATTGCGCGACCGGATGGACGGCGTTATTGACTGTTCACCATTTGGTTGCCGC    240
Lactiplantibacillus_pingfangensis_WP_137606249.1           GCTGGGCTATTGCGTGACCGGATGAGTGGCGTGATTGACTGCTCACCATTTGGTTGCCGG    240
Lactiplantibacillus_daoliensis_WP_137606249.1              GCTGGCCTATTGCGTGACCGGATGAGCGGCGTTATTGACTGTTCACCATTCGGTTGCCGG    240
Lactiplantibacillus_plajomi_WP_137645609.1                 GCCGGGTTATTACGGGACCGCATGGATGGGGTCATTGATTGTTCCCCATTTGGTTGCCGG    240
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     GCCGGGTTGTTACGTGACCGGATGGACGGCGTGATTGACTGTTCACCATTTGGCTGCCGG    240
Lactiplantibacillus_herbarum_WP_047999006.1                GCTGGCTTGTTGCGTGACCGGATGGACGGCGTCATTGACTGTTCACCGTTTGGTTGCCGA    240
Lactiplantibacillus_xiangfangensis_WP_057706829.1          GCCGGGTTATTGCGTGACCGGATGGACGGCGTTATCGACTGTTCACCATTCGGTTGCCGG    240
Lactiplantibacillus_pentosus_WP_003637857.1                GCAGGGTTATTGCGTGACCGGATGGACGGCGTCATTGACTGCTCACCATTTGGTTGCCGG    240
Lactiplantibacillus_plantarum_WP_003641031.1               GCTGGGTTATTACGTGATCGTATGGATGGCGTGATTGACTGCTCACCATTTGGTTGTCGG    240
Lactiplantibacillus_paraplantarum_WP_003641031.1           GCTGGATTATTGCGTGATCGGATGGATGGCGTGATCGACTGCTCACCATTTGGTTGCCGG    240
                                                           ** **   * ** ** ** ** ***    ** ** ** ** ** ** ** ** ** ** **

Lactiplantibacillus_daowaiensis_WP_137627809.1             ACTGGTTTTCATTTGATCACTTGGGGTGAACATAGCACTGAAGAAGTTGCTAAGGCCTTA    300
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         ACTGGCTTCCATTTGATTACTTGGGGTGAACACAGCACTGAAGAAGTGGCTAAAGCTTTG    300
Lactiplantibacillus_dongliensis_WP_137639430.1             ACTGGGTTCCATTTGATTACTTGGGGTGAACACAGTACGGAAGAAGTGGCCAAAGCATTG    300
Lactiplantibacillus_songbeiensis_WP_137634713.1            ACGGGGTTCCATTTGATTACTTGGGGTGAACACAGTACGGAAGAAGTTGCCAAAGCTTTG    300
Lactiplantibacillus_fabifermentans_WP_033613624.1          ACTGGTTTCCATTTGATCATGTGGGGCGAACACAGCACCGAAGAAGTTGCGAAAGCCTTG    300
Lactiplantibacillus_nangangensis_WP_137615776.1            ACTGGTTTTCATTTGATCACTTGGGGTGAGCACAGCACGGTGGAAGTGGCTAAAGCTTTG    300
Lactiplantibacillus_pingfangensis_WP_137606249.1           ACCGGGTTTCATTTGATCACATGGGGTGAACACAGCACTGAAGAAGTTGCTAAAGCTTTG    300
Lactiplantibacillus_daoliensis_WP_137606249.1              ACTGGTTTCCATTTGATCACTTGGGGTGAACACAGCACTGAAGAAGTTGCTAAAGCTTTG    300
Lactiplantibacillus_plajomi_WP_137645609.1                 ACCGGGTTCCATTTAATCACTTGGGGTGAACACGACACCGTTGAAGTCGCTAAGGCATTG    300
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     ACTGGTTTCCACTTGATTACTTGGGGTGAACACGATACGGTTGAAGTTGCCAAGGCGTTG    300
Lactiplantibacillus_herbarum_WP_047999006.1                ACTGGTTTTCATTTGATCACTTGGGGTGAACATGACACCGTTGAAGTTGCTAAGGCGCTG    300
Lactiplantibacillus_xiangfangensis_WP_057706829.1          ACTGGTTTCCATTTGATTACCTGGGGTGAACACGACACGGTTGAAGTGGCTAAAGCGTTG    300
Lactiplantibacillus_pentosus_WP_003637857.1                ACTGGTTTCCATTTGATTACATGGGGCGAACACGATACGGTCGAAGTTGCTAAGGCATTG    300
Lactiplantibacillus_plantarum_WP_003641031.1               ACTGGTTTTCATTTGATTACTTGGGGTGAACATGACACCGTTGAAGTTGCTAAGGCATTG    300
Lactiplantibacillus_paraplantarum_WP_003641031.1           ACTGGTTTTCATTTGATCACTTGGGGTGAACATGACACCGTGGAAGTTGCTAAGGCATTG    300
                                                           ** ** ** ** ** ** *  ***** ** **     ** *  ***** ** ** **  *

Lactiplantibacillus_daowaiensis_WP_137627809.1             AAGTCTTCTTTAGAATTTATCGCTGGTCCTGCTGAATGGAAAGACGTTCAAGGCACGACG    360
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         AAGTCATCACTTGAATTTATCGCGGGACCTGCTGAATGGTCAGACGTTCAAGGAACCACG    360
Lactiplantibacillus_dongliensis_WP_137639430.1             AAGTCTTCCTTAGAATTCATCGCTGGCCCAGCTAAGTGGTCTGATGTGCAAGGAACGACC    360
Lactiplantibacillus_songbeiensis_WP_137634713.1            AAATCTTCCTTAGAGTTCATCGCTGGTCCTGCTAAGTGGTCTGATGTTCAAGGGACGACC    360
Lactiplantibacillus_fabifermentans_WP_033613624.1          AAGTCTTCCTTAGAATTCATTGCTGGTCCTGCTGAATGGAAAGACGTTCAAGGGACGACC    360
Lactiplantibacillus_nangangensis_WP_137615776.1            AAGTCCTCATTAGAATTCATTGCCGGCCCTGCAGAATGGAAAGACGTTCAAGGGACGACC    360
Lactiplantibacillus_pingfangensis_WP_137606249.1           AAGTCCTCATTAGAATTCATTGCGGGTCCCGCTGAATGGAAAGACGTTCAAGGAACGACC    360
Lactiplantibacillus_daoliensis_WP_137606249.1              AAGTCCTCATTAGAATTCATCGCTGGTCCCGCTGAATGGAAAGACGTTCAAGGGACGACC    360
Lactiplantibacillus_plajomi_WP_137645609.1                 AAGTCGTCACTCGAATTCATCGCCGGTCCCGCTAAGTGGGAAGACGTTCAGGGGACCACG    360
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     AAGTCATCCCTTGAATTCATTGCCGGCCCAGCTAAGTGGGAAGACGTGCAGGGGACGACC    360
Lactiplantibacillus_herbarum_WP_047999006.1                AAATCCTCATTAGAATTCATTGCGGGTCCTGCTAAGTGGGAAGACGTTCAAGGAACGACT    360
Lactiplantibacillus_xiangfangensis_WP_057706829.1          AAGTCCTCGTTAGAATTTATTGCGGGTCCCGCTAAGTGGGAAGACGTGCAAGGCACAACC    360
Lactiplantibacillus_pentosus_WP_003637857.1                AAGTCCTCATTAGAATTCATTGCTGGCCCAGCTAAGTGGGAAGATGTTCAAGGAACGACC    360
Lactiplantibacillus_plantarum_WP_003641031.1               AAGTCCTCATTAGAATTCATTGCTGGCCCAGCTAAGTGGGAAGACGTACAAGGGACGACC    360
Lactiplantibacillus_paraplantarum_WP_003641031.1           AAGTCCTCATTAGAATTCATTGCTGGTCCAGCTAAGTGGGAAGATGTTCAAGGAACGACG    360
                                                           ** ** **  * ** ** ** ** ** ** **  * ***   ** ** ** ** ** **

Lactiplantibacillus_daowaiensis_WP_137627809.1             ATCGATAGCTGTGGGAACTACAAAGACCAGTCCTTATTCTCTGCTAAAGAATGGCTAAA    420
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         ATTGATAGTTGTGGGAACTACAAAGATCAGTCTTTGTTCTCAGCTAAAGAATGGCTAAG    420
Lactiplantibacillus_dongliensis_WP_137639430.1             ATCGATAGCTGTGGGAATTACAAAGACCATTCCTTATTCTCAGCTAAAGAATGGCTAAA    420
Lactiplantibacillus_songbeiensis_WP_137634713.1            ATTGATAGCTGTGGGAATTACAAAGATCATTCTTTATTCTCAGCTAAGGAATGGCTAAA    420
Lactiplantibacillus_fabifermentans_WP_033613624.1          ATCGACAGCTGTGGGAACTACAAAGATCATTCATTGTTCTCCGCCAAAGAATGGCAAAA    420
Lactiplantibacillus_nangangensis_WP_137615776.1            ATCGACAGTTGTGGGAACTACAAGGATCATTCCTTATTCTCAGCTAAAGAATGGCTAAA    420
Lactiplantibacillus_pingfangensis_WP_137606249.1           ATCGACAGTTGTGGGAACTACAAAGACCATTCGTTGTTCTCAGCGAAAGAATGGCTAAA    420
Lactiplantibacillus_daoliensis_WP_137606249.1              ATCGACAGTTGTGGGAATTATAAGGATCATTCCTTATTCTCCGCTAAAGAATGGCAAAA    420
Lactiplantibacillus_plajomi_WP_137645609.1                 ATCGATAGTTGTGGGAACTACAAGGACCAGTCCTTATTTTCAGCAAAAGAATGGCAAAA    420
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     ATTGACAGCTGTGGGAACTACAAAGACCAGTCCCTGTTCTCTGCTAAGGAATGGCTAAG    420
Lactiplantibacillus_herbarum_WP_047999006.1                ATCGACAGTTGTGGGAACTACAAGGATCAGTCATTATTCTCTGCTAAAGAATGGTCAAAA    420
Lactiplantibacillus_xiangfangensis_WP_057706829.1          ATCGACAGTTGTGGGAACTACAAGGACCAGTCCCTATTTTCTGCTAAGGAATGGCTAAA    420
Lactiplantibacillus_pentosus_WP_003637857.1                ATTGATAGCTGTGGGAATTACAAGGATCATTCCTTATTCTCCGCTAAAGAATGGCAAAA    420
Lactiplantibacillus_plantarum_WP_003641031.1               ATCGATAGCTGTGGGAATTACAAGGATCATTCTTTATTCTCAGCTAAGGAATGGCCAAG    420
Lactiplantibacillus_paraplantarum_WP_003641031.1           ATTGATAGTTGTGGGAATTATAAGGATCATTCGTTGTTCTCAGCTAAGGAATGGCTAAG    420
                                                           ** ** **                    **  * ** ** ** ** ****** * **

Lactiplantibacillus_daowaiensis_WP_137627809.1             TTGATTTTATCTGAAGGAATCTCTTCAGATCCATTTGTCCGGGAAGTTGTTGAATAA    477
Lactiplantibacillus_mudanjiangensis_WP_130844917.1         CTCATCTTATCTGAAGGCATTTCTTCTGATCCATTTGTCCGCAAAGTTGTGGAATAA    477
Lactiplantibacillus_dongliensis_WP_137639430.1             TTGATCTTGTCTCAAGGCATTTCCTCCGATCCATTCGTCCGGAAAGTCGTTGAATAA    477
Lactiplantibacillus_songbeiensis_WP_137634713.1            TTGATCTTGTCACAAGGCATTTCCTCAGATCCATTCGTTCGAAAAGTCGTTGAATAA    477
Lactiplantibacillus_fabifermentans_WP_033613624.1          TTGATCGTGTCACAAGGCATTTCATCAGACCCATTTGTCCGCAAAGTTGTTGAATAA    477
Lactiplantibacillus_nangangensis_WP_137615776.1            TTGATTTTGTCACAAGGTATTTCATCAGACCCATTTGTTCGCAAAGTCGTTGAATAA    477
Lactiplantibacillus_pingfangensis_WP_137606249.1           TTGATTTTGTCACAAGGAATTTCATCAGATCCATTTGTGCGCAAGGTCGTTGAATAA    477
Lactiplantibacillus_daoliensis_WP_137606249.1              TTGATCTTGTCACAAGGTATTTCATCAGATCCATTTGTTCGCAAGGTCGTTGAATAA    477
Lactiplantibacillus_plajomi_WP_137645609.1                 TTAATTTTGGCCGATGGGATCTCTTCCGACCCGTTTGTTCGCAAGGTCGTTGAATAA    477
Lactiplantibacillus_modestisalitolerans_WP_137642797.1     TTAATTCTCTCCCAGGGAATTTCTTCCGACCCATTTGTCCGCAAAGTCGTTGAATAA    477
Lactiplantibacillus_herbarum_WP_047999006.1                TTGATTCTTTCAAAAGGAATCTCATCTGACCCATTTGTCCGTAAAGTTGTTGAATAA    477
Lactiplantibacillus_xiangfangensis_WP_057706829.1          TTGATTCTTTCACAAGGAATTTCATCAGACCCATTCGTCCGTAAAGTGGTTGAATAA    477
Lactiplantibacillus_pentosus_WP_003637857.1                TTGATTCTTTCACAAGGTATTTCATCCGACCCATTTGTACGTAAGGTCGTTGAATAA    477
Lactiplantibacillus_plantarum_WP_003641031.1               TTGATTTTATCGCAAGGAATTTCATCGGATCCATTTGTACGCAAAGTCGTTGAATAG    477
Lactiplantibacillus_paraplantarum_WP_003641031.1           CTGATCTTATCACAAGGAATTTCATCGGATCCATTCGTTCGCAAAGTCGTTGAATAG    477
                                                            * **   *   *  * ** ** ** ** ** ** ** ** **   * ** ** *****
```

Fig. S6. Multiple sequence alignment of *L. plantarum* clade species. The conserved regions were highlighted by a black square and the F1 and R1 symbols refer to the sequence's regions for attachment of Luxplan1F and Luxplan1R primers, respectively.

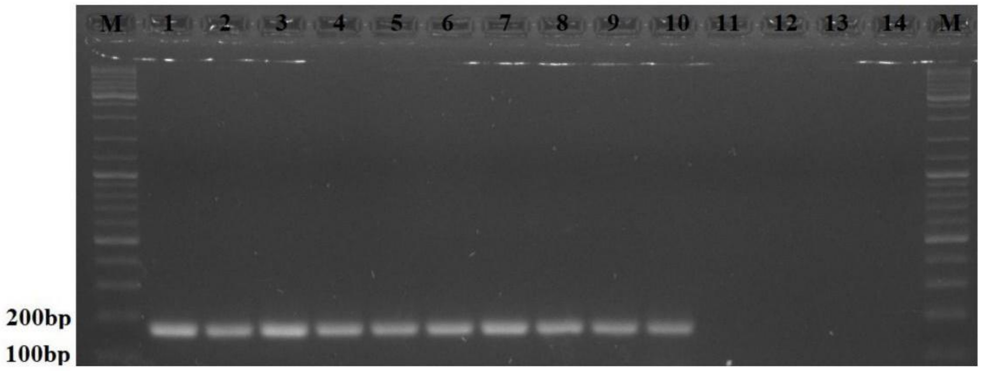


Fig. S7. Agarose gel electrophoresis of amplified *luxS* genes (amplicons with 162 bp) of *L. plantarum* species using specie-specific primers. Lanes: (M) Molecular size marker: (1) *L. plantarum* strain ATCC 14917^{T}; (2) *L. plantarum* strain CECT 4645; (3) *L. plantarum* strain LMG 9208; (4) *L. plantarum* strain PONK 16/4-5A3; (5) *L. plantarum* strain 4SP-5A2; (6) *L. plantarum* strain VSS4B; (7) *L. plantarum* strain VSS2A; (8) *L. plantarum* strain V27A2A04; (9) *L. plantarum* strain Lb2; (10) *L. Paraplantarum*^{T}; (11) *L. pentosus;* (12) *L. fermentum* strain LMG 6902^{T}; (13) *P. acidilactici* strain LMG 11384; (14) negative control.

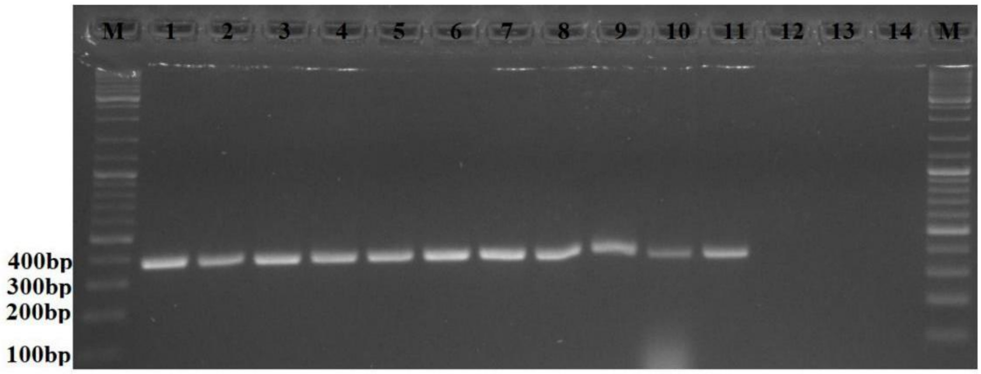


Fig. S8. Agarose gel electrophoresis of amplified *luxS* genes (amplicons with 376 bp) of *L. plantarum*, *L. paraplantarum* and *L. pentosus* species using clade-specific primers. Lanes: (M) Molecular size marker: (1) *L. plantarum* strain ATCC 14917^{T}; (2) *L. plantarum* strain CECT 4645; (3) *L. plantarum* strain LMG 9208; (4) *L. plantarum* strain PONK 16/4-5A3; (5) *L. plantarum* strain 4SP-5A2; (6) *L. plantarum* strain VSS4B; (7) *L. plantarum* strain VSS2A; (8) *L. plantarum* strain V27A2A04; (9) *L. plantarum* strain Lb2; (10) *L. Paraplantarum*^{T}; (11) *L. pentosus;* (12) *L. fermentum* strain LMG 6902^{T}; (13) *P. acidilactici* strain LMG 11384^{T}; (14) negative control.

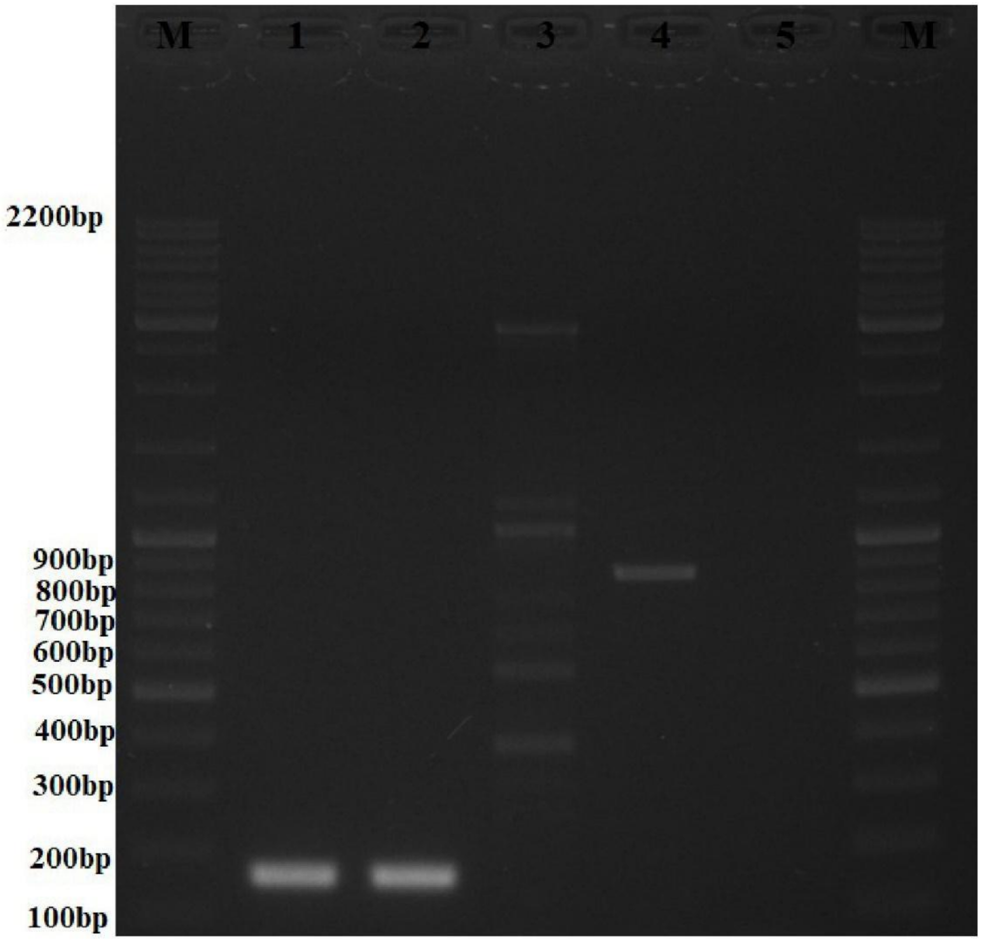
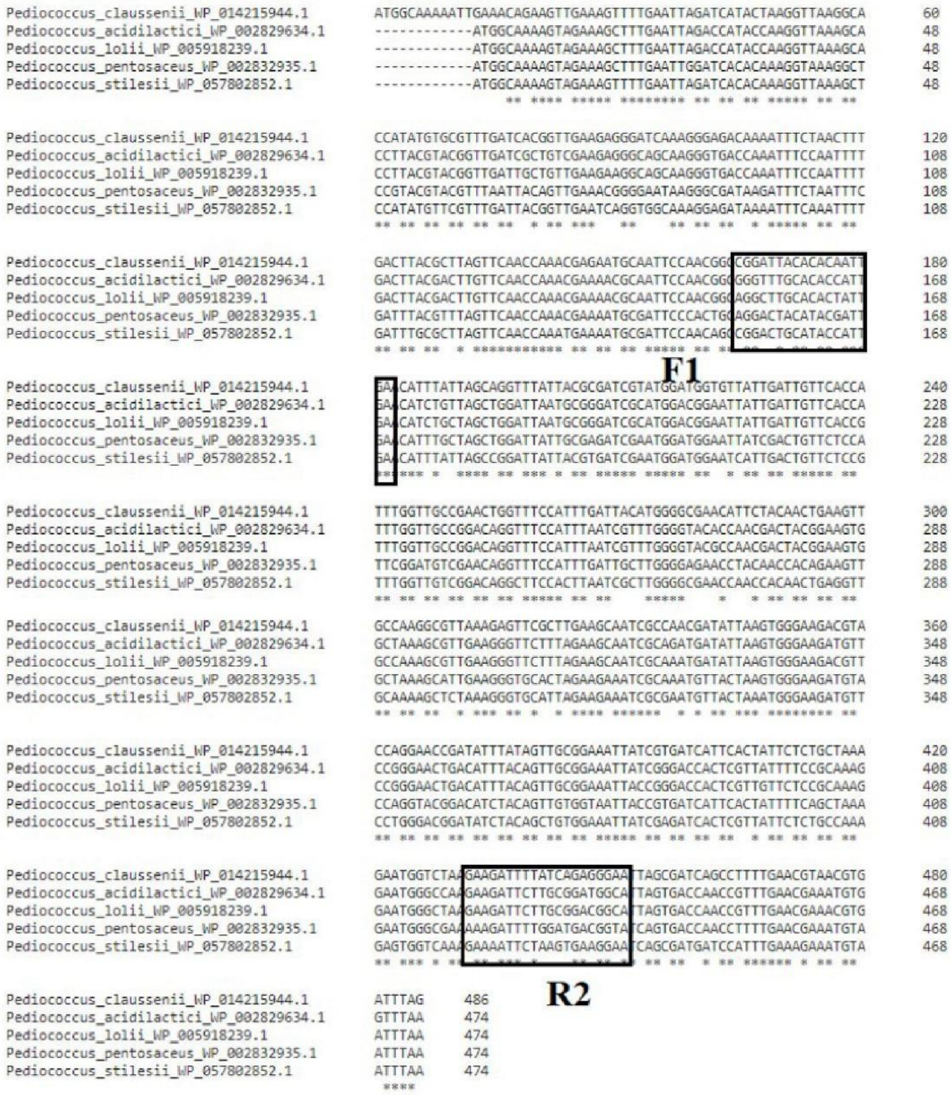


Fig. S9. Agarose gel electrophoresis of amplified *luxS* genes (amplicons with 190 bp) of *L. fermentum* using specie-specific primers. Lanes: (M) Molecular size marker: (1) *L. fermentum* strain LMG 6902^T; (2) *L. fermentum* strain Lb1; (3) *L. plantarum* strain ATCC 14917^T; (4) *P. acidilactici* strain LMG 11384^T; (5) negative control.



```
Pediococcus_claussenii_WP_014215944.1        ATGGCAAAAATTGAAACAGAAGTTGAAAGTTTTGAATTAGATCATACTAAGGTTAAGGCA      60
Pediococcus_acidilactici_WP_002829634.1      ------------ATGGCAAAAGTAGAAAGCTTTGAATTAGACCATACCAAGGTTAAAGCA      48
Pediococcus_lolii_WP_005918239.1             ------------ATGGCAAAAGTAGAAAGCTTTGAATTAGACCATACCAAGGTTAAAGCA      48
Pediococcus_pentosaceus_WP_002832935.1       ------------ATGGCAAAAGTAGAAAGCTTTGAATTGGATCACACAAAGGTAAAGGCT      48
Pediococcus_stilesii_WP_057802852.1          ------------ATGGCAAAAGTAGAAAGTTTTGAATTAGATCACACAAAGGTTAAAGCT      48
                                                             ** **** ***** ******** ** ** ** ***** ** **

Pediococcus_claussenii_WP_014215944.1        CCATATGTGCGTTTGATCACGGTTGAAGAGGGATCAAAGGGAGACAAAATTTCTAACTTT      120
Pediococcus_acidilactici_WP_002829634.1      CCTTACGTACGGTTGATCGCTGTCGAAGAGGGCAGCAAGGGTGACCAAATTTCCAATTTT      108
Pediococcus_lolii_WP_005918239.1             CCTTACGTACGGTTGATTGCTGTTGAAGAAGGCAGCAAGGGTGACCAAATTTCCAATTTT      108
Pediococcus_pentosaceus_WP_002832935.1       CCGTACGTACGTTTAATTACAGTTGAAACGGGGAATAAGGGCGATAAGATTTCTAATTTC      108
Pediococcus_stilesii_WP_057802852.1          CCATATGTTCGTTTGATTACGGTTGAATCAGGTGGCAAAGGAGATAAAATTTCAAATTTT      108
                                             ** ** ** ** ** **  * ** ***   **    ** ** **  * ***** ** **

Pediococcus_claussenii_WP_014215944.1        GACTTACGCTTAGTTCAACCAAACGAGAATGCAATTCCAACGGCCGGATTACACACAATT      180
Pediococcus_acidilactici_WP_002829634.1      GACTTACGACTTGTTCAACCAAACGAAAACGCAATTCCAACGGCGGGTTTGCACACCATT      168
Pediococcus_lolii_WP_005918239.1             GACTTACGACTTGTTCAACCAAACGAAAACGCAATTCCAACGGCAGGCTTGCACACTATT      168
Pediococcus_pentosaceus_WP_002832935.1       GATTTACGTTTAGTTCAACCAAACGAAAATGCGATTCCCACTGCAGGACTACATACGATT      168
Pediococcus_stilesii_WP_057802852.1          GATTTGCGCTTAGTTCAACCAAATGAAAATGCGATTCCAACAGCCGGACTGCATACCATT      168
                                             ** ** **  * *********** ** ** ** ***** ** ** **  * ** ** ***

Pediococcus_claussenii_WP_014215944.1        GAACATTTATTAGCAGGTTTATTACGCGATCGTATGGATGGTGTTATTGATTGTTCACCA      240
Pediococcus_acidilactici_WP_002829634.1      GAACATCTGTTAGCTGGATTAATGCGGGATCGCATGGACGGAATTATTGATTGTTCACCA      228
Pediococcus_lolii_WP_005918239.1             GAACATCTGCTAGCTGGATTAATGCGGGATCGCATGGACGGAATTATTGATTGTTCACCG      228
Pediococcus_pentosaceus_WP_002832935.1       GAACATTTGCTAGCTGGATTATTGCGAGATCGAATGGATGGAATTATCGACTGTTCTCCA      228
Pediococcus_stilesii_WP_057802852.1          GAACATTTATTAGCCGGATTATTACGTGATCGAATGGATGGAATCATTGACTGTTCTCCG      228
                                             ****** *  **** ** *** * ** ***** ***** **  * ** ** ***** **

Pediococcus_claussenii_WP_014215944.1        TTTGGTTGCCGAACTGGTTTCCATTTGATTACATGGGGCGAACATTCTACAACTGAAGTT      300
Pediococcus_acidilactici_WP_002829634.1      TTTGGTTGCCGGACAGGTTTCCATTTAATCGTTTGGGGTACACCAACGACTACGGAAGTG      288
Pediococcus_lolii_WP_005918239.1             TTTGGTTGCCGGACAGGTTTCCATTTAATCGTTTGGGGTACGCCAACGACTACGGAAGTG      288
Pediococcus_pentosaceus_WP_002832935.1       TTCGGATGTCGAACAGGTTTCCATTTGATTGCTTGGGGAGAACCTACAACCACAGAAGTT      288
Pediococcus_stilesii_WP_057802852.1          TTTGGTTGTCGGACAGGCTTCCACTTAATCGCTTGGGGCGAACCAACCACAACTGAGGTT      288
                                             ** ** ** ** ** ** ***** ** **    *****    *   * ** ** ** **

Pediococcus_claussenii_WP_014215944.1        GCCAAGGCGTTAAAGAGTTCGCTTGAAGCAATCGCCAACGATATTAAGTGGGAAGACGTA      360
Pediococcus_acidilactici_WP_002829634.1      GCTAAAGCGTTGAAGGGTTCTTTAGAAGCAATCGCAGATGATATTAAGTGGGAAGATGTT      348
Pediococcus_lolii_WP_005918239.1             GCCAAAGCGTTGAAGGGTTCTTTAGAAGCAATCGCAAATGATATTAAGTGGGAAGACGTT      348
Pediococcus_pentosaceus_WP_002832935.1       GCTAAAGCATTGAAGGGTGCACTAGAAGAAATCGCAAATGTTACTAAGTGGGAAGATGTA      348
Pediococcus_stilesii_WP_057802852.1          GCAAAAGCTCTAAAGGGTGCATTAGAAGAAATCGCGAATGTTACTAAATGGGAAGATGTT      348
                                             ** ** **  * *** ** *  * **** ******  * * ** *** ******** **

Pediococcus_claussenii_WP_014215944.1        CCAGGAACCGATATTTATAGTTGCGGAAATTATCGTGATCATTCACTATTCTCTGCTAAA      420
Pediococcus_acidilactici_WP_002829634.1      CCGGGAACTGACATTTACAGTTGCGGAAATTATCGGGACCACTCGTTATTTTCCGCAAAG      408
Pediococcus_lolii_WP_005918239.1             CCGGGAACTGACATTTACAGTTGCGGAAATTACCGGGACCACTCGTTGTTCTCCGCAAAG      408
Pediococcus_pentosaceus_WP_002832935.1       CCAGGTACGGACATCTACAGTTGTGGTAATTACCGTGATCATTCACTATTTTCAGCTAAA      408
Pediococcus_stilesii_WP_057802852.1          CCTGGGACGGATATCTACAGCTGTGGAAATTATCGAGATCACTCGTTATTCTCTGCCAAA      408
                                             ** ** ** ** ** ** ** ** ** ***** ** ** ** **  * ** ** ** **

Pediococcus_claussenii_WP_014215944.1        GAATGGTCTAAGAAGATTTTATCAGAGGGAATTAGCGATCAGCCTTTTGAACGTAACGTG      480
Pediococcus_acidilactici_WP_002829634.1      GAATGGGCCAAGAAGATTCTTGCGGATGGCATTAGTGACCAACCGTTTGAACGAAATGTG      468
Pediococcus_lolii_WP_005918239.1             GAATGGGCTAAGAAGATTCTTGCGGACGGCATTAGTGACCAACCGTTTGAACGAAACGTG      468
Pediococcus_pentosaceus_WP_002832935.1       GAATGGGCGAAAAAGATTTTGGATGACGGTATCAGTGACCAACCTTTTGAACGAAATGTA      468
Pediococcus_stilesii_WP_057802852.1          GAGTGGTCAAAGAAAATTCTAAGTGAAGGAATCAGCGATGATCCATTTGAAAGAAATGTA      468
                                             ** *** * ** ** *** *    ** ** ** ** **  * ** ****** * ** **

Pediococcus_claussenii_WP_014215944.1        ATTTAG      486
Pediococcus_acidilactici_WP_002829634.1      GTTTAA      474
Pediococcus_lolii_WP_005918239.1             ATTTAA      474
Pediococcus_pentosaceus_WP_002832935.1       ATTTAA      474
Pediococcus_stilesii_WP_057802852.1          ATTTAA      474
                                              ****
```

Fig. S10. Multiple sequence alignment of *P. acidilactici* type strains clade. The conserved regions were highlighted by a black square and the F1 and R1 refer to the LuxPedio_cladeF and LuxPedio_cladeR, respectively.

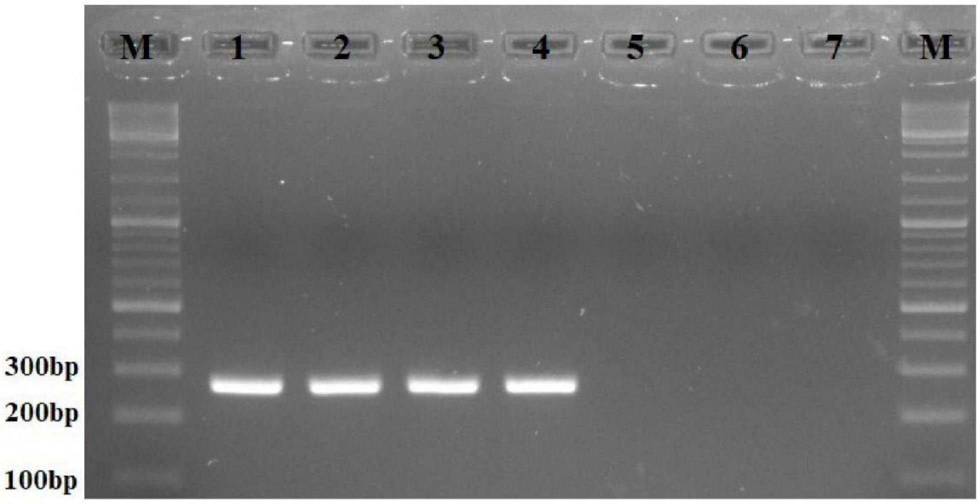


Fig. S11. Agarose gel electrophoresis of amplified *luxS* genes (amplicons with 164 bp) of *P. acidilactici* using clade-specific primers. Lanes: (M) Molecular size marker: (1) *P. acidilactici* strain LMG 11384^T; (2) *P. acidilactici* strain PONK 16/4 5A5; (3) *P. acidilactici* strain P1; (4) *P. acidilactici* strain P2; (5) *L. plantarum* strain ATCC 14917^T; (6) *L. fermentum* strain LMG 6902^T; (7) negative control.

Table S3. *Lp. plantarum* and *Lm. fermentum* genomes and the metrics that support their reclassification.

Accession number	Species name	Strain name	16S similarity against type strain	ANI against type strain
GCA_008084125.1	*Lp. plantarum*	CAB1-7	100%	99%
GCA_008084135.1	*Lp. plantarum*	LSI2-1	100%	99%
GCA_002900075.1	*Lp. plantarum*	ATCC 15578	94%	100%
GCA_003028835.1	*Lp. plantarum*	D1501	100%	99%
GCA_001807655.1	*Lm. fermentum*	HMSC24D01	100%	100%

Chapter 3

5. Chapter 3 - Comparative pangenomic analyses and biotechnological potential of cocoa-related *Acetobacter senegalensis* strains

Specific objectives

- To evaluate, through comparative genomic analysis, the metabolic potential of distinct genotypes of *Acetobacter senegalensis* strains isolated from a spontaneous cocoa fermentation in Brazil in comparison to *A. senegalensis* genomes publicly available on NCBI database obtained from different ecological niches. Additionally, to survey the presence and absence of genetic determinants involved with bacterial tolerance to ethanol, low pH and high-temperature conditions. Finally, to identify genetic determinants involved with quorum sensing and quorum quenching.

The article was published in the Journal *Antonie van Leeuwenhoek* (doi: https://doi.org/10.1007/s10482-021-01684-7), which is a dedicated microbiology Journal. The authorization for reuse in Thesis and Dissertations is depicted in the Attachment D.

Antonie van Leeuwenhoek (2022) 115:111–123
https://doi.org/10.1007/s10482-021-01684-7

ORIGINAL PAPER

Comparative pangenomic analyses and biotechnological potential of cocoa-related *Acetobacter senegalensis* strains

O. G. G. Almeida · M. P. Gimenez · E. C. P. De Martinis

Received: 2 September 2021 / Accepted: 6 November 2021 / Published online: 24 November 2021

Abstract *Acetobacter senegalensis* belongs to the group of acetic acid bacteria (AAB) that present potential biotechnological applications, for production of D-gluconate, cellulose and acetic acid. AAB can overcome heat and acid stresses by using strategies involving the overexpression of heat-shock proteins and enzymes from the complex pyrroquinoline-ADH, besides alcohol dehydrogenases (ADH). Nonetheless, the isolation of *A. senegalensis* and other AAB from food may be challenging due to presence of viable but non-culturable (VBNC) cells and due to uncertainties about nutritional requirements. To contribute for a better understanding of the ecology of AAB, this paper reports on the pangenome analysis of five strains of *A. senegalensis* recently isolated from a Brazilian spontaneous cocoa fermentation. The results showed biosynthetic clusters exclusively found in some cocoa-related AAB, such as those related to terpene pathways, which are important for flavour development. Genes related to oxidative stress were conserved in all the genomes, with multiple clusters. Moreover, there were genes coding for ADH and putative ABC transporters distributed in core, shell and cloud genomes, while chaperonin-encoding genes were present only in the core and soft-core genomes. Regarding quorum sensing, a response regulator gene was in the shell genome, and the gene encoding for acyl-homoserine lactone efflux protein was in the soft-core genome. There were quorum quenching-related genes, mainly encoding for lactonases, but also for acylases. Moreover, *A. senegalensis* did not have determinants of virulence or antibiotic resistance, which are good traits for strains intended to be applied in food fermentation.

Keywords Acetic acid bacteria · *Acetobacter senegalensis* · Pangenome analysis

Introduction

The group of acetic acid bacteria (AAB) refers to obligate aerobic, mainly Gram-negative, catalase-positive, oxidase-negative, ellipsoidal or rod-shaped bacteria able to oxidize ethanol into acetic acid (Gomes et al. 2018). The AAB *Acetobacter senegalensis* has been reported as a thermotolerant bacterium with optimal growth at 35 °C and maximum at 40 °C (Ndoye et al. 2007). Members of this species can produce several industrial commodities, such as gluconic acid (Shafiei et al. 2017), bacterial cellulose (Aswini et al. 2020) and vinegar (Tran et al.

O. G. G. Almeida · M. P. Gimenez ·
E. C. P. De Martinis (✉)
Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Departamento de Análises Clínicas, Toxicológicas e Bromatológicas, Universidade de São Paulo, Avenida do Café s/n, Ribeirão Preto, São Paulo 14040-903, Brazil
e-mail: edemarti@usp.br

2018). The isolation of AAB may be troublesome considering the presence of viable but non culturable (VBNC) cells (Shafiei et al. 2017; De Roos and De Vuyst 2018), summed to the limited knowledge on their nutritional requirements (Wu et al. 2012).

Despite the drawbacks of culture-based strategies, it is well established this genus is highly correlated with alcoholic-related niches (Lynch et al. 2019) due to its metabolic fitness for ethanol oxidisation, as well as heat and acid tolerance. In AAB, the oxidisation of ethanol occurs in two steps that involve an outer cell membrane multienzymatic complex composed by alcohol dehydrogenase (ADH) and pyrroquinoline quinone (PQQ) that converts ethanol to acetaldehyde, which is further oxidised to acetic acid by the action of the enzyme aldehyde dehydrogenase (ALDH) (Gomes et al. 2018; Qiu et al. 2021). Non-dissociated acetic acid can freely cross cell membranes, and its dissociation in the cytoplasm can cause a collapse of proton motive force (Shafiei et al. 2013), hampering cell viability. One microbial strategy to counteract the toxicity of acetic acid is based on its conversion to acetyl-CoA, which can enter the tricarboxylic cycle to be completely oxidised to carbon dioxide and water (Qiu et al. 2021).

Several other mechanisms in AAB can aid in the tolerance to stressful conditions. In *A. senegalensis* LMG23690^{T} it has been shown ethanol is able to elicit up-regulation of heat-shock proteins such as DnaK, GroEL (Hsp60), HtpG (Hsp90), GrpE (Hsp20/Alpha/HspA), besides the chaperonin ClpB (Shafiei et al. 2019). Additionally, proteomic analysis elucidated the participation of ABC transporter proteins and aconitases related to acetic acid resistance (Nakano and Fukaya 2008). Additionally, the overexpression of ADH and the up-regulation of the pathway for biosynthesis of terpenes, which modifies the composition of membrane lipids, may confer tolerance to extreme temperatures (Wang et al. 2015b).

Bacterial adaptation to stressful conditions can also involve quorum sensing (QS) mechanisms, that refer to the synchronization of gene expression mediated by autoinducer molecules (AIs) that are released in situations of high cell-densities (Ng and Bassler 2009). Recently, a metagenomic study revealed effectors and antagonists related to QS in AAB, that potentially straighten the oscillation of acylases, transporters and response regulators along cocoa fermentation (Almeida et al. 2020).

Specifically, in AAB, there are few studies on the importance of QS for metabolism and adaptation to harsh environments. For *Gluconacetobacter intermedius*, the genes *ginI-ginR* activate *ginA* in high cell-densities. This later gene inhibits the production of acetic and gluconic acids, by controlling the expression of *gltA* (glycosyltransferase), *pdeA* (cyclic-di-GMP phosphodiesterase) and *pdeB* (phosphodiesterase/diguanylate cyclase). Moreover, *nagA* (N-acetylglucosamine-6-phosphate deacetylase) is upregulated by *ginA* and promotes exponential growth (Iida et al. 2008, 2009). Similarly, response regulators have also been shown to be important for acetate tolerance in *Acetobacter* spp. (Wang et al. 2015b).

On the other hand, a Quorum Quenching (QQ) mechanism has been described in *Komagataeibacter europaeus*, with strong inhibition of cellulose synthesis by disruption of QS signalling, mediated by the enzyme GqqA (Grandclément et al. 2015; Valera et al. 2016).

Based on the literature, it is evidenced the importance of AAB for diverse fermentative processes, as well as the gap of knowledge on the mechanisms involved in niche adaptation and stress tolerance, which will likely be better understood as growing genetic information builds up in public databases (Qiu et al. 2021). In this scenario, this study describes five novel high-quality complete genomes of *Acetobacter senegalensis* and provides a pangenome analysis to fully evaluate the metabolic pathways and properties of interest for biotechnological applications of this species.

Material and methods

Bacterial isolates

Eight *A. senegalensis* isolates were obtained from a single Brazilian cocoa spontaneous fermentation and were preliminary identified by the sequencing of *16S rRNA* gene (Baker et al. 2003). Genotyping was performed using the $(GTG)_5$-repPCR method (Versalovic et al. 1994). Thus, five genotypes were selected for WGS, named as strains MRS7, GYC10, GYC12, GYC19 and GYC27.

Whole genome sequencing and genome assembly

DNA was extracted using the EasyPure® Bacteria Genomic DNA kit (TransGen Biotech) according to the customer's guidelines. The DNA samples were sent to the company "GenOne Soluções em Biotecnologia" (Rio de Janeiro, Brazil), for WGS carried on the Illumina® NovaSeq6000 platform in the paired-end mode (PE) 2 × 150 bp. The raw reads are available at NCBI bioproject: PRJNA756875. The PE reads were quality-filtered using the bbduk tool (Bushnell 2014). The filtered reads were assembled into genomes using spades version 3.13.0 (Prjibelski et al. 2020). Plasmids were assembled using the plasmidspades version 3.13.0 (Antipov et al. 2016). The assemblies are available at GenBank genomes database (GYC10: JAIMFT000000000; GYC12: JAIMFQ000000000; GYC19: JAIMFR000000000; GYC27: JAIMFS000000000 and MRS7:JAIMFP000000000). The assemblies' quality was evaluated using the QUAST tool (Gurevich et al. 2013) version 5.0.2. The completeness was determined using the BUSCO pipeline (Simão et al. 2015) version 4.0.6 (parameter: rhodospirillales_odb10 database).

Selection of publicly available genomes and pangenome analysis

Publicly available *A. senegalensis* genomes were downloaded from the GenBank and the taxonomic affiliation was confirmed by the average nucleotide identity (ANI) using pyani (Pritchard et al. 2016). Besides, two high-quality metagenome-assembled genomes (MAGs) named bin1 and bin17 were downloaded (Almeida and De Martinis 2021). The MAGs were polished using the MAGpurify pipeline (Nayfach et al. 2019). After polishing, from the bin1 MAG 28 contigs were removed (final genome size of 2.9 Mb) and 59 contigs were discarded from bin17 MAG (final genome size of 2.863 Mb). Then, the assemblies harbouring plasmids were split into chromosomal and plasmidial files and processed separately. All the genomes were annotated using Prokka (Seemann 2014). Pangenome analysis was performed using theget_homologues (Contreras-Moreira and Vinuesa 2013) and the get_phylomarkers pipelines (Vinuesa et al. 2018) following the standard tutorial (https://vinuesa.github.io/get_phylomarkers/#get_phylomarkers-tutorial). The pangenome ML tree was visualized on the iTOL software (Letunic and Bork 2019).

Metabolic predictions

Genomic metabolic predictions were annotated on the Rast*k* environment (Brettin et al. 2015) and on the MicrobeAnnotator pipeline (Ruiz-Perez et al. 2021) (–light mode). The clusters of biosynthetic genes were predicted using antiSMASH version 6.0 (Blin et al. 2021).

Search for carbohydrate-active enzymes, bacteriocins, virulence and antibiotic resistance genes

Carbohydrate-active enzymes were searched using the dbCAN2 tool (Huang et al. 2018). Only the annotations consensually determined by the three methods implemented in the pipeline were reported. The enzymes were classified in glycosyl hydrolases (GH), glycosyltransferases (GT), carbohydrate-binding module (CBM) and auxiliary activities (AA). Bacteriocins, virulence, antibiotics and metal resistance genes were searched using the following databases implemented on ABRICATE pipeline (Seeman): Bactibase (Hammami et al. 2007), ARG-ANNOT (Gupta et al. 2014), CARD (Alcock et al. 2020), bacmet2 (Pal et al. 2014), Resfinder (Zankari et al. 2012), MegaRes (Doster et al. 2020) and VFDB (Chen et al. 2016).

Determination of the localization of genes of interest in core and accessory genomes

The PpanGGOLiN pipeline (Gautreau et al. 2020) was used to search in the pangenomes genes related to alcohol, heat and acetic acid tolerance (e.g. *groES/EL, grpE,* ABC transporter, aconitase, *adh*, *adhA*, *clpB, pqq, dnaJ/K*) and QS/QQ (e.g. acyl-homoserine lactones, *luxR*, acylases, lactonases).

Results and discussion

Five distinct genotypes of cocoa-related *A. senegalensis* isolates were selected by PCR fingerprinting (results not shown) and sequenced by WGS for comparison with publicly available strains (Table 1).

Genome assemblies and completeness

Figure 1A shows *A. senegalensis* presents uniform distribution of GC content, genome size and number of CDS, in agreement with former public genomes (GC contents of 55.4 to 55.7%). Moreover, GC content of plasmidial contigs ranged at 55–56%, the size was 0.08–0.42 Mb, and had 14 to 374 coding-DNA sequences. Figure 1B shows non-fragmented high-quality genomes were obtained for the novel strains, with 99.9% average completeness (compared to 99.39% for public data).

Taxonomy assignment of *A. senegalensis* strains

Two-independent methods were used for taxonomic assignment (Fig. 2A), and the core genome estimated by three independent methods revealed a total of 1,665 single-copy core genes (Fig. 2B), yielding 375 high-quality genes filtered to build the cgML tree, which revealed there was no correlation of any *A. senegalensis* strain with its respective isolation source (Fig. 2A). The ANI metrics (Fig. 2C) indicated 100% of nucleotide similarity between the strains *A. senegalensis* MRS7 and GYC19, with more than 99% of similarity among those and *A. senegalensis* A3, with an overall similarity $\geq$ 97%.

Pangenome reveals a set of metabolic pathways and clusters of biosynthetic genes distributed in *A. senegalensis* genomes

Pangenome size of *A. senegalensis* was composed of 6,485 gene families, divided as 1,710 core, 567 strict soft-core, 1,160 shell and 3,048 cloud genes. From these, a ML pangenome tree was built, including data of predicted metabolic pathways (Fig. 3A), and it revealed there was no clustering based on isolation sources or patterns for niche-specific adaptations.

At functional level, there were no remarkable differences in the numbers of genes per metabolic pathways in *A. senegalensis*, and genetic determinants related to osmotic, oxidative and periplasmic stresses were found in all strains (Fig. 3A), except for the potential to tolerate osmotic stress more evident in *A. senegalensis* strains 108B, DmL050, A3, LMG23690^{T} and bin17. On the other hand, important genetic determinants for adaptation to oxidative stress were found in *A. senegalensis* GYC12, 108B, DmL050, A3, MRS7, GYC19 and GYC27.

In AAB, adaptation to several stressors is highly linked to the ability to ferment sugars (Yang et al. 2019) and to metabolism of amino acids (Xia et al. 2016, 2020; Yin et al. 2017). Figure 3A shows only *A. senegalensis* GYC12, GYC10 and the type strain presented a gene for sucrose utilisation. On the other hand, the majority of *A. senegalensis* strains studied were capable of metabolizing maltose, maltodextrin, trehalose, lactate, mannose and glycogen, besides presenting the ability to synthesize butanol and to ferment butyrate (Fig. 3A). In the literature, it has been described that maltooligosaccharides and

Table 1 Genomes and metagenome-assembled genomes used in this study

Strain	Source	Country	Bioproject or reference
108B	Cocoa fermentation	Ghana	PRJEB7128
A3	Industrial ethanol fermentation	Brazil	PRJNA577465
bin17*	Cocoa fermentation	Brazil	Almeida and De Martinis (2021)
bin17*	Cocoa fermentation	Brazil	Almeida and De Martinis (2021)
DmL050	Insect (*Drosophila melanogaster*)	USA	PRJNA253128
GYC10	Cocoa fermentation	Brazil	PRJNA756875
GYC12	Cocoa fermentation	Brazil	PRJNA756875
GYC19	Cocoa fermentation	Brazil	PRJNA756875
GYC27	Cocoa fermentation	Brazil	PRJNA756875
LMG23690^{T}	Mango fruit	Belgium	PRJNA288385
MRS7	Cocoa fermentation	Brazil	PRJNA756875

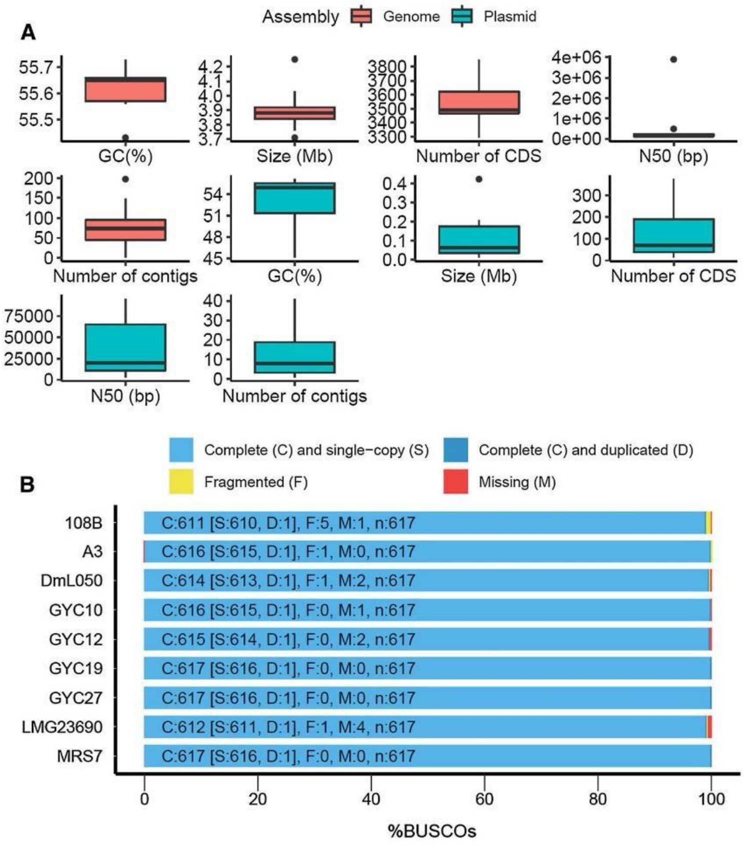


Fig. 1 Quality assembly metrics of **A** genomes and plasmids and **B** comparison of the completeness of the novel five distinct *A. senegalensis* strains regarding those publicly available

trehalose present convergent pathways in *Acetobacter* spp., and AAB carry the enzymes maltooligosyl trehalose synthase and maltooligosyl trehalase, which catalyse the metabolism of maltooligosaccharides into trehalose, that ends up in the central carbohydrate metabolism (De Roos et al. 2020). It has been shown trehalose protects AAB against osmotic stress (Zhang et al. 2015), although the effect may be limited due to the inhibition of the biosynthesis of this carbohydrate in high concentrations of acetic acid (Yang et al. 2019).

Some authors suggested in mixed culture of lactic acid bacteria (LAB) and AAB, the production of lactate by LAB could facilitate AAB growth (Pelicaen et al. 2019). Besides, butyrate is generated by the lactate consume through the action of lactate dehydrogenase (Esquivel-Elizondo et al. 2017), explaining the strains and MAGs' potentials to participate in butyrate fermentation. The utilisation of lactate, fructose, glucose and trehalose as carbon sources in AAB may stimulate cellulose production (Son et al. 2001). In fact, the detection of multiple genes for glycosyl hydrolases in all the genomes studied (Fig. 3A) correlates with the production of levans in AAB mediated by glycosyl hydrolases of the family 32 (Jakob et al. 2019), and levans are required for bacterial cellulose production (La China et al. 2018). Besides, glycosyl transferases participate of acetan

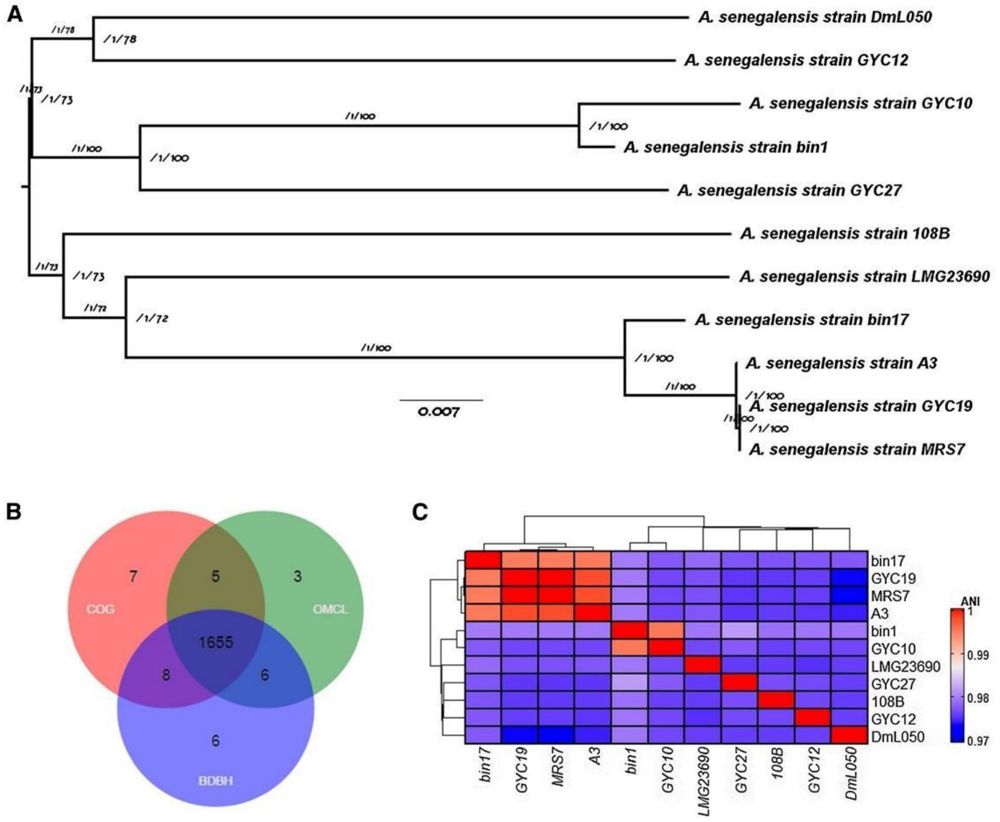


Fig. 2 Assessment of *A. senegalensis* taxonomy. **A** Core genome Maximum-likelihood (ML) tree depicting the proximal relationships among *A. senegalensis* strains based on SNP polymorphisms. The cgML tree was built based on the filtering of the best single-copy core genes (SCGs) filtered from **B** the 1655 SCGs previously determined by three independent clustering algorithms: OMCL, BDBH and COG triangles. In **C** the Average Nucleotide Identity (ANI) was measured to assess the unambiguous identification of the five novel strains: MRS7, GYC10, GYC12, GYC19 and GYC27 inside the *A. senegalensis* species

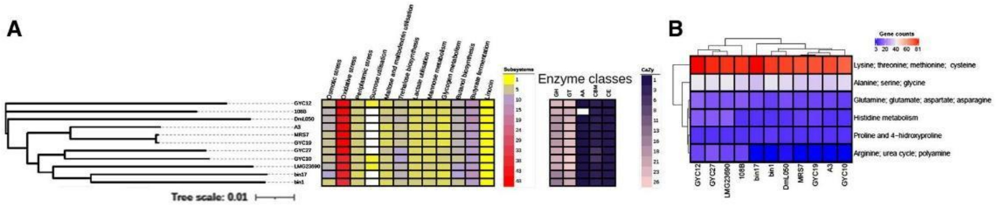


Fig. 3 Pangenome and main metabolic pathways. **A** Pangenome Maximum-likelihood (ML) tree depicting association of *A. senegalensis* strains and carbohydrate-related metabolic pathways. Next, **B** clustering analysis of strains versus amino acids metabolism

biosynthesis in AAB (Ishida et al. 2002), which is a water-soluble polysaccharide with commercial applications and fundamental for biofilm formation by AAB (Trček et al. 2021).

Other enzyme classes and modules were depicted in *A. senegalensis* genomes (Fig. 3A). Auxiliary activities are involved with plant cell wall degradation and aid in the access to carbohydrates by GH, GT and CE enzymes (Levasseur et al. 2013). Only the *A. senegalensis* 108B did not present any gene related to AA activity. Regarding CBM, the number of genes coding for these proteins was higher in the *A. senegalensis* strains GYC10 and 27. CBMs improve the metabolism of complex polysaccharides such as cellulose and starch (Carvalho et al. 2015). Conversely, the putative enzymatic repertoire concerning CE was conserved in all the strains. CEs are involved with modification, breakdown and assembly of glyco and non-glyco conjugated carbohydrates (Nakamura et al. 2017). The detection of CAZY enzymes in plasmids was relevant only in the GYC12 plasmid, which presented GH (n = 2), GT (n = 1) and polysaccharide lyases (PL; n = 1), this last representing a group of enzymes that metabolizes uronic acid-containing polysaccharides (Chakraborty et al. 2016).

Genes involved with lysine, threonine, methionine and cysteine were enriched in all genomes, while genes encoding for alanine, serine and glycine were slightly more abundant in GYC12, LMG23690^{T} and bin1 (Fig. 3B). Besides, histidine metabolizing-genes were particularly abundant in the genomes of the *A. senegalensis* strains GYC12, GYC17 and LMG23690^{T}. Genes involved with arginine and urea cycle were more abundant in the strains GYC27 and 108B (Fig. 3B).

A great number of genes involved in the metabolism of those amino acids potentially enhance bacterial fitness in acidic environments: lysine decarboxylase activity leads to cadaverine production, which is a polyamine that can raise the cellular pH (Wang et al. 2015a). Analogously, genes related to deamination of amino acids results in increased intracellular ammonia concentration, that neutralizes acids and maintain intracellular pH (Yin et al. 2017). Similar mechanisms to regulate intracellular pH are performed by glycine deamination (Xia et al. 2016) and of urea decarboxylation (Wang et al. 2015a). However, a transcriptomic study showed increased ethanol concentrations downregulated pathways for decomposition of urea (Yang et al. 2019). Besides, arginine-related repertoire might be advantageous for AAB, considering proteins decorated with salt-bridge Arg residues are more thermostable (Matsutani et al. 2011).

In this research, only the metabolic pathways with at least 80% of completeness for carbohydrate and amino acid metabolism were considered for comparisons (Fig. 4). All but bin1 *A. senegalensis* strains presented the arginine and proline metabolism pathway complete. With regard to the metabolism of aromatic amino acids, *A. senegalensis* bin17 and LMG23690^{T} presented ca. 87% of the repertoire for tryptophan biosynthesis, but the type strain did not present genes for shikimate pathway. Another exception was *A. senegalensis* bin17, with less than 50% of the genes needed for the synthesis of valine and leucine. Moreover, only *A. senegalensis* LMG23690^{T} and GYC12 strains presented 100% of the genes required for glutamate synthesis from histidine. Conversely, the plasmids found in *A. senegalensis* strains GYC12 and GYC27 presented 100% and 80% of completeness for glutamate synthesis from histidine (data not shown). All strains presented the complete pathway for betaine biosynthesis and only the strain 108B presented genes involved with threonine biosynthesis from aspartate conversion (Fig. 4).

In relation to carbohydrates metabolism, the most complete pathways were related to the pentose phosphate cycle, ribose 5P synthesis and to the citrate cycle (this latter except for the MAGs), while pyruvate oxidation to acetyl-CoA was complete only in the bin17 MAG (Fig. 4).

Regarding the potential of *A. senegalensis* to produce secondary metabolites, only the cocoa-related strains bin1, bin17 and 108B presented biosynthetic gene clusters in their genomes, and no plasmid carried a biosynthetic gene cluster (BCG). A cluster for terpene synthesis was found in all three strains and it was composed by genes encoding for squalene-hopene synthase, hydroxysqualene dehydroxylase, presqualene diphosphate synthase and hydroxysqualene synthase. Another cluster, involved with redox-cofactor PQQ, was found only in the *A. senegalensis* bin17 and 108B. This cluster was formed by the genes encoding for PqqA peptide cyclase, PqqA binding protein and bifunctional coenzyme PQQ synthesis protein C/D. Moreover, the PqqA peptide cyclase and the bifunctional coenzyme PQQ synthesis protein C/D formed a

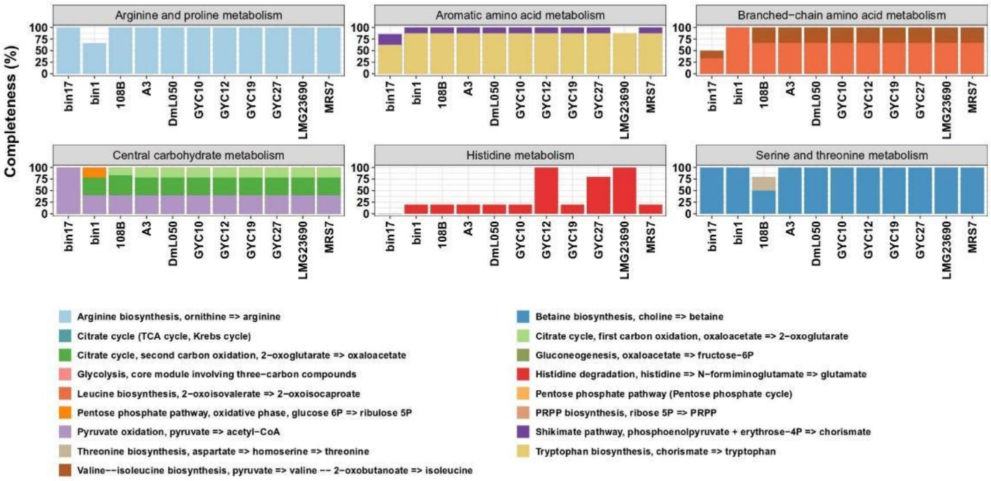


Fig. 4 Amino acids and carbohydrate-related metabolic pathways completeness. Only the metabolic pathways presenting more than 80% of completeness were plotted

distinct cluster in *A. senegalensis* bin1, characterizing a ranthipeptide cluster—cysteine enriched peptides.

A homoserine lactone cluster also was found in all three genomes of *A. senegalensis*, and presented a unique gene annotated as a hypothetical protein. Besides, non-ribosomal peptide synthetase (NRPS) and NRPS-like clusters were found only in the *A. senegalensis* bin1 and 108B, composed by a di-modular non-ribosomal peptide synthase and a O-mycaminosyltylonolide 6-deoxyallosyltransferase—this last one present only in the NRPS-like cluster. Another unspecified ribosomally synthesised and post-translationally modified peptide product (RiPP-like) cluster was conserved in the three *A. senegalensis* strains and it was composed by the linocin M-18 bacteriocin coding-gene.

BCGs related to terpenes biosynthesis are related to heat stress resistance in *Acetobacter* as they integrate biological membranes (Siedenburg and Jendrossek 2011), allowing to maintain its fluidity in higher temperatures (Wang et al. 2015a). Redox-cofactor BCG is another interesting adaptation. The production of PQQ is closely-related to bacterial adaptation to higher temperatures and lower pH because when PQQ is synthesized in higher quantities it can balance the levels of acetic acid and low-pH through energy production, allowing cell growth by providing energy to circumvent cellular injuries (Gao et al. 2021).

Homoserine lactones are well described in *Gluconobacter* spp. and they correlate with bacterial cellulose synthesis and the control of acetic acid production in a possible QS-mediated signalling (Liu et al. 2019). On the other hand, NRPS-like BCG has been related to the synthesis of several compounds such as toxins, antibiotics, siderophores and pigments (Martínez-Núñez and López 2016), but its importance for AAB is still not fully understood.

Finally, neither antibiotic nor metal-resistance genes were found in any bacterial genome or associated plasmid, evidencing the safety of the surveyed strains to be applied in industrial processes.

Genomic localization of adaptive-response genes for alcohol, acetic acid and oxidative stress tolerance in *A. senegalensis* genomes

The genes shared by all the strains and MAGs were referred as core genes, and they were determined based on homology and vicinity of gene families in the context of shared genomic regions (Fang et al. 2008). Paralogous genes trend to conserve the function but could change the sequence composition along the evolution time (Koonin 2005), which results in their aggregation in distinct clusters in pangenome analysis (Almeida et al. 2021). For this reason, the genes with several copy numbers and with the same functional

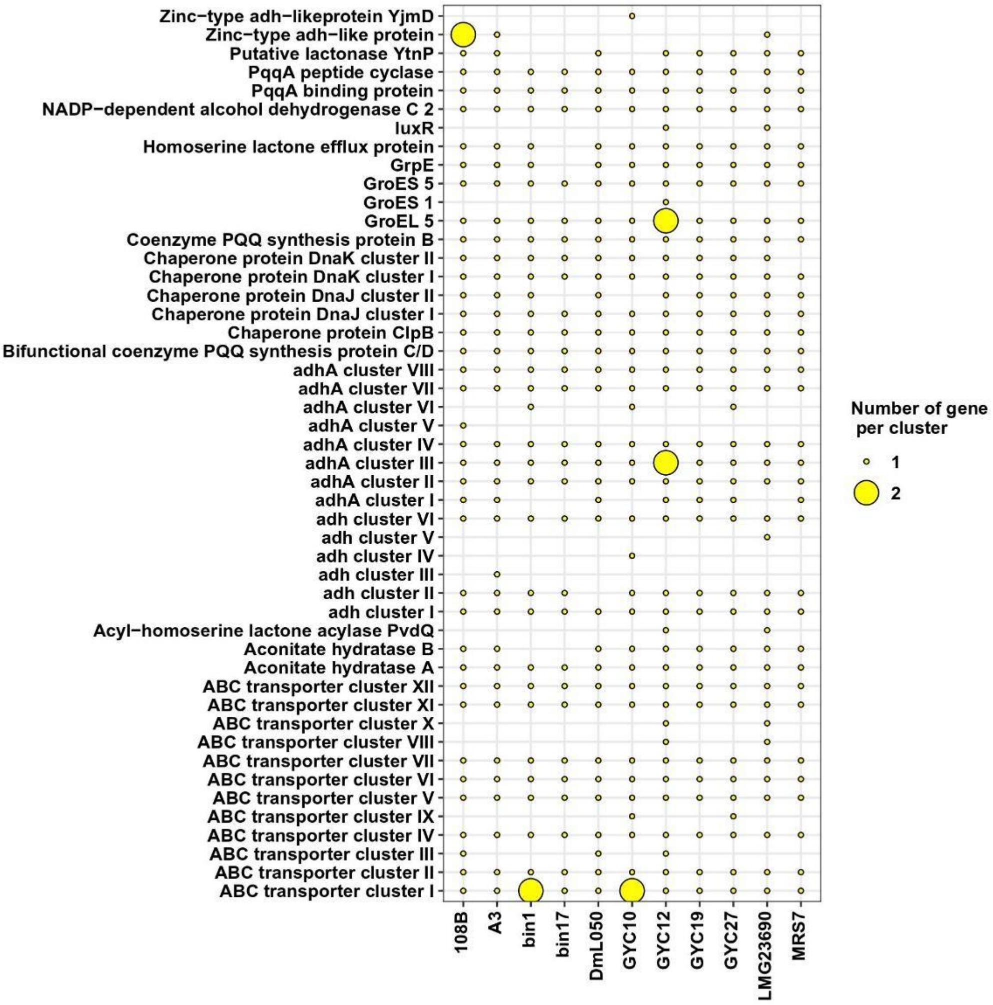
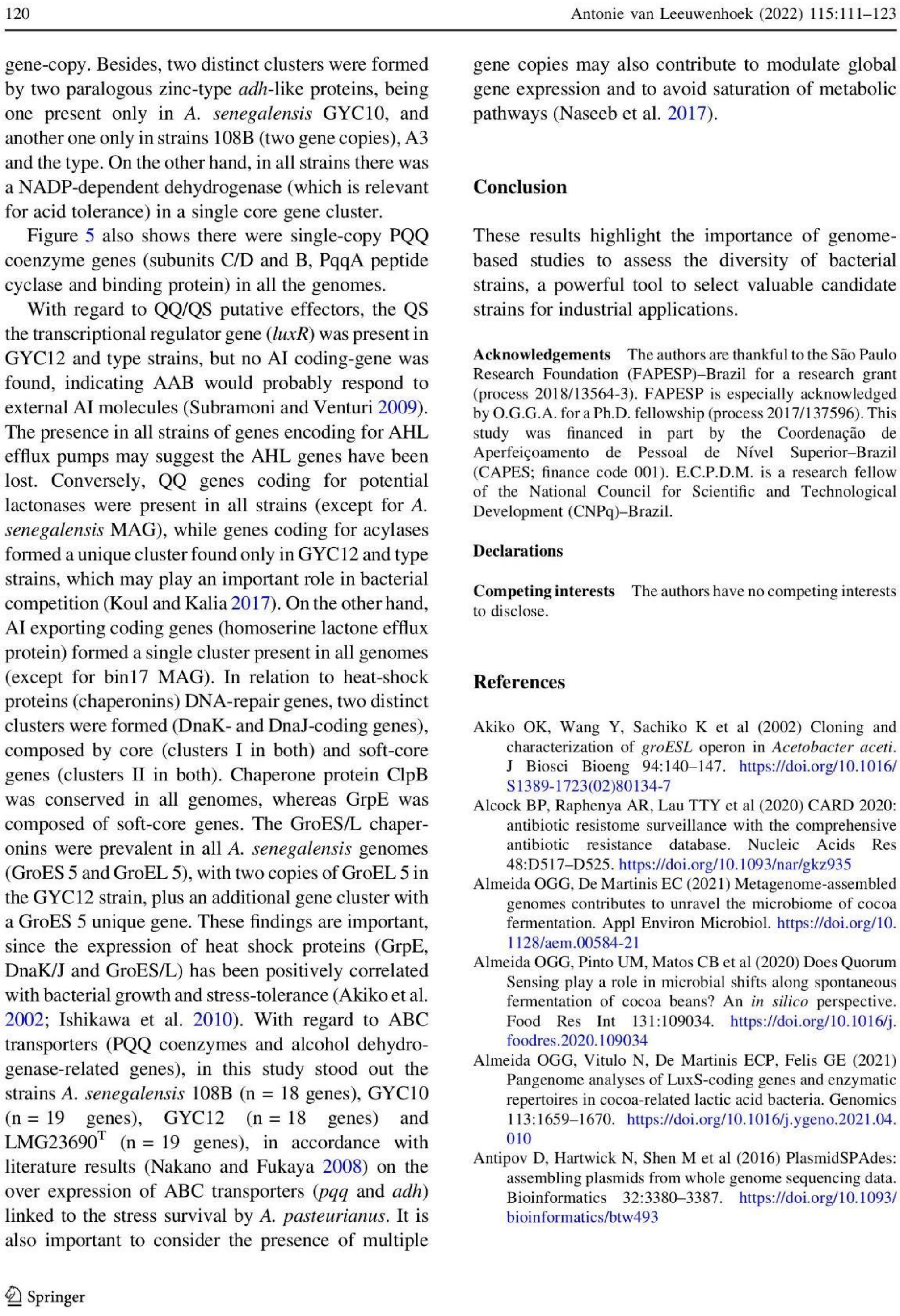


Fig. 5 Classification of core, shell and cloud genes involved with adaptive responses of*A. senegalensis* in harsh environments

annotation were split in different gene clusters, as shown in Fig. 5. A high diversity of ABC transporters were observed, and the clusters I-II, IV-VII and XI-XII were conserved in all the genomes, with a variable number of copies. Aconitases were split in two gene clusters: aconitase hydratase A (100% of *A. senegalensis* strains) and aconitase hydratase B (absent only in the MAGs). Genes coding for alcohol dehydrogenase (*adh*) were divided into five clusters, being the cluster I and VI composed by core and cluster II composed by soft-core genes. The remaining clusters III, IV and V were formed by cloud genes. Another ADH gene family, *adhA* genes, were also spread out in eight gene clusters. The *adhA* gene clusters II-IV and VII-VIII were formed by core genes, and the *A. senegalensis* GYC12 presented two *adhA* copies (gene cluster III). The clusters I and VI were composed by shell genes, while the cluster V was a strain-specific



gene-copy. Besides, two distinct clusters were formed by two paralogous zinc-type *adh*-like proteins, being one present only in *A. senegalensis* GYC10, and another one only in strains 108B (two gene copies), A3 and the type. On the other hand, in all strains there was a NADP-dependent dehydrogenase (which is relevant for acid tolerance) in a single core gene cluster.

Figure 5 also shows there were single-copy PQQ coenzyme genes (subunits C/D and B, PqqA peptide cyclase and binding protein) in all the genomes.

With regard to QQ/QS putative effectors, the QS the transcriptional regulator gene (*luxR*) was present in GYC12 and type strains, but no AI coding-gene was found, indicating AAB would probably respond to external AI molecules (Subramoni and Venturi 2009). The presence in all strains of genes encoding for AHL efflux pumps may suggest the AHL genes have been lost. Conversely, QQ genes coding for potential lactonases were present in all strains (except for *A. senegalensis* MAG), while genes coding for acylases formed a unique cluster found only in GYC12 and type strains, which may play an important role in bacterial competition (Koul and Kalia 2017). On the other hand, AI exporting coding genes (homoserine lactone efflux protein) formed a single cluster present in all genomes (except for bin17 MAG). In relation to heat-shock proteins (chaperonins) DNA-repair genes, two distinct clusters were formed (DnaK- and DnaJ-coding genes), composed by core (clusters I in both) and soft-core genes (clusters II in both). Chaperone protein ClpB was conserved in all genomes, whereas GrpE was composed of soft-core genes. The GroES/L chaperonins were prevalent in all *A. senegalensis* genomes (GroES 5 and GroEL 5), with two copies of GroEL 5 in the GYC12 strain, plus an additional gene cluster with a GroES 5 unique gene. These findings are important, since the expression of heat shock proteins (GrpE, DnaK/J and GroES/L) has been positively correlated with bacterial growth and stress-tolerance (Akiko et al. 2002; Ishikawa et al. 2010). With regard to ABC transporters (PQQ coenzymes and alcohol dehydrogenase-related genes), in this study stood out the strains *A. senegalensis* 108B (n = 18 genes), GYC10 (n = 19 genes), GYC12 (n = 18 genes) and LMG23690^T (n = 19 genes), in accordance with literature results (Nakano and Fukaya 2008) on the over expression of ABC transporters (*pqq* and *adh*) linked to the stress survival by *A. pasteurianus*. It is also important to consider the presence of multiple gene copies may also contribute to modulate global gene expression and to avoid saturation of metabolic pathways (Naseeb et al. 2017).

Conclusion

These results highlight the importance of genome-based studies to assess the diversity of bacterial strains, a powerful tool to select valuable candidate strains for industrial applications.

Acknowledgements The authors are thankful to the São Paulo Research Foundation (FAPESP)–Brazil for a research grant (process 2018/13564-3). FAPESP is especially acknowledged by O.G.G.A. for a Ph.D. fellowship (process 2017/137596). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brazil (CAPES; finance code 001). E.C.P.D.M. is a research fellow of the National Council for Scientific and Technological Development (CNPq)–Brazil.

Declarations

Competing interests The authors have no competing interests to disclose.

References

Akiko OK, Wang Y, Sachiko K et al (2002) Cloning and characterization of *groESL* operon in *Acetobacter aceti*. J Biosci Bioeng 94:140–147. https://doi.org/10.1016/S1389-1723(02)80134-7

Alcock BP, Raphenya AR, Lau TTY et al (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48:D517–D525. https://doi.org/10.1093/nar/gkz935

Almeida OGG, De Martinis EC (2021) Metagenome-assembled genomes contributes to unravel the microbiome of cocoa fermentation. Appl Environ Microbiol. https://doi.org/10.1128/aem.00584-21

Almeida OGG, Pinto UM, Matos CB et al (2020) Does Quorum Sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An *in silico* perspective. Food Res Int 131:109034. https://doi.org/10.1016/j.foodres.2020.109034

Almeida OGG, Vitulo N, De Martinis ECP, Felis GE (2021) Pangenome analyses of LuxS-coding genes and enzymatic repertoires in cocoa-related lactic acid bacteria. Genomics 113:1659–1670. https://doi.org/10.1016/j.ygeno.2021.04.010

Antipov D, Hartwick N, Shen M et al (2016) PlasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 32:3380–3387. https://doi.org/10.1093/bioinformatics/btw493

Aswini K, Gopal NO, Uthandi S (2020) Optimized culture conditions for bacterial cellulose production by *Acetobacter senegalensis* MA1. BMC Biotechnol 20:1–16. https://doi.org/10.1186/s12896-020-00639-6

Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555. https://doi.org/10.1016/j.mimet.2003.08.009

Blin K, Shaw S, Kloosterman AM et al (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:29–35. https://doi.org/10.1093/nar/gkab335

Brettin T, Davis JJ, Disz T et al (2015) RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep. https://doi.org/10.1038/srep08365

Bushnell B (2014) BBTools software package. https://jgi.doe.gov/data-and-tools/bbtools/

Carvalho CC, Phan NN, Chen Y, Reilly PJ (2015) Carbohydrate-binding module tribes. Biopolymers 103:203–214. https://doi.org/10.1002/bip.22584

Chakraborty S, Rani A, Dhillon A, Goyal A (2016) Polysaccharide Lyases. Elsevier, Amsterdam

Chen L, Zheng D, Liu B et al (2016) VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res 44:D694–D697. https://doi.org/10.1093/nar/gkv1239

Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79:7696–7701. https://doi.org/10.1128/AEM.02411-13

De Roos J, De Vuyst L (2018) Acetic acid bacteria in fermented foods and beverages. CurrOpinBiotechnol 49:115–119. https://doi.org/10.1016/j.copbio.2017.08.007

De Roos J, Verce M, Weckx S, De Vuyst L (2020) Temporal Shotgun Metagenomics Revealed the Potential Metabolic Capabilities of Specific Microorganisms During Lambic Beer Production. Front Microbiol. https://doi.org/10.3389/fmicb.2020.01692

Doster E, Lakin SM, Dean CJ et al (2020) MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. Nucleic Acids Res 48:D561–D569. https://doi.org/10.1093/nar/gkz1010

Esquivel-Elizondo S, Ilhan ZE, Garcia-Peña EI, Krajmalnik-Brown R (2017) Insights into Butyrate Production in a Controlled Fermentation System via Gene Predictions. mSystems. https://doi.org/10.1128/msystems.00051-17

Fang G, Rocha EPC, Danchin A (2008) Persistence drives gene clustering in bacterial genomes. BMC Genomics 9:4. https://doi.org/10.1186/1471-2164-9-4

Gao L, Wu X, Xia X, Jin Z (2021) Fine-tuning ethanol oxidation pathway enzymes and cofactor PQQ coordinates the conflict between fitness and acetic acid production by *Acetobacter pasteurianus*. MicrobBiotechnol 14:643–655. https://doi.org/10.1111/1751-7915.13703

Gautreau G, Bazin A, Gachet M et al (2020) PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol 16:1–27. https://doi.org/10.1371/journal.pcbi.1007732

Gomes RJ, Borges MF, Rosa MF et al (2018) Acetic acid bacteria in the food industry: systematics, characteristics and applications. Food Technol Biotechnol 56:139–151. https://doi.org/10.17113/ftb.56.02.18.5593

Grandclément C, Tannières M, Moréra S et al (2015) Quorum quenching: role in nature and applied developments. FEMS Microbiol Rev 40:86–116. https://doi.org/10.1093/femsre/fuv038

Gupta SK, Padmanabhan BR, Diene SM et al (2014) ARG-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 58:212–220. https://doi.org/10.1128/AAC.01310-13

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hammami R, Zouhir A, Ben Hamida J, Fliss I (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiol 7:6–11. https://doi.org/10.1186/1471-2180-7-89

Huang L, Zhang H, Wu P et al (2018) DbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. Nucleic Acids Res 46:D516–D521. https://doi.org/10.1093/nar/gkx894

Iida A, Ohnishi Y, Horinouchi S (2008) Control of acetic acid fermentation by quorum sensing via N-acylhomoserine lactones in *Gluconacetobacter intermedius*. J Bacteriol 190:2546–2555. https://doi.org/10.1128/JB.01698-07

Iida A, Ohnishi Y, Horinouchi S (2009) Identification and characterization of target genes of the GinI/GinR quorum-sensing system in *Gluconacetobacter intermedius*. Microbiology 155:3021–3032. https://doi.org/10.1099/mic.0.028613-0

Ishida T, Sugano Y, Shoda M (2002) Novel glycosyltransferase genes involved in the acetan biosynthesis of *Acetobacter xylinum*. Biochem Biophys Res Commun 295:230–235. https://doi.org/10.1016/S0006-291X(02)00663-0

Ishikawa M, Okamoto-Kainuma A, Jochi T et al (2010) Cloning and characterization of grpE in *Acetobacter pasteurianus* NBRC 3283. J Biosci Bioeng 109:25–31. https://doi.org/10.1016/j.jbiosc.2009.07.008

Jakob F, Quintero Y, Musacchio A et al (2019) Acetic acid bacteria encode two levan sucrase types of different ecological relationship. Environ Microbiol 21:4151–4165. https://doi.org/10.1111/1462-2920.14768

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Koul S, Kalia VC (2017) Multiplicity of quorum quenching enzymes: a potential mechanism to limit quorum sensing bacterial population. Indian J Microbiol 57:100–108. https://doi.org/10.1007/s12088-016-0633-1

La China S, Zanichelli G, De Vero L, Gullo M (2018) Oxidative fermentations and exopolysaccharides production by acetic acid bacteria: a mini review. Biotechnol Lett 40:1289–1302. https://doi.org/10.1007/s10529-018-2591-7

Letunic I, Bork P (2019) Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:256–259. https://doi.org/10.1093/nar/gkz239

Levasseur A, Drula E, Lombard V et al (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels 6:1–14. https://doi.org/10.1186/1754-6834-6-41

Liu LP, Huang LH, Ding XT et al (2019) Identification of quorum-sensing molecules of N-acyl-homoserine lactone in *Gluconacetobacter* strains by liquid chromatography-tandem mass spectrometry. Molecules 24:2694. https://doi.org/10.3390/molecules24152694

Lynch KM, Zannini E, Wilkinson S et al (2019) Physiology of acetic acid bacteria and their role in vinegar and fermented beverages. Compr Rev Food Sci Food Saf 18:587–625. https://doi.org/10.1111/1541-4337.12440

Martínez-Núñez MA, López VEL (2016) Nonribosomal peptides synthetases and their applications in industry. Sustain Chem Process 4:1–8. https://doi.org/10.1186/s40508-016-0057-6

Matsutani M, Hirakawa H, Nishikura M et al (2011) Increased number of Arginine-based salt bridges contributes to the thermotolerance of thermotolerant acetic acid bacteria, *Acetobacter tropicalis* SKU1100. Biochem Biophys Res Commun 409:120–124. https://doi.org/10.1016/j.bbrc.2011.04.126

Nakamura AM, Nascimento AS, Polikarpov I (2017) Structural diversity of carbohydrate esterases. Biotechnol Res Innov 1:35–51. https://doi.org/10.1016/j.biori.2017.02.001

Nakano S, Fukaya M (2008) Analysis of proteins responsive to acetic acid in *Acetobacter*: Molecular mechanisms conferring acetic acid resistance in acetic acid bacteria. Int J Food Microbiol 125:54–59. https://doi.org/10.1016/j.ijfoodmicro.2007.05.015

Naseeb S, Ames RM, Delneri D, Lovell SC (2017) Rapid functional and evolutionary changes follow gene duplication in yeast. Proc R Soc B Biol Sci. https://doi.org/10.1098/rspb.2017.1393

Nayfach S, Shi ZJ, Seshadri R et al (2019) New insights from uncultivated genomes of the global human gut microbiome. Nature 568:505–510. https://doi.org/10.1038/s41586-019-1058-x

Ndoye B, Cleenwerck I, Engelbeen K et al (2007) *Acetobacter senegalensis* sp. nov., a thermotolerant acetic acid bacterium isolated in Senegal (sub-Saharan Africa) from mango fruit (*Mangifera indica* L.). Int J Syst Evol Microbiol 57:1576–1581. https://doi.org/10.1099/ijs.0.64678-0

Ng WL, Bassler BL (2009) Bacterial quorum-sensing network architectures. Annu Rev Genet 43:197–222. https://doi.org/10.1146/annurev-genet-102108-134304

Pal C, Bengtsson-Palme J, Rensing C et al (2014) BacMet: Antibacterial biocide and metal resistance genes database. Nucleic Acids Res 42:737–743. https://doi.org/10.1093/nar/gkt1252

Pelicaen R, Gonze D, Teusink B et al (2019) Genome-scale metabolic reconstruction of *Acetobacter pasteurianus* 386B, a candidate functional starter culture for cocoa bean fermentation. Front Microbiol. https://doi.org/10.3389/fmicb.2019.02801

Pritchard L, Glover RH, Humphris S et al (2016) Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods 8:12–24. https://doi.org/10.1039/c5ay02550h

Prjibelski A, Antipov D, Meleshko D et al (2020) Using SP Ades de novo assembler. Curr Protoc Bioinforma 70:1–29. https://doi.org/10.1002/cpbi.102

Qiu X, Zhang Y, Hong H (2021) Classification of acetic acid bacteria and their acid resistant mechanism. AMB Express 11:29. https://doi.org/10.1186/s13568-021-01189-6

Ruiz-Perez CA, Conrad RE, Konstantinidis KT (2021) MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinform 22:1–16. https://doi.org/10.1186/s12859-020-03940-5

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Seemann T Abricate. https://github.com/tseemann/abricate

Shafiei R, Delvigne F, Babanezhad M, Thonart P (2013) Evaluation of viability and growth of *Acetobacter senegalensis* under different stress conditions. Int J Food Microbiol 163:204–213. https://doi.org/10.1016/j.ijfoodmicro.2013.03.011

Shafiei R, Zarmehrkhorshid R, Mounir M et al (2017) Influence of carbon sources on the viability and resuscitation of *Acetobacter senegalensis* during high-temperature gluconic acid fermentation. Bioprocess Biosyst Eng 40:769–780. https://doi.org/10.1007/s00449-017-1742-x

Shafiei R, Leprince P, Sombolestani AS et al (2019) Effect of sequential acclimation to various carbon sources on the proteome of *Acetobacter senegalensis* LMG 23690^{T} and its tolerance to downstream process stresses. Front Microbiol 10:1–14. https://doi.org/10.3389/fmicb.2019.00608

Siedenburg G, Jendrossek D (2011) Squalene-hopene cyclases. Appl Environ Microbiol 77:3905–3915. https://doi.org/10.1128/AEM.00300-11

Simão FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Son H-J, Heo M-S, Kim Y-G, Lee S-J (2001) Optimization of fermentation conditions for the production of bacterial cellulose by a newly isolated *Acetobacter* sp.A9 in shaking cultures. Biotechnol Appl Biochem 33:1. https://doi.org/10.1042/ba20000065

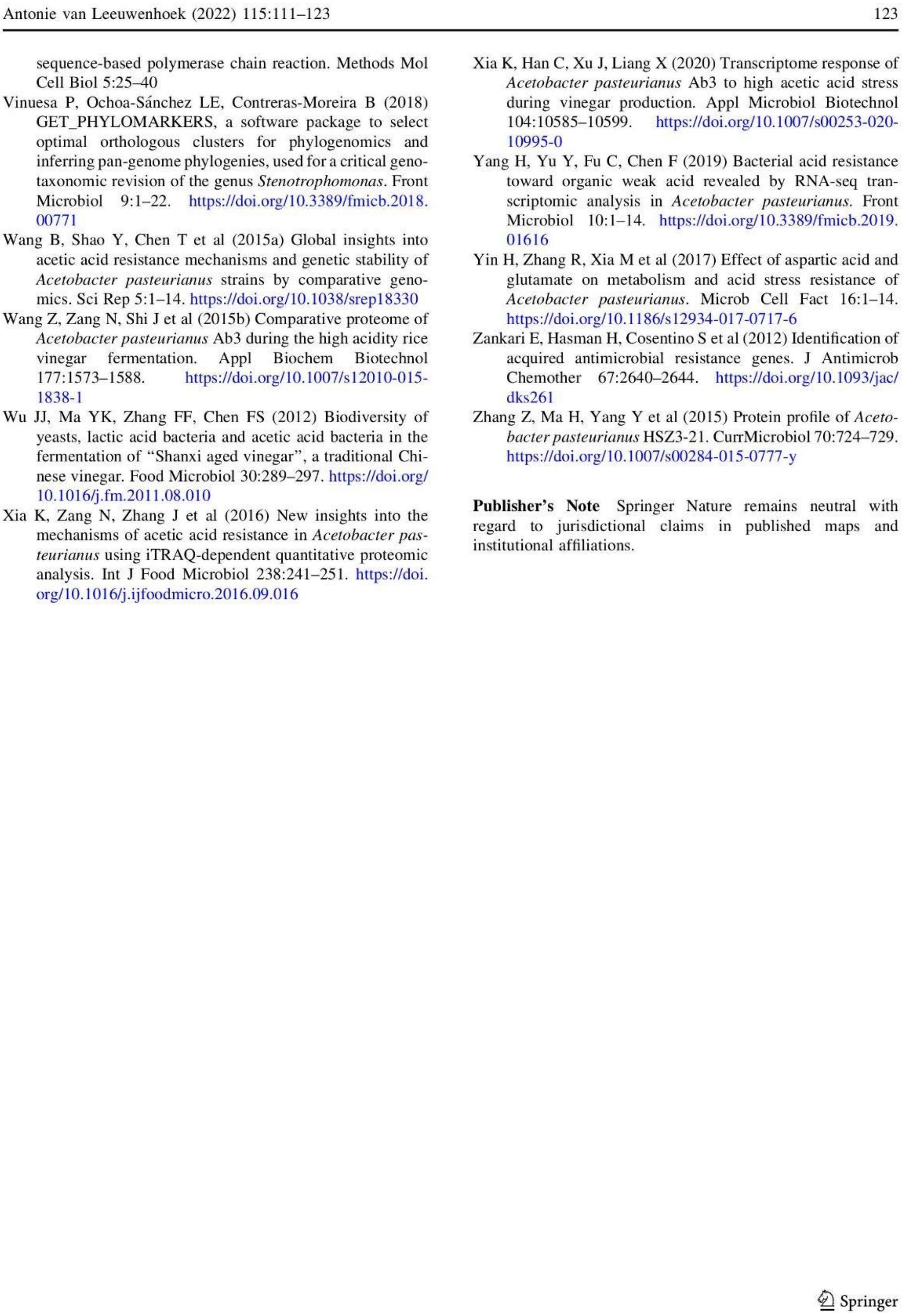
Subramoni S, Venturi V (2009) LuxR-family "solos": Bachelor sensors/regulators of signalling molecules. Microbiology 155:1377–1385. https://doi.org/10.1099/mic.0.026849-0

Tran TN, Phong HX, ThaoAnh BT et al (2018) High-temperature production of acerola vinegar using thermotolerant *Acetobacter senegalensis* A28. Vietnam J Sci Technol Eng 60:13–18. https://doi.org/10.31276/vjste.60(3).13

Trček J, Dogsa I, Accetto T, Stopar D (2021) Acetan and acetan-like polysaccharides: genetics, biosynthesis, structure, and viscoelasticity. Polymers (basel) 13:1–16. https://doi.org/10.3390/polym13050815

Valera MJ, Mas A, Streit WR, Mateo E (2016) GqqA, a novel protein in *Komagataeibacter europaeus* involved in bacterial quorum quenching and cellulose formation. Microb Cell Fact 15:1–15. https://doi.org/10.1186/s12934-016-0482-y

Versalovic J, Schneider M, De Bruijn FJ, Lupski J (1994) Genomic fingerprinting of bacteria using repetitive



sequence-based polymerase chain reaction. Methods Mol Cell Biol 5:25–40

Vinuesa P, Ochoa-Sánchez LE, Contreras-Moreira B (2018) GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus *Stenotrophomonas*. Front Microbiol 9:1–22. https://doi.org/10.3389/fmicb.2018.00771

Wang B, Shao Y, Chen T et al (2015a) Global insights into acetic acid resistance mechanisms and genetic stability of *Acetobacter pasteurianus* strains by comparative genomics. Sci Rep 5:1–14. https://doi.org/10.1038/srep18330

Wang Z, Zang N, Shi J et al (2015b) Comparative proteome of *Acetobacter pasteurianus* Ab3 during the high acidity rice vinegar fermentation. Appl Biochem Biotechnol 177:1573–1588. https://doi.org/10.1007/s12010-015-1838-1

Wu JJ, Ma YK, Zhang FF, Chen FS (2012) Biodiversity of yeasts, lactic acid bacteria and acetic acid bacteria in the fermentation of "Shanxi aged vinegar", a traditional Chinese vinegar. Food Microbiol 30:289–297. https://doi.org/10.1016/j.fm.2011.08.010

Xia K, Zang N, Zhang J et al (2016) New insights into the mechanisms of acetic acid resistance in *Acetobacter pasteurianus* using iTRAQ-dependent quantitative proteomic analysis. Int J Food Microbiol 238:241–251. https://doi.org/10.1016/j.ijfoodmicro.2016.09.016

Xia K, Han C, Xu J, Liang X (2020) Transcriptome response of *Acetobacter pasteurianus* Ab3 to high acetic acid stress during vinegar production. Appl Microbiol Biotechnol 104:10585–10599. https://doi.org/10.1007/s00253-020-10995-0

Yang H, Yu Y, Fu C, Chen F (2019) Bacterial acid resistance toward organic weak acid revealed by RNA-seq transcriptomic analysis in *Acetobacter pasteurianus*. Front Microbiol 10:1–14. https://doi.org/10.3389/fmicb.2019.01616

Yin H, Zhang R, Xia M et al (2017) Effect of aspartic acid and glutamate on metabolism and acid stress resistance of *Acetobacter pasteurianus*. Microb Cell Fact 16:1–14. https://doi.org/10.1186/s12934-017-0717-6

Zankari E, Hasman H, Cosentino S et al (2012) Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261

Zhang Z, Ma H, Yang Y et al (2015) Protein profile of *Acetobacter pasteurianus* HSZ3-21. CurrMicrobiol 70:724–729. https://doi.org/10.1007/s00284-015-0777-y

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 4

6. Chapter 4 - Investigating interspecific quorum sensing influence on cocoa fermentation quality through defined microbial cocktails

Specific Objectives

- To survey lab scale fermentation inoculated with defined cocktails of yeasts, LAB, and AAB strains in order to evaluate the expression pattern of the *luxS* along fermentations and to compare to a control fermentation (with no cocktail inoculation). Besides, to measure enzymatic activity of 16 enzyme families of amilases, β-glucanases, cellulases, arabinofuranosidases, invertases, endoglucanases, xylanases, β-glucosidases, arabinases, xyloglucanases, esterases, β-xylosidases, pectinases, mannanases, cellobiohydrolases, and lipases to visualize enzyme shifts during fermentation and its relationship with the total microbiota determined by metataxonomics. Finally, to detect VOCs related to cocoa beans' fermentation quality and to correlate sensorial profiles with the *luxS* gene expression, aiming to dissect whether the influence of QS on cocoa fermentation quality.

The manuscript was submitted to the *Food Science International* Journal, which allows manuscript's inclusion in the Thesis before and pre-print publication official publication on the journal, as explained in this link. The manuscript was made available online as pre-print in: https://doi.org/10.1101/2022.06.14.496151.

Investigating interspecific quorum sensing influence on cocoa fermentation quality through defined microbial cocktails

Almeida, O.G.G[1].; Pereira, M. G.[2]; Bighetti-Trevisan, R. L[3].; Santos, E[1].S.; De Campos, E. G[4,7].; Felis, G.E[5].; Guimarães, L.H.S[6].; Polizeli, M.L.T.M[6]; De Martinis, B. S[7].; De Martinis, E.C.P[1]*.

[1] Universidade de São Paulo, Faculdade de Ciências Farmacêuticas de Ribeirão Preto. Departamento de Análises Clínicas, Toxicológicas e Bromatológicas.

[2]Universidade do Estado de Minas Gerais, Unidade Passos.

[3]Universidade de São Paulo, Faculdade de Odontologia de Ribeirão Preto. Departamento de Biologia Básica e Oral.

[4]Appalachian State University, Department of Chemistry and Fermentation Sciences, Boone, NC, United States

[5]University of Verona, Department of Biotechnology, Verona, Italy

[6]Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto. Departamento de Biologia.

[7]Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto. Departamento de Química.

*Corresponding author: Universidade de São Paulo, Faculdade de Ciências Farmacêuticas de Ribeirão Preto. Departamento de Análises Clínicas, Toxicológicas e Bromatológicas. Address: Avenida do Café, s/n – Campus da USP. Bairro Monte Alegre. Ribeirão Preto – SP. Postal Code: 14040-903. Email: edemarti@usp.br.

Abstract

The fermentation of cocoa beans is a key process to supply high quality ingredients for the chocolate industry. In spite of several attempts to obtain standardised microbial cultures for cocoa fermentation, it is still a spontaneous process. It has been suggested lactobacilli present potential for quorum sensing (QS) regulation in cocoa fermentation, and in the present research, laboratory scale fermentations were carried out to further elucidate possible QS influence on microbial shifts and fermented seeds quality. The experimental design comprised the 96 hours-fermentations designated as F0 (control), F1 (yeasts, lactic acid bacteria, and acetic acid bacteria), F2 (yeasts and acetic acid bacteria), F3 (yeasts only), with evaluation of the microbial succession by plate counting, determination of enzymatic activities by classical methods and qualitative evaluation of flavour compounds by gas-chromatography (GC-MS) with headspace sampling. Besides, QS was estimated by quantification of the expression of *luxS* genes by Reverse Transcriptase Real Time PCR analysis using selected primers. The results demonstrated that microbial successions were displayed in lab conditions, but no statistical difference in terms of microbial enumeration and α-diversity metrics were observed among the experimental and control fermentations. Moreover, enzymatic activities were not correlated to the total microbiota, indicating the seeds' endogenous hydrolases protagonist enzymes secretion and activity. Regarding *luxS* genes measuring for the species *Lactiplantibacillus plantarum* and *Limosilactobacillus fermentum*, genes were active in fermentation in the start to the end phase and to the beginning to the middle phase of fermentation, respectively. Correlation analysis among *luxS* expression and volatile metabolites evidenced *Lp. plantarum* association with detrimental compounds for fermentation quality. This data contributes to our previous research which monitored fermentations to survey enzymatic changes and QS potential along the process and sheds light of QS-related strategies of lactobacilli dominance in cocoa fermentations.

Key words: cocoa fermentation, quorum sensing, starter cultures, *luxS* gene.

Introduction

Cocoa fermentation is a spontaneous process characterised by a succession of yeasts, lactic acid bacteria (LAB) and acetic acid bacteria (AAB), which are responsible for generating desirable sensory characteristics in the raw material for chocolate production, encompassing colour, aroma, flavour, and texture (De Vuyst & Weckx, 2016). However, such biological process is complex, as the microbial composition may vary geographically and according to the fermentation methods and fermentative processes used (Bortolini et al., 2016). Although many studies indicated *Lactiplantibacillus plantarum* and *Limosilactobacilus fermentum* species dominate in several fermentations carried out in diverse geographic regions, the determination of a core microbiome for cocoa fermentation is challenging (Viesser et al., 2021). In Brazil, for instance, it was observed a superior diversity of LAB and AAB in comparison to Ghana which exhibited a varied repertoire of LAB instead. On the other hand, other producing regions such Nicaragua, Colombia, Cameroon, and Ivory Coast showed lower bacterial diversities due to the paucity of NGS data (Viesser et al., 2021).

Besides, little is still known about the intraspecific variation among the strains of cocoa dominant species for the selection of those with metabolic repertoires of interest in order to standardise the process (Ouattara & Niamké, 2021). The search for a starter culture is not a novelty, as so many works have proposed a multitude of candidate strains with interesting metabolic traits (Farrera et al., 2021; Magalhães da Veiga Moreira et al., 2017; Ooi et al., 2020; Saunshi et al., 2020; Visintin et al., 2017). However, the criteria for the selection of these microorganisms are yet a subject of much discussion and there is no guideline for this, since they are based, mostly, on the correlation of the sensory profiles generated after the inoculation of strains in fermentations, inhibition of detrimental and pathogenic microbiota and mobilisation of pulp components (Chagas Junior et al., 2021).

Among the characteristics of a well-defined starter culture is its potential to tolerate adverse conditions and to compete with undesirable microorganisms, replacing them and assuring safety. Microorganisms must overcome stressors by the expression of heat shock proteins, production of exopolysaccharides, releasing of antimicrobial compounds and harbouring specific metabolic pathways that assure their permanency in an environment (De Roos & De Vuyst, 2018; García-Ríos et al., 2021; Ogunremi et al., 2022; Romanens et al., 2019). The robustness of a microbial species resides in all these characteristics gathered in the fermentation.

Recent studies relate quorum sensing (QS) to the adaptation of bacteria to stressful conditions. QS is a phenomenon associated to the synchronisation of gene expression in bacteria in response to the density of bacterial populations leading to a multitude of downstream metabolic responses such as phenotypic modulation and physiological changes to maintain homeostasis (Johansen & Jespersen, 2017; Wu et al., 2020; Yang et al., 2021). The *luxS* gene is the universal gene marker for interspecific quorum sensing as it is responsible for the conversion of 4,5-dihydroxy-2,3-pentanedione

(DPD) molecule to the autoinducer 2 (AI-2) that will sensitise adjacent cells in the environment, tiggering QS. The AI-2 is produced in the methyl-cycle from the S-adenosylmethionine (SAM) precursor molecule. The pathway starts by the transferring of a methyl group from SAM to methyl-transferases and substrates producing S-adenosylhomocysteine (SAH). Then, SAH has an adenine removed by a nucleosidase (pfs) enzyme leading to the production of a S-ribosyl homocysteine (SRH), whose reaction is catalysed by the LuxS enzyme. The SRH is finally converted into homocysteine and DPD. As this last is an unstable molecule, it can spontaneously cyclize into several DPD derivatives and sensibilize several bacteria (Wu et al., 2020; Yang et al., 2021). As this gene is widespread in the bacterial kingdom as it is part of the activated methyl-cycle, it could be a clue of a universal interspecies communicator (Tobias et al., 2020), being and excellent marker to assess bacterial communication by *in silico* methods.

In this sense, a metagenomic study of spontaneous cocoa fermentation carried out by our research group showed that lactic acid bacteria, with emphasis on lactobacilli, were correlated to higher amounts of *luxS*-derivate gene reads along cocoa fermentation (Almeida et al., 2020). That study showed that as the lactobacilli load increased, the *luxS* gene counts were also increased. Another study, carried out on kimchi's fermentation, showed that the most dominant LAB bacteria showed the expression of *luxS*, while non-dominant bacteria did not (Park et al., 2016). Other studies also demonstrate that the expression of the *luxS* gene in LAB can result in adaptive benefits in the face of unfavourable conditions related to low pH, high temperatures and nutrient depletion (Gu et al., 2018; Jiang et al., 2021).

Thus, in view of starter cultures drawing QS characterization of strains, accompanied with sensorial screening and other metrics already used in literature, could be a valuable tool to select robust candidates in the aim of maintaining products identity balanced with standardisation. On the other hand, the information relating QS to the quality of fermented foods is scarce and must be evaluated in depth. In this context, this work aims to determine through lab-scale fermentation the influence of interspecific QS on the dominance of LAB in cocoa fermentation and if it influences the quality of fermented seeds by the use of defined yeasts, LAB, and AAB cocktails. To test this hypothesis, production of volatile metabolites, enzymatic activity changes, and microbial composition were evaluated by chemistry based and NGS methods. Besides, *luxS* gene expression was monitored, quantified along the fermentations, and analysed by correlation and regression to answer the following question: "Does quorum sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans?".

Materials and methods

Selected strains

All the strains used in this study were obtained from a spontaneous cocoa fermentation carried out in Bahia state, Brazil (Almeida et al. 2020). The fungal strains selected were *Pichia kudriavzevii* strain PCA1, *Pichia kluyveri* strain PCA4 and four *Saccharomyces cerevisiae* with distinct genotypes as determined by the microsatellite amplification (Vaudano & Garcia-Moruno, 2008), strains F3, F6, F11 and F12. These strains were selected since *Pichia* and *Saccharomyces* species are often reported in cocoa fermentation.

Regarding LAB, the strains *Lactiplantibacillus plantarum* Lb2 and *Limosilactobacillus fermentum* Lb1 were selected as they were characterised in terms of potential interspecific QS by genomic-centred analysis (Almeida et al., 2021). Finally, five genotypically diverse *Acetobacter senegalensis* strains named MRS7, GYC10, GYC12, GYC19 and GYC27 were selected due to their genetic potential to adapt in harsh fermentative environments (Almeida et al., 2022).

Inoculum preparation

The yeasts were cultured on a 50 ml of yeast extract (1%), peptone (2%) and dextrose (2%) medium (YPD) and incubated for 24h at 30ºC under shaker agitation at 110 rpm. LAB were cultured on a 50 ml of De Man, Rogosa & Sharpe (MRS, Oxoid, Basingstoke, UK) broth, incubated at 30ºC for three days in anaerobic jars. AAB were cultured on a 50 ml of a modified broth of glucose (5%) and yeast extract (1%) (GYC medium) and incubated for 24h at 30ºC under shaker agitation at 110 rpm. Then, the strains were enumerated to standardise the inoculum load on fermentations. A concentration of $1x10^6$ UFC/g was determined using a growth curve as the optimal concentration for each yeast and bacterial species selected. The strains were recovered from their culture media by centrifugation for 15 minutes at 7,500g. Then, each strain's pellet was resuspended in saline solution and centrifuged for 15 minutes at 7,500g twice. Dedicated volumes were added in sterile Boeco® flasks to compose the cocktails. The yeast, LAB and AAB cocktails totalized a volume of 300 ml, 200 ml and 100ml, respectively.

Fermentation experimental design

Cocoa fruits (mix of cultivars Criolo and Forastero) were obtained from "Companhia de Entrepostos e Armazéns Gerais de São Paulo - CEAGESP" distributor located at Ribeirão Preto city (Brazil) and were used in all fermentations. To carry out the fermentation, the fruits were opened with a sterile stainless-steel knife to perform the peeling. A total of 5 kg of fresh-harvested cocoa seeds was placed in each box measuring 40 cm long, 12 cm high and 29 cm width which were partially covered with the lid along the entire fermentation. This was considered the start time of fermentation (time 0h). After the first 48 h, the seeds were tuned for aeration, being this process repeated every 24h as performed in field conditions.

The experimental design was composed by a control fermentation, named F0, which was composed only by cocoa seeds without replicates, and three main fermentations performed in biological duplicates, named: F1, F2 and F3. The experimental design was depicted in Fig. 1. The C1 cocktail was prepared in a volume of 300 ml of *Pichia kudriavzevii* PCA1, *Pichia kluyveri* PCA4, and *Saccharomyces cerevisiae* strains F3, F6, F11, and F12 in equimolar concentrations of 1x10^6 yeasts/g of each species. The C2 cocktail was prepared in a volume of 200 ml of *Lactiplantibacillus plantarum* Lb2 and *Limosilactobacillus fermentum* Lb1 in equimolar concentrations of 1x10^6 bacteria/g of seeds of each species. The C3 cocktail was prepared in a volume of 100 ml of *Acetobacter senegalensis* strains MRS7, GYC10, GYC12, GYC19 and GYC27 in concentrations of 1x10^6 bacteria/g of each species.

In the F1 fermentation, all cocktails C1, C2, and C3 were added at the times 0h, 24h, and 72h, respectively. In the fermentation F2, the C2 cocktail was not inoculated, but only the C1 and C3 cocktails were added at the times 0h and 72h, respectively. Finally, F3 fermentation was inoculated only with the C1 cocktail at the time 0h. Boxes temperature and pH were measured continuously using a digital thermohygrometer (Kasvi, Paraná, Brazil) and a pHmeter (Quimis, São Paulo, Brazil).

Sampling and enumeration of microorganisms from lab-scale cocoa fermentations

During the fermentations the samples were collected aseptically in intervals of 0h, 24h, 48h, 72h and 96h for enumeration of presumptive yeasts, LAB, and AAB. For this, selective media for these microorganisms were used: Amphotericin B (Sigma-Aldrich, Massachusetts, USA) (5mg/l) in MRS (Oxoid) and GYC culture media and chloramphenicol (Sigma-Aldrich) (200 mg/l) in YPD. 225 ml of 0.1% peptone water was added to 25 g of each sample and then it was manually homogenised for 2 min. Then, serial dilutions were performed in 0.1% peptone water. After the process, the samples were plated in three different types of culture media (GYC, MRS (Oxoid) and YPD), containing antibiotic and/or antifungal. The plates were incubated for 48 h and then counting was performed.

Crude extract preparation and enzymatic activities quantification

To determine the total protein content and enzymatic activity catalysed by cocoa microbiota, 17 g of seeds were homogenised in 30 ml of distilled water by vortex for complete homogenization of the pulp. This mixture was placed in 50 ml Falcon® tubes which were kept refrigerated at 4°C up to 8°C. Enzyme activities were determined using the direct substrates for the enzyme. One unit (U) was defined as the amount of enzyme necessary to hydrolyze 1 μmol of substrate per minute under the previously described conditions.

Lipase activity was determined using *p*-nitrophenylpalmitate (Sigma-Aldrich) as substrate (Pencreac'h & Baratti, 1996). Pectinase activity by the 3',5'-dinitrosalicylic acid (DNS) (Sigma-Aldrich) method (Miller, 1959), using 1% galacturonic acid solution dissolved in 50 mM sodium acetate buffer, pH 5.0. Amylase activity was determined using the reaction by the DNS method (Miller, 1959), using 1% starch solution dissolved in 50 mM sodium acetate buffer, pH 5.0. Arabinase activity was dosed using an unbranched arabinan as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Arabinofuranosidase activity was detected using the batch method (Kersters-Hilderson et al., 1982), using the synthetic substrate and *p*-nitrophenyl-α-L-arabinofuranoside (Sigma-Aldrich) (PNP-ara). Cellulase activity was performed with the substrate avicel 1% (w/v) through the formation of reducing sugars by the DNS method (Miller, 1959). Invertase activity was measured using sucrose as substrate 1% (w/v) through the formation of reducing sugars by the DNS method (Miller, 1959). β-glucanase activity was measured using β-glucan syrup 0.5% (w/v) as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Endoglucanase activity was measured using Carboximetil celulose (Sigma-Aldrich) (CMC) 1% (w/v) as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Xyloglucanase activity was measured using xyloglucan 1% (w/v) as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Mannanase activity was measured using “Locust bean gum” 1% (w/v) as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Xylanase activity was measured using xylan Beechwood (Sigma-Aldrich) 1% (w/v) as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Esterase activity was measured using *p*-nitrophenyl-acetate (PNP-acetate) (Sigma-Aldrich) 1% (w/v) as substrate through the formation of reducing sugars by the DNS method (Miller, 1959). Cellobiohydrolase activity was measured using the synthetic substrate *p*-nitrophenyl-cellobioside (Sigma-Aldrich) (PNP-cellobioside). β-glucosidase activity was measured using the synthetic substrate *p*-nitrophenyl-glucopyranoside (Sigma-Aldrich) (PNP-Glu). β-xylosidase activity was measured using the synthetic substrate *p*-nitrophenyl-xylopyranoside (Sigma-Aldrich) (PNP-xylo).

Protein amounts were determined using the Bradford (1976) method and a bovine serum albumin curve as a standard reference.

Volatile compounds identification

Prior to analysis, frozen cocoa samples (pulp and seed) were removed from the freezer, weighted (1.00 to 1.05 g) and inserted into a headspace vial. Samples were manually subjected to incubation in a dry block heater for 30 minutes, at 90ºC. After the incubation, 0.5 mL of the vapour phase was collected with a gas tight syringe and manually injected into the GC-MS for analysis.

Analyses were performed using an Agilent (Santa Clara, CA, United States) 7890A GC coupled to an Agilent 5975C MS (Santa Clara, CA, United States). A capillary column HP-5MS (30 m X 0.25 mm, 0.25 mm) was used for the chromatographic separation. The method used in these analyses was based on previously published methods available in the literature (Instituto Adolfo Lutz, 2008; Tait et al., 2014). The separation was performed using a temperature program as follows: 40°C for 5 min, increased to 220°C at 8°C/min e isothermal for 2 min. Injection was performed at

250°C, in split mode, with a split ratio of 1:10. Helium was used as carrier gas, at a flow rate of 1 ml/min. The mass spectrometer operated in Full Scan mode (m/z 50 – 650). The temperatures of the MS source and quadrupole were 230 and 150°C, respectively.

Data obtained from the analyses were performed using the AMDIS from NIST (Version 2.66, August 2008), available within the GC-MS software. The following settings were adopted for data analysis and processing using the AMDIS software: (a) type of analysis: simple; (b) minimum match factor set at 60 and (c) parameters of resolution, sensitivity and shape requirements were used as the default setting (Medium). Library used for the searches was the NIST Mass Spectral Library (Version 2.0, October 2009). The criteria for the search in the NIST library was a match of 700 or higher (NIST, 2008). Only metabolites detected in both replicates were summarised and reported.

Metagenomic DNA extraction and sequencing

Metagenomic DNA extraction was performed according to the protocol described by our research group in the reference (Almeida et al., 2020). In summary, three random seeds were selected, and their pulps were scraped. Then, DNA extraction was performed according to the recommendations of the manufacturer of the Zymobiomics DNA MINI KIT (California, USA). The extracted DNA was quantified by Qubit fluorometer (ThermoFisher) and nanodrop (ThermoFisher) and sent to the "BPI Biotecnologia EPP" facility for 16S *rRNA* V3-V4 region (forward primer: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG3'; reverse primer: 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAA TCC3') and *ITS* region amplification (primer 86F: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGAATCATCGAATCTTTGAA 3'; 4R: 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCTCCGCTTATTGATATGC 3') and sequencing on the MiniSeq platform (Illumina®) to generate paired-end (PE) 2x250 bp reads following the facility's protocols. Raw sequencing data was made publicly available on NCBI Bioprojects PRJNA842267 and PRJNA842340.

Total RNA extraction for qRT-PCR

Total RNA extraction was performed according to the protocol described by Verce et al. (2021), with modifications. Briefly, three cocoa beans were homogenised in 30 ml of RNAprotect (Qiagen, Venio, The Netherlands) in a 50 ml falcon and vortexed for 30 seconds to disrupt the mucilaginous pulp from the seeds. Then, the seeds were removed with the aid of sterile forceps. Then, the tubes were centrifuged at 6,000g for 30 minutes. Subsequently, the supernatant was discarded, and 2 ml of sorbitol wash buffer (1.5 M sorbitol, 50 mM Tris-base and nuclease-free water) were added, followed by vigorous vortex, and centrifugation for 10 minutes at 6,000g. The supernatant was discarded and 3.5 ml of RLT buffer with B-mercaptoethanol (10 µl for every 350 µl of RLT buffer – Qiagen) was added. The tubes were vortexed for 20 seconds and then centrifuged for three minutes at maximum speed (about 10,000g). Then, the supernatant was eluted on the RNeasy Mini Kit (Qiagen) gDNA strip elution column until the entire contents of each tube were exhausted. Then, 2.0 ml of 70% alcohol was added to the eluted volume, which was gently homogenised with the pipette. Finally, the remaining steps were performed according to the RNeasy Mini Kit manufacturer's instructions for RNA obtaining. The integrity of the extracted RNA was evaluated in a 1% (m/v) agarose gel and the material was quantified using the nanodrop (ThermoFisher, Massachusetts, USA).

luxS gene quantification by qRT-PCR

For relative quantification by qRT-PCR, the *luxS* genes of *Lm*. *fermentum* and *Lp. plantarum* species were amplified in parallel with the respective 16S *rRNA* genes for gene expression normalisation. The chosen primers were already described in literature (Almeida et al., 2021; Gu et al., 2018; Schwendimann et al., 2015) and are presented in the Table 1. The cDNA preparation was performed following the instructions of the manufacturer of the ProtoScript II Reverse Transcriptase (New England Biolabs, Massachusetts, USA). Random primers (ThermoFisher) were used to initiate the reverse transcription reaction. Quantification of relative *luxS* gene expression was performed by real-time qPCR using the qPCRBIO SyGreen Mix (PCR Biosystems, Pennsylvania, USA), according to the manufacturer’s instructions. The reactions were carried out in triplicate in the Realplex^{4} Epgradient Mastercycler (Eppendorf, Hamburg, Germany). The $2^{-\Delta\Delta Ct}$ method (Livak & Schmittgen, 2001) was applied to calculate the relative expression levels of *luxS* genes, using the 16S *rRNA* Ct values as normalizer.

Bioinformatic analysis of microbial composition

The high-quality forward reads (R1) were selected for downstream analysis on Qiime2 environment (Bolyen et al., 2019). Initially, the delivered raw reads were assessed in terms of quality using the bbduk tools (Bushnell, 2014) to remove adapters/barcodes fragments and reads with phred scores lower than 30 (parameters: hdist=1 tpe tbo qtrim=rl trimq=30 maq=30). Then, the reads were imported and processed with the dada2 (Callahan et al., 2016) plugin for chimera removal and de-replication. The taxonomy was assigned by training the full 16S *rRNA* region of the 99% SSU NR database obtained from the SILVA repository (Quast et al., 2013). The same initial steps were followed for fungi taxonomic assignment, but the database chosen was the UNITE 99% NR version 10.05.2021 (Nilsson et al., 2019).

The processed ASVs (amplicon sequence variants) data was imported to the RStudio environment using the R package qiime2R (Jordan & Bisanz, 2018). Data visualisation (i.e., alpha-diversity, beta-diversity, and composition) was generated using dedicated microbiome packages such as phyloseq (McMurdie & Holmes, 2013), qiime2R

(Jordan & Bisanz, 2018), and microbiomeutilities (Shetty & Lahti, 2020). Statistical analysis was performed using the Wilcox test to compare the means of fermentation alpha-diversities.

Monitoring of inoculated starters in fermentation

To confirm the starters added in the defined times were present during fermentations (F1, F2 and F3) the *16S rRNA* and *ITS* genes sequenced by the Sanger method were used as reference in a blast analysis to correlate the ASVs *16S rRNA* reads with their cognate sequence. Yeasts DNA was extracted from two-days incubated cultures grew on YPD medium. From cultures, an aliquot of 2.0 ml was selected, and the DNA was obtained using the Wizard Genomic DNA Purification Kit (Promega, Milan, Italy), following the manufacturer's instructions. DNA concentration was determined using a Nanodrop®. *ITS* 5.8S *rRNA* gene was amplified using the primers ITS1 (TCCGTAGGTGAACCTGCGG) and ITS4 (TCCTCCGCTTATTGATATGC) (White, 1990). The PCR reaction was performed as described by Esteve-Zarzoso et al. (1999). Finally, PCR products were purified using the kit Gene elute^{TM} Gel extraction kit (Sigma Aldrich), re-surveyed on Nanodrop® and delivered for Sanger DNA sequenced at Eurofins Genomics (Ebersberg, Germany).

LAB DNA was extracted from pure cultures grew overnight. From the culture, an aliquot of 1.0 ml was selected and centrifuged at 14,000 rpm at 4ºC for 5 min. The underneath was discarded, and the pellet was pre-treated for disruption of the bacterial cell wall with 300 µl of lysozyme solution in Tris-EDTA buffer (TE) (10 mg/ml, Sigma-Aldrich/Merck), incubated at 37ºC for 60 min. The resulting suspension was then centrifuged at 14,000 rpm at 4ºC to remove cellular fragments and the genomic DNA finally was extracted using the Wizard® Genomic DNA Purification Kit reference A1125 (Promega, Madison, WI, USA) following the bulla guidelines. The pure DNA was quality-assessed and quantified using Nanodrop® (Thermo Fisher Scientific, USA). Full *16S rRNA* gene was amplified using the primers E8F (5'AGAGTTTGATCCTGGCTCAG3') and E1541R (5'AAGGAGGTGATCCANCCRCA3') (Baker et al., 2003). PCR was performed as described by Rathnayake et al. (2010). The PCR products were purified using the kit Gene elute^{TM} Gel extraction kit (Sigma Aldrich, MA, USA), re-surveyed on Nanodrop® and delivered for Sanger DNA sequenced at Eurofins Genomics (Ebersberg, Germany).

A. senegalensis strains' DNA was extracted from overnight cultures on BHI medium (OXOID, Basingstoke, UK) from a 1,0 ml aliquot of suspension was selected for DNA extraction. The DNA was extracted using the kit Illustrative Bacteria Genomic Mini Spin Kit (GE Life Sciences, Switzerland) according to manufacturer's recommendations. The pure DNA was quality-assessed and quantified using Nanodrop® (Thermo Fisher Scientific, USA). Full *16S rRNA* was amplified using the primers 27F (5'AGAGTTTGATCMTGGCTCAG3') and (5'GCTTACCTTGTTACGACTT3') according to the protocol of Tulini et al. (2016). The amplicons were visualised on an agarose gel (0.8%) electrophoresis, and amplicons were purified using the Gel Band Purification kit (GE Life Sciences), re-surveyed on Nanodrop® and delivered to DNA Sanger sequencing at "laboratório de sequenciamento de ácidos nucleicos da FCFRP-USP" (Brazil). Bacterial and fungal sequencing amplicons were made publicly available on the GenBank database. Accession numbers ON623883-ON623889 and ON738686-ON738691, respectively.

A threshold of 97% of identity was established as minimal to assign an ASV to the respective species based on *16S rRNA* and *ITS* gene. The sequences presenting alignments with no more than 97% of similarity were blasted against the NCBI nr-database in order to relate the ASV with a cognate species.

Statistical analysis of correlation and regression

Correlation among ASVs and measured variables (i.e., volatile compounds and *luxS* gene expression) was performed using the R package microbiomeSeq (Ssekagiri et al., 2017). Regression analyses among *luxS* gene expression versus LAB loads was performed using the R package ggpmisc (Aphalo, 2020).

Results

In this study, four lab-scale fermentations were performed to unravel whether the influence of quorum sensing on microbial shifts has possible impacts on sensorial attributes development along cocoa fermentation. The experiments were planned to aim to understand the interaction of the autochthonous microbiota and selected starter cultures employed, as the raw material was not sterile to simulate field conditions. All fermentations exhibited microbial successions as determined by microbial enumeration and NGS. In terms of pH and temperature, no significant differences were observed for these variables, being the pH variation of 3.30 up to 3.45 for the experimental fermentations and 3.2 up to 4.5 for the non-inoculated one (F0). Moreover, temperature was not different among fermentations, reaching a maximum of 26ºC in laboratory conditions (Supplementary Fig. A.1). Regarding microbiota, microbial loads presented by experimental fermentations (F1, F2 and F3) were similar to the F0 fermentation as shown in Fig. 2A and were not differentially significant (Wilcoxon test p-value > 0.05). Besides, NGS corroborated enumeration analysis, revealing no significant differences (p-value > 0.05) in α-diversity dynamics caused by starters inoculation (Fig. 2B). Nonetheless, F1 and F3 fermentations presented augmented median values of observed and expected (Chao1) indexes, meaning these fermentations were more diverse in terms of intrasample microbial composition (a slightly enhanced number of different microorganisms detected). In contrast, in terms of equitability (Shannon measurements), F2 fermentation presented lower variations (boxplot heights). In addition to the observed number of ASVs and Chao1 estimates, the equitability of F2 fermentation suggests a low variation in microbial diversity and composition over the fermentation period (Fig. 2B). In summary, even with the addition of defined cocktails in F1, F2 and F3 fermentations, in general it was observed a restricted microbial diversity, which not necessarily means the composition was not impacted by the addition of starter cultures, but fermentations were not dissimilar regarding changes in microbial richness.

Microbial composition was partially influenced by starters

While discrete fluctuations of microbial diversity and low diversity exhibited by experimental fermentations, the inoculation of different cocktails resulted in distinct bacterial profiles, except for F2 and F3 which presented a slight semblance. In comparison to the non-inoculated fermentation (F0), characterised by a heterogeneous composition of non-dominant taxa (glued as <0.01% of representativity) (Fig. 3A), F1 fermentation exhibited a dominance of LAB, specifically lactobacilli (older genus *"Lactobacillus"*, Zheng et al. 2020) in the middle to the last phases of fermentation (48h-96h) (Fig. 3A). As in this fermentation all cocktails (composed by yeasts, LAB, and AAB) were inoculated, it was expected to observe higher amounts of AAB bacteria in the range of 72h-96h of fermentation, which was observed only in the final period of fermentation. F2 and F3 fermentations were characterised by the dominance of *Pantoea* (enterobacteria) from the middle to the last phases (48h-96h) and AAB at 96h of fermentation (Fig. 3A). The presence of AAB in the F2 fermentation was expected since only with the C3 cocktail was inoculated in that fermentation, but for the F3 fermentation the AAB dominance was unexpected as only the C1 cocktail was added. Another interesting observation was the ability of all cocktails (C1, C2, and C3) to replace accompanying detrimental microbiota, such as enterobacteria which was detected only in the start of fermentations (0h) for the inoculated fermentations (F1, F2, and F3) (Fig. 3A). Besides, *Gluconobacter* was detected after 72h of fermentation in all experiments. ASVs related to this genus were also detected in F0 fermentation (Fig. 3A).

Regarding the monitoring of starters fluctuations along fermentations, the taxonomic assignment for this level was satisfactory for the identification of two ASVs derived from the *Lm. fermentum* Lb1 strain, which dominated the middle to the last phases of F1 and were detected also in F2 (96h) and F3 (96h-lower amounts) fermentations (Fig. 3B). The *Lp. plantarum* Lb2 strain was related to three ASVs, being the ASVs "b" and "c" enriched in the F2 and F3 fermentations, which were not inoculated with the C2 cocktail. Other lactobacilli-related ASVs were classified by blast analysis against the NCBI nr-database since they do not present similarity with *16S rRNA* gene of the strains inoculated in fermentations. These lactobacilli were classified only in the genus level as *Liquorilactobacillus* spp (Fig. 3B). Moreover, it was not detected as lactobacilli ASVs in the control (F0) fermentation.

As expected, due to the latter inoculation (48h) of the C3 cocktail, *Acetobacter* spp. were detected in the final of fermentations (96h) along with *Gluconobacter* spp (Fig. 3A). The addition of a defined cocktail of *A. senegalensis* strains (C3-cocktail) was not enough to allow *Acetobacter* dominance in F1 and F2 fermentations, as this genus was also determined in the last phases of F0 and F3 fermentations as well (Fig. 3A).

The monitoring of the inoculated strains through the similarity of *16S rRNA* genes and the ASVs sequences was achieved and demonstrated the presence of six *Acetobacter* spp. ASVs, being four of them, originated from the inoculated *A. senegalensis* strains (Fig. 3C). However, no differentiation of ASVs originating from a cognate strain was possible as blast analysis evidenced a high similarity among the *16S rRNA* genes of these strains. Besides, some ASVs identified as *Acetobacter* sp. on Qiime2 were identified as *Gluconobacter* sp. by blast analysis against the NCBI nr-database (Fig. 3C). Only the ASV “c” belonging to the inoculated *A. senegalensis* strains was detected in significant amounts at the period of 96h of fermentation in most experimental fermentations (Fig. 3C).

The taxonomic composition of fungi was more diverse than bacteria and varied among fermentations. F0 fermentation was dominated by ASVs related to the *Saccharomyces* genus from the middle to the end of fermentation (48h-96h), while the inoculated fermentations presented a robust dominance of this genus along the entire time range (Fig. 3D). F1, F2, and F3 fermentations presented higher amounts of *Lasiodiplodia* at the start, being the middle and the end of fermentations marked by *Saccharomyces* dominance. *Diaporthe* ASVs were observed in the range of 24h-96h in F1 fermentation in association of *Candida* ASVs fluctuations in the period 0h-96h (Fig. 3D). The remaining F2 and F3 fermentation were virtually equivalent in terms of fungal composition.

Concerning the fungal species of interest (whose genera *Saccharomyces* and *Pichia* were inoculated in fermentations), only three ASVs were assigned to *Saccharomyces* genus and to the *S. cerevisiae* species, being impossible to distinguish from which strains these sequences were from (Fig. 3E). Only the “c” ASV was dominant in all fermentations, while the ASVs “a” and “b” were detected in low amounts (Fig. 3E). Concerning *Pichia*, taxonomic assignment was unable to assign the *Pichia* ASVs to the species level (Fig. 3F). Besides, *Pichia* ASVs exhibited a huge diversity of ASVs, summing at least 19 amplicon variants with uneven prevalence in fermentations (Fig. 3F).

Lp. plantarum species trend to express more *luxS* than *Lm. fermentum*

The Fig. 4 presents the data related to the expression of the *luxS* gene throughout the evaluated fermentations. Regarding the *luxS* gene expressed by the species *Lp. plantarum* (Fig. 4A), it was observed that the addition of the C2 cocktail (composed only by the strains *Lp. plantarum* Lb2 and *Lm. fermentum* Lb1) was able to enrich the expression of this gene in the F1 fermentation from the middle (48h) to the end (96h) of the fermentation. In relation to 48h of fermentation, the increase in gene expression was about 66.67 times greater than in the same period determined for F0 fermentation. Gene expression was linearly increasing at 72h and 96h fermentation times, which in relation to F0 fermentation was about 1,250 and 166.67 times greater, respectively. For the F2 and F3 fermentations, the expression of the *luxS* gene species-specific of *Lp. plantarum* was detected in the range of 72h and 96h of fermentation, being higher than the F0 fermentation only at the end of fermentation (96h) (Fig. 4A). During this period, the F2 and F3 fermentations showed higher expression of *luxS* related to the *Lp. plantarum* species are about 2.67 and 6.67 times greater, respectively, in relation to the F0 fermentation in the same period (Fig. 4A).

Curiously, the expression of the *luxS* gene by the species *Lm. fermentum*, even with the addition of the C2 cocktail, was not higher in relation to the F0 fermentation (Fig. 4B). Furthermore, only in F0 fermentation the expression of this gene was enormously increased but detected in large amounts only in the 72h period of fermentation (Fig. 4B). According to the data presented, *Lp. plantarum* tends to show greater expression of the *luxS* gene than the species *Lm. fermentum*. However, no pattern could be observed since there was no significant difference (Wilcoxon Test p-value>0.05) among fermentations in terms of gene expression. Furthermore, even comparing the density of lactic acid bacteria over time (presumptive LAB plate enumeration data) with the increase in *luxS* gene expression (Figs. 4C and 4D), in all fermentations as LAB cell density increases, there is an increase in *luxS* gene expression of both lactobacilli species, except for F1 fermentation, in which *luxS* expression by *Lm. fermentum* was inversely related to the increase in cell density.

Metabolic changes along fermentation and production of volatile compounds

Enzymatic degradation of cocoa pulp is crucial for flavour development as the infiltration of external compounds in consonance with endogenous hydrolases lead to a multitude of VOCs which directly impact sensorial attributes of fermented seeds (Aprotosoaie et al., 2016; Batista et al., 2016). In this study, 16 enzyme families were dosed: amilases, β-glucanases, cellulases, arabinofuranosidases, invertases, endoglucanases, xylanases, β-glucosidases, arabinases, xyloglucanases, esterases, β-xylosidases, pectinases, mannanases, cellobiohydrolases, and lipases (Fig. 5).

Higher enzymatic activities were determined for cellulases (72h), arabinofuranosidases (24h), invertases (24h), endoglucanases (24h), xylanases (24h), β-glucosidases (48h), arabinases (72h), mannanases (24h), and cellobiohydrolases (72h) in F1 when compared to the F0 fermentation. Amilases (72h), β-glucanases (72h), arabinofuranosidases (72h), invertases (72h), endoglucanases (72h), arabinases (48h), esterases (72h), and pectinases (48h) were augmented in F2 in relation to F0 fermentation. Arabinofuranosidases (72h), invertases (96h), arabinases (72h), and cellobiohydrolases (48h) were higher in F3 fermentation than F0. Interestingly, lipase activity was pronounced in F0 and F3 fermentations, and xyloglucanase activity was higher in the control fermentation than others (Fig. 5).

Additionally, the higher enzymatic activity determined in F1, F2 and F3 fermentations were proportional to the microbial shifts caused by inoculation of starter cultures. In F1, for instance, the augmented activities of cellulases (72h), arabinofuranosidases (24h), invertases (24h), endoglucanases (24h), xylanases (24h), β-glucosidases (48h), arabinases (72h), mannanases (24h), and cellobiohydrolases (72h) coincided to the peak of LAB dominance in that fermentation period (24h-72h), while in F2 the prevalence of higher enzymatic activities in the range of 48h-72h coincided to the period of AAB inoculation (Fig. 5).

To track the possible microbial actors related to enzymatic changes in cocoa seeds along fermentation a Kendall correlation was performed among genera and measured enzymatic activities. Few genera of the inoculated strains were significatively correlated to enzymatic changes (Fig. 6). *Pichia* and Lactobacilli were negatively associated with invertases and pectinases in F0, while for the experimental fermentations no significant association was observed (Fig. 6). Nevertheless, even with no statistical meaning, it's worth to emphasise that in most fermentations *Pichia* was positively correlated with esterase activity (F0 and F1), while *Saccharomyces* was positively correlated with amilase (F1 and F2), arabinase (F1), β-glucosidase (F1), cellulase (F1), cellobiohydrolase (F2 and F3), endoglucanase (F1), invertase (F1), mannanase (F1), pectinase (F1) and xylanase (F1). Lactobacilli exhibited higher correlation to arabinase (F3), β-glucanase (F1), endoglucanase (F1) and mannanase (F3). *Acetobacter* instead was more related to arabinose (F0), invertase (F0), mannanase (F0 and F3), pectinase (F0), β-xylosidase (F2) and esterase (F2) (Fig. 6).

The results of microbial metabolic activities were not only measured by enzymatic ability to breakdown complex sugars, but by the production of VOCs as well. Table 2 describes the VOCs identified by a qualitative HPLC analysis in both replicates of each inoculated fermentation (F1, F2 and F3) and the control one (F0). A total of 89 compounds were identified. Most of them representing alcohols and phenols (n=28), aldehydes and ketones (n=25), esters (n=13), and sulfur compounds (n=3) (Table 2) distributed unequally in fermentations. 20 compounds were not identified as playing a significative role, based on literature reviewing (Aprotosoaie et al., 2016; Misnawi & Ariza, 2011; Owusu et al., 2010; Rodriguez-Campos et al., 2011), on flavour production in cocoa fermentation, and so were represented by a trace in the Table 2.

Fig. 7 shows the scores of surveyed fermentations and demonstrates that even with the addition of defined cocktails the risk of generation of detrimental flavour precursors occurs. In all fermentations, but F2, is possible to visualise a tendency for sweetness-, fruity- and vegetal-like VOCs precursors at the beginning of fermentations, while at the end of the processes the emergence of undesirable compounds, such as wine-like and malty odors were detected. Nevertheless, some desired sensorial attributes were exalted. For example, in F0 fermentation the detection of undesirable compounds was balanced by the augment of fruity and chocolate sweetness precursors (Fig. 7). In F1 instead, at the end of fermentation, there was a reduction of chocolate notes, emergence of undesirable and vinegar-like VOCs, with the concomitant increasing of fruity notes. In F2 the tendency of production of undesirable VOCs was counterbalanced by the enhancement of fruity- and chocolate-like VOCs (Fig. 7). In F3 fermentation, there was an augment of undesirable compounds, decreasing of chocolate-like precursors and an enhancement of fruity-like VOCs (Fig. 7). Statistical analysis based on the Wilcoxon test showed the differences observed in the start to the end point of fermentations were not significant (p-value > 0.05).

In order to infer whether QS has a potential contribution for cocoa fermentation quality a Kendall correlation was performed among *luxS* expression amounts (at the beginning and end of fermentations), and the qualitative presence/absence of sensorial precursors determined by HPLC (at the beginning and end of fermentations) (Supplementary Fig. A.2). From the data derived from this study, it was impossible to precisely determine if QS directly impacted the quality of fermentations as no significant correlation ($p < 0.05$) was observed among *luxS* gene expression and sensorial profiles. However, the production of some precursors such as vegetal-, vinegar-, fruity-like (this presenting $p < 0.05$) and undesirable (i.e., malty- and wine-like) compounds was correlated with *Lp. plantarum luxS* gene expression. Conversely, Vegetal-like notes were more associated with *Lm. fermentum luxS* expression (Supplementary Fig. A.2).

Discussion

Cocoa fermentation has a huge importance for the development of sensorial traits in the raw material for chocolate's production. Several studies have been conducted to propose novel starter cultures to standardise the process, relating the dominance of candidate strains with the development of sensorial attributes along fermentation (Batista et al., 2016; García-Ríos et al., 2021; Magalhães da Veiga Moreira et al., 2017; Mota-Gutierrez et al., 2019; H. G. Ouattara et al., 2020). Following this direction, the present research aimed to investigate the possible contribution of QS for bacterial dominance, specifically lactic acid bacteria, in fermentation and if this could result in fermented beans with superior quality in terms of production of aroma precursors.

To scrutinise this hypothesis, yeasts, LAB, and AAB isolates from a spontaneous cocoa fermentation (Almeida et al., 2020) were selected as they are intrinsic to the cocoa and fermentation ecosystem. The yeasts *S. cerevisiae* strains F3, F6, F11 and F12 and the strains *Pichia kudriavzevii* PCA1 and *Pichia kluyveri* strain PCA4 were identified and genotypically characterised in this work, while the LAB *Lm. fermentum* Lb1 and *Lp. plantarum* Lb2, the AAB *A. senegalensis* strains MRS7, GYC10, GYC12, GYC19 and GYC27 were previously evaluated and genomic characterised (de Almeida et al., 2021; Almeida et al., 2022). These yeasts were selected to add contrast of fungal diversity and due to the valuable importance of fungi for cocoa fermentation (Crafack et al., 2013; Lefeber et al., 2012; Magalhães da Veiga Moreira et al., 2017), while the LAB and AAB were selected given the QS potential teased previously (Almeida et al., 2021) and adaptive potential to stressors (Almeida et al., 2022), respectively.

Cocoa fermentation is hard to standardise, but several studies emphasise the dominance of *S. cerevisiae*, *Lp. plantarum*, *Lm. fermentum* and *Acetobacter* species as core members (Lefeber et al., 2011; Meersman et al., 2013; Papalexandratou et al., 2011; Verce et al., 2021), with evidence to *Lm. fermentum* to the detriment of *Lp. plantarum*. In this work we employed a deep marker analysis based on the monitoring of ASVs along fermentation to assess whether the inoculated strains participated or not in fermentations. However, the technology of ASVs determination is not always able to distinguish among closely related strains but refers to the truly biological meaning of a sequence variant. In other words, ASVs can separate sequencing errors from biological variations in sequences (Callahan et al., 2017). In cocoa fermentation, ASVs were already employed by oligotyping methods (ASVs), which revealed a huge intraspecific variability of yeasts and bacteria (LAB and AAB) along fermentation (Díaz-Muñoz et al., 2021; Pacheco-Montealegre et al., 2020; Verce et al., 2021). Therefore, to discriminate if the ASVs might represent intraspecific strains variability or if a given strain originated several ASVs, the comparison of gene markers (*16S rRNA* and *ITS*) against the ASVs provides a useful resolution (Verce et al., 2021; Díaz-Muñoz et al., 2021).

The inoculation of *Pichia kudriavzevii* strain PCA1 and *Pichia kluyveri* strain PCA4 produced no significant changes in fermentations (F1, F2 and F3), as these strains were not dominant and no *Pichia* ASVs found matched the inoculated ones, suggesting these strains cannot replace indigenous *Pichia* species. The high intraspecies diversity of *Pichia* spp. (H. G. Ouattara & Niamké, 2021; Pereira et al., 2017) could explain the high numbers of different ASVs measured in this study. Regarding *S. cerevisiae*, it was shown that the inoculated strains were dominant in F1, F2 and F3 fermentations, but it was impossible to distinguish which strain dominated the scenarios as they presented high similarity in their *ITS* gene sequences. As observed, even without inoculation of any cocktail, the F0 fermentation presented the same *S. cerevisiae* strains, suggesting these strains are indigenous strains present in cocoa producing regions of Brazil and could be used for starter development. These observations are in accordance with the literature, as *Saccharomyces cerevisiae* strains trend to dominate fermentations from different regions and *Pichia* sp., especially *Pichia kudriavzevii* are more variable in genetic background, being dispersed in several regions (Ouattara et al., 2021).

Although no statistical difference was observed among the experimental fermentations, the F1 fermentation presented the highest loads of ASVs derived from *Lm. fermentum* reads in comparison to the *Lp. plantarum* ones. The *Lm. fermentum* dominance in cocoa fermentation is not novel and it was already explained in terms of physiological adaptations. It was demonstrated that heterofermentative *Lm. fermentum* isolates obtained from cocoa fermentation (indigenous strains) may be able to metabolise citric acid, oppositely to the type strain, as this characteristic seems species-specific and corresponds to the reality in the cocoa ecosystem: the pulp is rich in citric acid (Lefeber et al., 2010).

In cocoa fermentation, *Lp. plantarum* and *Lm. fermentum* species preferentially consume fructose in relation to glucose. This leads to mannitol reduction and acetic acid production by a heterofermentative metabolism. Thus, the metabolization of citric acid aids the dominance of *Lm. fermentum*, which could augment the quality of fermentation by replacing non-citrate fermenters (depectinizing yeasts), which capture voraciously carbohydrates and convert them directly to ethanol. Therefore, more production of acetic acid from fructose consumption and mannitol reduction allows the regeneration of NAD+ for extra ATP production, aiming for bacterial proliferation (Lefeber et al., 2010). Nevertheless, lactobacilli metabolism is inhibited in acid titles above 5%, limiting their growth and colonisation (Ouattara et al., 2016).

Regarding the temporal distribution of *Lp. plantarum* (beginning-middle of fermentation) and *Lm. fermentum* (middle-end of fermentation), surveys of spontaneous fermentations have already observed the shifts displayed by these species (Camu et al., 2007; Papalexandratou et al., 2013). As observed for *S. cerevisiae* strains, the same *Lp. plantarum* Lb2 and *Lm. fermentum* Lb1 strains were detected in the F0 fermentation, suggesting these strains are dominant in cocoa producing regions.

However, some lactobacilli ASVs were identified as *Liquorilactobacillus* spp, a genus commonly found in cocoa fermentation (Almeida & De Martinis, 2021). As the novel taxonomic classification was proposed (Zheng et al., 2020), the older genus *"Lactobacillus"* was now split in several genera, among them *Limosilactobacillus*, *Lactiplantibacillus,* and *Liquorilactobacillus*. The huge intra-genera variability explored by Zheng et al. (2020) explains the reason of previously classified *"Lactobacillus"* species on SILVA database were identified as *Liquorilactobacillus* spp. in this work, as to confirm the species level assignment, each ASV was blasted against the *16S rRNA* genes of the starter cultures strains and against the NCBI non-redundant database, which is already updated for the novel taxonomy. So, as reported by a metagenome-assembled genomes survey (Almeida & De Martinis, 2021), many microorganisms may still be overlooked in cocoa fermentation, and the new taxonomy sheds light into the huge variability of previous taxa which were named "*Lactobacillus"*, but in fact belong to other LAB genera.

Acetobacter species are dominant in cocoa fermentation (Miescher Schwenninger et al., 2016) and *A. senegalensis* was already proposed as a potential starter candidate (Illeghems et al., 2016). In this work, four of six *A. senegalensis* ASVs identified matched to the inoculated strains but no differentiation between the strains from C2 cocktail was possible due to higher similarities in their *16S rRNA* gene sequences. The determination of *A. senegalensis* strains was masked by the ambiguous identification of *Gluconobacter* species presenting high similarities for the *16S rRNA* gene as, during blast analysis, some *Acetobacter* ASVs were identified as *Gluconobacter* spp. However, the results raised in this study show no dominance of the inoculated *Acetobacter senegalensis* strains, which were detected in lower amounts, even presenting an interesting metabolic potential to circumvent adverse conditions in fermentative environments (Almeida et al., 2022).

Metabolization of cocoa seeds and VOCs production

To the best of the authors' knowledge, no study was still performed to measure enzymatic activities directly in cocoa fermentation and correlate it to microbial dynamics. In the present research, no significant association was made regarding microorganisms and enzymatic activities, which suggests the main enzymatic profiles determined derivate from the pulp and cotyledons metabolism (endogenous hydrolases) as disseminated in literature (De Vuyst & Weckx, 2016). A study showed the potential of some yeasts strains isolated from cocoa fermentation to produce β-glucosidases, xylanases, pectinases, cellulases, and lipases (Delgado-Ospina et al., 2020). Another study demonstrated the influence of *Bacillus* spp. to produce pectinases along cocoa fermentation (H. G. Ouattara et al., 2008). Besides, the inoculated *Lp. plantarum* Lb2 and *Lm. fermentum* Lb1 presented a remarkable potential to metabolise carbohydrates as several glycoside hydrolases, glycosyl transferases and carbohydrate esterases were annotated in their genomes (Almeida et al., 2021). The participation of microorganisms in enzymatic activity in cocoa fermentation seems to be complementary to the endogenous hydrolases, mostly acting as activators of these effectors along fermentation as no enzymatic activity was directly and statistically correlated to the total microbiota.

A satisfactory cocoa fermentation is crucial for the development of aroma and taste of chocolate. For this, microbiota stimulates the biochemical transformations inside the beans, leading to the formation of desirable precursors (Castro-Alayo et al., 2019). A study reported the inoculation of defined starters elevated the levels of acids and esters in fermented beans, while decreased the amounts of aldehydes and ketones, and alcohols (Moreira et al., 2021). The opposite was observed in this study, the starters inoculation slightly augmented the levels of alcohols. Aldehydes and ketones were discreetly diminished in F0 (0h to 96h) and F1 (0h to 96h), and augmented in F2 (0h to 96h), while in F3 no changes were observed. Esters were also detected and were enhanced (0h to 96h) in F0, F1, F2, and no changes were observed in F3. Alcohols are not specifically related to quality of fermentation as they reflect the microbial metabolism along the process, while aldehydes and ketones play a role in the development of flavour due to their carbonylic compounds. Following, esters are the second most important compounds for cocoa fermentation and are related to the fruity-like aroma in fermented beans (Aprotositae et al., 2015).

The tendency of increasing desirable sensorial compounds (aldehydes, ketones, and esters) was observed at the end of all fermentations but was accompanied also by the emergence of detrimental metabolites: wine- and malty-like compounds. Besides, the expression of *luxS* of *Lp. plantarum* was associated with detrimental metabolites for fermented seeds quality. However, the absence of statistical significance of difference among the fermentations entangles the delimitation of the relationship of QS-related bacteria and the quality of cocoa fermentation.

Quorum sensing and importance for fermentation

The importance of QS for bacterial stability and dominance in cocoa fermentation was first hypothesised by our research group. By the monitoring of a spontaneous cocoa fermentation carried out in Bahia state of Brazil, a metagenomic analysis predicted several taxa implied with interspecific QS, with evidence to lactobacilli (Almeida et al., 2020). In that work, it was proposed that novel studies could shed light into microbial ecology of cocoa fermentation by the evaluation of QS. As no *Acetobacter* spp. present *luxS* genes (Almeida et al., 2020; Almeida et al. 2022), the survey in this study was performed only for lactobacilli. It was observed *luxS* genes were highly expressed in F1 fermentation, most because of the inoculation of a high-density of lactobacilli cells. QS occurs in high-cell densities, and it is responsible for modulation of specific phenotypes that may confer adaptive advantages in harsh environments (Schluter et al., 2016). Initially, it was argued that QS could play a significant role in bacterial shifts in cocoa fermentation, impacting microbial succession (Almeida et al., 2020).

The data raised by this study brings a new vision, as it demonstrates the activity of QS along fermentation as argued previously (Almeida et al., 2020) and reveals the main player in this process: *Lp. plantarum*. While in F0 the bacterial diversity was low and in F2 and F3 fermentations there was a dominance of *Pantoea* genus, the addition of C2 cocktail induced the prevalence of lactobacilli, probably replacing *Enterobacteriaceae* members, which could confer stability and safety to the final product. It is worth highlighting that the *Pantoea* members were previously correlated with *luxS* genes (Almeida et al., 2020), and their replacement by *Lp. plantarum* could indicate a sum of forces that might proportionate *Lp. plantarum* dominance in detriment of *Enterobacteriaceae*.

By the other hand, in the view of some authors, the dominance of LAB in fermentation could be detrimental, due to the higher loads of lactic acid produced. Since it is a non-volatile compound, it may impact the taste of chocolate. Besides, research groups evidence the absence of LAB is not impeditive for fermentation (Ho et al., 2015, 2018). However, there is no consensus in literature about LAB detrimental properties, as many authors advocate these bacteria could be useful to replace undesirable microorganisms (Marwati et al., 2021) and provide good aroma notes (Viesser et al., 2020).

It's worth to evidence of a pangenome survey based on 404 and 63 genomes of *Lp. plantarum* and *Lm. fermentum* species showed that *Lp. plantarum* species present multiple *luxS* gene homologues distributed in six gene clusters, while *Lm. fermentum* species present two gene clusters (Almeida et al., 2021). This can explain why *Lm. fermentum luxS* gene expression was incipient in comparison to *Lp. plantarum* genes. The augmented number of *luxS* genes in *Lp. plantarum* genomes could enhance its gene expression and confer adaptive fitness to outperform other lactobacilli in the same environment. Future studies focusing on primer design to monitor other lactobacilli species along cocoa fermentation are needed to compare the *luxS* expression patterns.

Still it is hard to affirm QS has an influence on cocoa fermentation quality as no significant changes were determined by the addition of QS-related bacteria in fermentation and changes in VOCs. Besides, no significant association of enzymatic activities and lactobacilli was found. As shown, QS could be related only to dominance of *Lp. plantarum*, while the quality is a sum of factors involving all microbiota in that environment and the intrinsic activity of cotyledons endogenous hydrolases (De Vuyst & Weckx, 2016). At the same time, more studies are needed to understand if LAB really impacts positively or negatively the fermentation, and if negatively, the disruption of QS communication could help to diminish lactobacilli loads in fermentation, especially of *Lp. plantarum* species. On the other hand, QS could aid the augmentation of lactobacilli to replace detrimental microorganisms.

Conclusion

This was the first study to monitor enzymatic QS activities along cocoa fermentations. Although novel data was provided, this study was impacted by the intrinsic limitation of lab-scale fermentations, which not always can capture the real conditions of field such as the presence of a in house microbiota (from cocoa boxes, vessels, banana leaves and utensils), regional temperature and humidity and cocoa mass volume. Nevertheless, the findings were corroborated by literature in terms of microbial diversity and the microbial shifts during fermentation. The activity of *luxS* genes in all fermentations, but especially in that with an extra repertoire of lactobacilli, demonstrate the previous predictions of a metagenomic study, suggesting these experiments could be performed once in field conditions to compare the same observations. Moreover, it was impossible to fully link QS with cocoa fermentation quality, as additional tests are needed. However, the expression of *luxS* by lactobacilli *in situ* conditions is an interesting finding, as if future studies demonstrate LAB are detrimental to the process, due to lactic acid accumulation, strategies of QS disruption may be employed to limit these microorganisms. On the other hand, if these bacteria are crucial to the process, the stimulation of QS signalling would be useful to standardise the process favouring bacterial dominance. Finally, another limited knowledge in cocoa fermentation science is the absence of data regarding the microbial enzymatic activities and their extent in cocoa fermentation. This work showed and hypothesised that the main enzymatic activities related to changes in beans are displayed by endogenous hydrolases of the seeds, as no significant correlation of enzymes profiles and microbial composition was obtained. The data provided in this work could inspirate future studies to fill these new gaps.

CRediT authorship contribution statement

O.G.G. Almeida: Conceptualization, Formal analysis, Investigation, Methodology, Writing original draft, Writing - review & editing**. M.G. Pereira:** Conceptualization, Methodology, Writing - review & editing**. R. L. Bighetti-Trevisan:** Methodology, Writing, - review & editing**. E. Santos:** Methodology. **E.G. De Campos:** Methodology. **G.E. Felis:** Methodology, Writing- review**. L.H.S. Guimarães:** Methodology, Writing-

review. **M.L.T.M. Polizeli:** Methodology, Writing – review. **B.S. De Martinis:** Methodology, Writing – review. **E.C.P. De Martinis:** Conceptualization, Methodology, Writing – review & editing, Resources, Supervision.

Conflict of interests

The authors declare no conflict of interests.

Funding

This work was supported by a grant from The São Paulo Research Foundation, Brazil (FAPESP, grant #18/13564-3). O.G.G.A. is also grateful to FAPESP for Ph.D. fellowships (Processes # 17/13759-6 and # 18/26719-5). This research was also was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001. ECPDM is grateful for a Researcher Fellowship from CNPq (PQ-2, Proc.# 306330/2019-9).

References

Almeida, O. G.G., & De Martinis, E. C. P. (2021). Metagenome-Assembled Genomes Contribute to unraveling of the Microbiome of Cocoa Fermentation. *Applied and Environmental Microbiology*, *87*(16), 1–18. https://doi.org/10.1128/AEM.00584-21

Almeida, O. G.G., Gimenez, M. P., & De Martinis, E. C. P. (2022). Comparative pangenomic analyses and biotechnological potential of cocoa-related *Acetobacter senegalensis* strains. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *115*(1), 111–123. https://doi.org/10.1007/s10482-021-01684-7

Almeida, O. G.G., Pinto, U. M., Matos, C. B., Frazilio, D. A., Braga, V. F., von Zeska-Kress, M. R., & De Martinis, E. C. P. (2020). Does Quorum Sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An in silico perspective. *Food Research International*, *131*(January), 109034. https://doi.org/10.1016/j.foodres.2020.109034

Almeida, O.,G.,G., Vitulo, N., De Martinis, E. C. P., & Felis, G. E. (2021). Pangenome analyses of LuxS-coding genes and enzymatic repertoires in cocoa-related lactic acid bacteria. *Genomics*, *113*(4), 1659–1670. https://doi.org/10.1016/j.ygeno.2021.04.010

Aphalo, P. (2020). Learn R: As a Language. In *The R Series. Boca Raton and London: Chapman and Hall/CRC Press* (p. 350).

Aprotosoaie, A. C., Luca, S. V., & Miron, A. (2016). Flavor Chemistry of Cocoa and Cocoa Products-An Overview. *Comprehensive Reviews in Food Science and Food Safety*, *15*(1), 73–91. https://doi.org/10.1111/1541-4337.12180

Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, *55*(3), 541–555. https://doi.org/10.1016/j.mimet.2003.08.009

Batista, N. N., Ramos, C. L., Dias, D. R., Pinheiro, A. C. M., & Schwan, R. F. (2016). The impact of yeast starter cultures on the microbial communities and volatile compounds in cocoa fermentation and the resulting sensory attributes of chocolate. *Journal of Food Science and Technology*, *53*(2), 1101–1110. https://doi.org/10.1007/s13197-015-2132-5

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bortolini, C., Patrone, V., Puglisi, E., & Morelli, L. (2016). Detailed analyses of the bacterial populations in processed cocoa beans of different geographic origin, subject to varied fermentation conditions. *International Journal of Food Microbiology*, *236*, 98–106. https://doi.org/10.1016/j.ijfoodmicro.2016.07.004

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*(1), 248–254. https://doi.org/https://doi.org/10.1016/0003-2697(76)90527-3

Bushnell, B. (2014). *BBTools*. http://bbtools.jgi.doe.gov/

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Camu, N., De Winter, T., Verbrugghe, K., Cleenwerck, I., Vandamme, P., Takrama, J. S., Vancanneyt, M., & De Vuyst, L. (2007). Dynamics and biodiversity of populations of lactic acid bacteria and acetic acid bacteria involved in spontaneous heap fermentation of cocoa beans in Ghana. *Applied and Environmental Microbiology*, *73*(6), 1809–1824. https://doi.org/10.1128/AEM.02189-06

Castro-Alayo, E. M., Idrogo-Vásquez, G., Siche, R., & Cardenas-Toro, F. P. (2019). Formation of aromatic compounds precursors during fermentation of Criollo and Forastero cocoa. *Heliyon*, *5*(1). https://doi.org/10.1016/j.heliyon.2019.e01157

Chagas Junior, G. C. A., Ferreira, N. R., & Lopes, A. S. (2021). The microbiota diversity identified during the cocoa fermentation and the benefits of the starter cultures use: an overview. *International Journal of Food Science and Technology*, *56*(2), 544–552. https://doi.org/10.1111/ijfs.14740

Crafack, M., Mikkelsen, M. B., Saerens, S., Knudsen, M., Blennow, A., Lowor, S., Takrama, J., Swiegers, J. H., Petersen, G. B., Heimdal, H., & Nielsen, D. S. (2013). Influencing cocoa flavour using *Pichia kluyveri* and *Kluyveromyces marxianus* in a defined mixed starter culture for cocoa fermentation. *International Journal of Food Microbiology*, *167*(1), 103–116. https://doi.org/10.1016/j.ijfoodmicro.2013.06.024

De Roos, J., & De Vuyst, L. (2018). Acetic acid bacteria in fermented foods and beverages. *Current Opinion in Biotechnology*, *49*, 115–119. https://doi.org/10.1016/j.copbio.2017.08.007

De Vuyst, L., & Weckx, S. (2016). The cocoa bean fermentation process: from ecosystem analysis to starter culture development. *Journal of Applied Microbiology*, *121*(1), 5–17. https://doi.org/10.1111/jam.13045

Delgado-Ospina, J., Triboletti, S., Alessandria, V., Serio, A., Sergi, M., Paparella, A., Rantsiou, K., & Chaves-López, C. (2020). Functional biodiversity of yeasts isolated from Colombian fermented and dry Cocoa beans. *Microorganisms*, *8*(7), 1–17. https://doi.org/10.3390/microorganisms8071086

Díaz-Muñoz, C., Van de Voorde, D., Comasio, A., Verce, M., Hernandez, C. E., Weckx, S., & De Vuyst, L. (2021). Curing of Cocoa Beans: Fine-Scale Monitoring of the Starter Cultures Applied and Metabolomics of the Fermentation and Drying Steps. *Frontiers in Microbiology*, *11*(January). https://doi.org/10.3389/fmicb.2020.616875

Esteve-Zarzoso, B., Belloch, C., Uruburu, F., & Querol, A. (1999). Identification of yeasts by RFLP analysis of the 5.8S rRNA gene and the two ribosomal internal transcribed spacers. *International Journal of Systematic Bacteriology*, *49*(1), 329–337. https://doi.org/10.1099/00207713-49-1-329

Farrera, L., Colas, A., Noue, D., Strub, C., Guibert, B., Kouame, C., Grabulos, J., Montet, D., & Teyssier, C. (2021). Towards a Starter Culture for Cocoa Fermentation by the Selection of Acetic Acid Bacteria. *Fermentation*, *7*, 1–19.

García-Ríos, E., Lairón-Peris, M., Muñiz-Calvo, S., Heras, J. M., Ortiz-Julien, A., Poirot, P., Rozès, N., Querol, A., & Guillamón, J. M. (2021). Thermo-adaptive evolution to generate improved *Saccharomyces cerevisiae* strains for cocoa pulp fermentations. *International Journal of Food Microbiology*, *342* (September 2020). https://doi.org/10.1016/j.ijfoodmicro.2021.109077

Gu, Y., Li, B., Tian, J., Wu, R., & He, Y. (2018). The response of LuxS/AI-2 quorum sensing in Lactobacillus fermentum 2-1 to changes in environmental growth conditions. *Annals of Microbiology*, *68*(5), 287–294. https://doi.org/10.1007/s13213-018-1337-z

Ho, V. T. T., Fleet, G. H., & Zhao, J. (2018). Unravelling the contribution of lactic acid bacteria and acetic acid bacteria to cocoa fermentation using inoculated organisms. *International Journal of Food Microbiology*, *279*(April), 43–56. https://doi.org/10.1016/j.ijfoodmicro.2018.04.040

Ho, V. T. T., Zhao, J., & Fleet, G. (2015). The effect of lactic acid bacteria on cocoa bean fermentation. *International Journal of Food Microbiology*, *205*, 54–67. https://doi.org/10.1016/j.ijfoodmicro.2015.03.031

Illeghems, K., Pelicaen, R., De Vuyst, L., & Weckx, S. (2016). Assessment of the contribution of cocoa-derived strains of *Acetobacter ghanensis* and *Acetobacter senegalensis* to the cocoa bean fermentation process through a genomic approach. *Food Microbiology*, *58*, 68–78. https://doi.org/10.1016/j.fm.2016.03.013

Instituto Adolfo Lutz. (2008). *Métodos físico-químicos para análise de alimentos* (O. Zenebon, N. S. Pascuet, & P. Tiglea (eds.); 4th ed.). http://www.ial.sp.gov.br/resources/editorinplace/ial/2016_3_19/analisedealimentosial_2008.pdf

Jiang, L., Luo, Y., Cao, X., Liu, W., Song, G., & Zhang, Z. (2021). LuxS quorum sensing system mediating *Lactobacillus plantarum* probiotic characteristics. *Archives of Microbiology*, *203*(7), 4141–4148. https://doi.org/10.1007/s00203-021-02404-5

Johansen, P., & Jespersen, L. (2017). Impact of quorum sensing on the quality of fermented foods. *Current Opinion in Food Science*, *13*, 16–25. https://doi.org/10.1016/j.cofs.2017.01.001

Jordan, & Bisanz. (2018). *qiime2R: Importing QIIME2 artifacts and associated data into R sessions.* https://github.com/jbisanz/qiime2R

Kersters-Hilderson, H., Claeyssens, M., Van Doorslaer, E., Saman, E., & De Bruyne, C. K. B. T.-M. in E. (1982). [60] β-d-xylosidase from *Bacillus pumilus*. In *Complex Carbohydrates Part D* (Vol. 83, pp. 631–639). Academic Press. https://doi.org/https://doi.org/10.1016/0076-6879(82)83062-0

Lefeber, T., Gobert, W., Vrancken, G., Camu, N., & De Vuyst, L. (2011). Dynamics and species diversity of communities of lactic acid bacteria and acetic acid bacteria during spontaneous cocoa bean fermentation in vessels. *Food Microbiology*, *28*(3), 457–464. https://doi.org/10.1016/j.fm.2010.10.010

Lefeber, T., Janssens, M., Camu, N., & De Vuyst, L. (2010). Kinetic analysis of strains of lactic acid bacteria and acetic acid bacteria in cocoa pulp simulation media toward development of a starter culture for cocoa bean fermentation. *Applied and Environmental Microbiology*, *76*(23), 7708–7716. https://doi.org/10.1128/AEM.01206-10

Lefeber, T., Papalexandratou, Z., Gobert, W., Camu, N., & De Vuyst, L. (2012). On-farm implementation of a starter culture for improved cocoa bean fermentation and its influence on the flavour of chocolates produced thereof. *Food Microbiology*, *30*(2), 379–392. https://doi.org/10.1016/j.fm.2011.12.021

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods*, *25*(4), 402–408. https://doi.org/10.1006/meth.2001.1262

Magalhães da Veiga Moreira, I., de Figueiredo Vilela, L., da Cruz Pedroso Miguel, M. G., Santos, C., Lima, N., & Freitas Schwan, R. (2017). Impact of a Microbial Cocktail Used as a Starter Culture on Cocoa Fermentation and Chocolate Flavor. *Molecules (Basel, Switzerland)*, *22*(5). https://doi.org/10.3390/molecules22050766

Marwati, T., Purwaningsih, Djaafar, T. F., Sari, A. B. T., & Hernani. (2021). Inhibition the growth of fungi and improving the quality of cocoa beans through fermentation using lactic acid bacteria. *IOP Conference Series: Earth and Environmental Science*, *807*(2). https://doi.org/10.1088/1755-1315/807/2/022048

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061217

Meersman, E., Steensels, J., Mathawan, M., Wittocx, P. J., Saels, V., Struyf, N., Bernaert, H., Vrancken, G., & Verstrepen, K. J. (2013). Detailed analysis of the microbial population in Malaysian spontaneous cocoa pulp fermentations reveals a core and variable microbiota. *PLoS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0081559

Miescher Schwenninger, S., Freimüller Leischtfeld, S., & Gantenbein-Demarchi, C. (2016). High-throughput identification of the microbial biodiversity of cocoa bean fermentation by MALDI-TOF MS. *Letters in Applied Microbiology*, *63*(5), 347–355. https://doi.org/10.1111/lam.12621

Miller, G. L. (1959). Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Analytical Chemistry*, *31*(3), 426–428. https://doi.org/10.1021/ac60147a030

Misnawi, & Ariza, B. T. S. (2011). Use of gas Chromatography-Olfactometry in combination with solid phase micro extraction for cocoa liquor aroma analysis. *International Food Research Journal*, *18*(2), 829–835.

Moreira, I., Costa, J., Vilela, L., Lima, N., Santos, C., & Schwan, R. (2021). Influence of S. cerevisiae and P. kluyveri as starters on chocolate flavour. *Journal of the Science of Food and Agriculture*, *101*(10), 4409–4419. https://doi.org/10.1002/jsfa.11082

Mota-Gutierrez, J., Barbosa-Pereira, L., Ferrocino, I., & Cocolin, L. (2019). Traceability of functional volatile compounds generated on inoculated cocoa fermentation and its potential health benefits. *Nutrients*, *11*(4). https://doi.org/10.3390/nu11040884

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

Ogunremi, O. R., Freimüller Leischtfeld, S., & Miescher Schwenninger, S. (2022). MALDI-TOF MS profiling and exopolysaccharide production properties of lactic acid bacteria from Kunu-zaki - A cereal-based Nigerian fermented beverage. *International Journal of Food Microbiology*, *366*(January), 109563. https://doi.org/10.1016/j.ijfoodmicro.2022.109563

Ooi, T. S., Ting, A. S. Y., & Siow, L. F. (2020). Influence of selected native yeast starter cultures on the antioxidant activities, fermentation index and total soluble solids of Malaysia cocoa beans: A simulation study. *Lwt*, *122*(December 2019), 108977. https://doi.org/10.1016/j.lwt.2019.108977

Ouattara, D., Ouattara, H., Adom, J., Goualié, B., Koua, G., Doué, G., & Niamke, S. (2016). Screening of Lactic Acid Bacteria Capable to Breakdown Citric Acid during Ivorian Cocoa Fermentation and Response of Bacterial Strains to Fermentative Conditions. *British Biotechnology Journal*, *10*(3), 1–10. https://doi.org/10.9734/bbj/2016/19279

Ouattara, H. G., Elias, R. J., & Dudley, E. G. (2020). Microbial synergy between *Pichia kudriazevii* YS201 and *Bacillus subtilis* BS38 improves pulp degradation and aroma production in cocoa pulp simulation medium. *Heliyon*, *6*(1). https://doi.org/10.1016/j.heliyon.2020.e03269

Ouattara, H. G., Koffi, B. L., Karou, G. T., Sangaré, A., Niamke, S. L., & Diopoh, J. K. (2008). Implication of *Bacillus* sp. in the production of pectinolytic enzymes during cocoa fermentation. *World Journal of Microbiology and Biotechnology*, *24*(9), 1753–1760. https://doi.org/10.1007/s11274-008-9683-9

Ouattara, H. G., & Niamké, S. L. (2021). Mapping the functional and strain diversity of the main microbiota involved in cocoa fermentation from Cote d'Ivoire. *Food Microbiology*, *98*(June 2020). https://doi.org/10.1016/j.fm.2021.103767

Owusu, M., Petersen, M. a, & Heimdal, H. (2010). Assessment of Aroma of Chocolate Produced From Two Ghanaian Cocoa Fermentation Types. *Expression of Multidisciplinary Flavour Science*, *January*, 363–366.

Pacheco-Montealegre, M. E., Dávila-Mora, L. L., Botero-Rute, L. M., Reyes, A., & Caro-Quintero, A. (2020). Fine Resolution Analysis of Microbial Communities Provides Insights Into the Variability of Cocoa Bean Fermentation. *Frontiers in Microbiology*, *11*(April), 1–15. https://doi.org/10.3389/fmicb.2020.00650

Papalexandratou, Z., Camu, N., Falony, G., & De Vuyst, L. (2011). Comparison of the bacterial species diversity of spontaneous cocoa bean fermentations carried out at selected farms in Ivory Coast and Brazil. *Food Microbiology*, *28*(5), 964–973. https://doi.org/10.1016/j.fm.2011.01.010

Papalexandratou, Z., Lefeber, T., Bahrim, B., Lee, O. S., Daniel, H. M., & De Vuyst, L. (2013). *Hanseniaspora opuntiae, Saccharomyces cerevisiae, Lactobacillus fermentum,* and *Acetobacter pasteurianus* predominate during well-performed Malaysian cocoa bean box fermentations, underlining the importance of these microbial species for a successful cocoa bean fermentation process. *Food Microbiology*, *35*(2), 73–85. https://doi.org/10.1016/j.fm.2013.02.015

Park, H., Shin, H., Lee, K., & Holzapfel, W. (2016). Autoinducer-2 properties of kimchi are associated with lactic acid bacteria involved in its fermentation. *International Journal of Food Microbiology*, *225*, 38–42. https://doi.org/10.1016/j.ijfoodmicro.2016.03.007

Pencreac'h, G., & Baratti, J. C. (1996). Hydrolysis of p-nitrophenyl palmitate in n-heptane by the *Pseudomonas cepacia* lipase: A simple test for the determination of lipase activity in organic media. *Enzyme and Microbial Technology*, *18*(6), 417–422. https://doi.org/https://doi.org/10.1016/0141-0229(95)00120-4

Pereira, G. V. M., Alvarez, J. P., Neto, D. P. de C., Soccol, V. T., Tanobe, V. O. A., Rogez, H., Góes-Neto, A., & Soccol, C. R. (2017). Great intraspecies diversity of *Pichia kudriavzevii* in cocoa fermentation highlights the importance of yeast strain selection for flavor modulation of cocoa beans. *Lwt*, *84*, 290–297. https://doi.org/10.1016/j.lwt.2017.05.073

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. https://doi.org/10.1093/nar/gks1219

Rathnayake, I. V. N., Megharaj, M., Bolan, N., & Naidu, R. (2010). Tolerance of Heavy Metals by Gram Positive Soil Bacteria. *International Journal of Environmental Engineering*, *2*(5), 191–195.

Rodriguez-Campos, J., Escalona-Buendía, H. B., Orozco-Avila, I., Lugo-Cervantes, E., & Jaramillo-Flores, M. E. (2011). Dynamics of volatile and non-volatile compounds in cocoa (Theobroma cacao L.) during fermentation and drying processes using principal components analysis. *Food Research International*, *44*(1), 250–258. https://doi.org/10.1016/j.foodres.2010.10.028

Romanens, E., Freimüller Leischtfeld, S., Volland, A., Stevens, M., Krähenmann, U., Isele, D., Fischer, B., Meile, L., & Miescher Schwenninger, S. (2019). Screening of lactic acid bacteria and yeast strains to select adapted anti-fungal co-cultures for cocoa bean fermentation. *International Journal of Food Microbiology*, *290*(September 2018), 262–272. https://doi.org/10.1016/j.ijfoodmicro.2018.10.001

Saunshi, Y. B., Sandhya, M. V. S., Rastogi, N. K., & Murthy, P. S. (2020). Starter consortia for on-farm cocoa fermentation and their quality attributes. *Preparative Biochemistry and Biotechnology*, *50*(3), 272–280. https://doi.org/10.1080/10826068.2019.1689508

Schluter, J., Schoech, A. P., Foster, K. R., & Mitri, S. (2016). The Evolution of Quorum Sensing as a Mechanism to Infer Kinship. *PLoS Computational Biology*, *12*(4), 1–18. https://doi.org/10.1371/journal.pcbi.1004848

Shetty, S. A., & Lahti, L. (2020). *Microbiomeutilities:Utilities for Microbiome Analytics*. https://microsud.github.io/microbiomeutilities/

Schwendimann, L., Kauf, P., Fieseler, L., Gantenbein-Demarchi, C., & Miescher Schwenninger, S. (2015). Development of a quantitative PCR assay for rapid detection of *Lactobacillus plantarum* and *Lactobacillus fermentum* in cocoa bean fermentation. Journal of Microbiological Methods, 115, 94–99. https://doi.org/10.1016/j.mimet.2015.05.022

Ssekagiri, A., Sloan, W. T., & Ijaz, U. Z. (2017). *MicrobiomeSeq*. https://github.com/umerijaz/microbiomeSeq

Tait, E., Perry, J. D., Stanforth, S. P., & Dean, J. R. (2014). Identification of Volatile Organic Compounds Produced by Bacteria Using HS-SPME-GC–MS. *Journal of Chromatographic Science*, *52*(4), 363–373. https://doi.org/10.1093/chromsci/bmt042

Tobias, N. J., Brehm, J., Kresovic, D., Brameyer, S., Bode, H. B., & Heermann, R. (2020). New Vocabulary for Bacterial Communication. *ChemBioChem*, *21*(6), 759–768. https://doi.org/10.1002/cbic.201900580

Tulini, F. L., Hymery, N., Haertlé, T., Le Blay, G., & De Martinis, E. C. P. (2016). Screening for antimicrobial and proteolytic activities of lactic acid bacteria isolated from cow, buffalo and goat milk and cheeses marketed in the southeast region of Brazil. *Journal of Dairy Research*, *83*(1), 115–124. https://doi.org/DOI: 10.1017/S0022029915000606

Vaudano, E., & Garcia-Moruno, E. (2008). Discrimination of Saccharomyces cerevisiae wine strains using microsatellite multiplex PCR and band pattern analysis. *Food Microbiology*, *25*(1), 56–64. https://doi.org/10.1016/j.fm.2007.08.001

Verce, M., Schoonejans, J., Hernandez Aguirre, C., Molina-Bravo, R., De Vuyst, L., & Weckx, S. (2021). A Combined Metagenomics and Metatranscriptomics Approach to Unravel Costa Rican Cocoa Box Fermentation Processes Reveals Yet Unreported Microbial Species and Functionalities. *Frontiers in Microbiology*, *12*(February), 1–24. https://doi.org/10.3389/fmicb.2021.641185

Viesser, J. A., de Melo Pereira, G. V., de Carvalho Neto, D. P., Favero, G. R., de Carvalho, J. C., Goés-Neto, A., Rogez, H., & Soccol, C. R. (2021). Global cocoa fermentation microbiome: revealing new taxa and microbial functions by next generation sequencing technologies. *World Journal of Microbiology and Biotechnology*, *37*(7), 1–17. https://doi.org/10.1007/s11274-021-03079-2

Viesser, J. A., de Melo Pereira, G. V., de Carvalho Neto, D. P., Vandenberghe, L. P. d. S., Azevedo, V., Brenig, B., Rogez, H., Góes-Neto, A., & Soccol, C. R. (2020). Exploring the contribution of fructophilic lactic acid bacteria to cocoa beans fermentation: Isolation, selection and evaluation. *Food Research International*, *136*(June), 109478. https://doi.org/10.1016/j.foodres.2020.109478

Visintin, S., Ramos, L., Batista, N., Dolci, P., Schwan, F., & Cocolin, L. (2017). Impact of *Saccharomyces cerevisiae* and *Torulaspora delbrueckii* starter cultures on cocoa beans fermentation. *International Journal of Food Microbiology*, *257*(June), 31–40. https://doi.org/10.1016/j.ijfoodmicro.2017.06.004

White, T. J. (1990). Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics. In *PCR Protocols, a Guide to Methods and Applications* (pp. 315–322).

Wu, S., Liu, J., Liu, C., Yang, A., & Qiao, J. (2020). Quorum sensing for population-level control of bacteria and potential therapeutic applications. *Cellular and*

Molecular Life Sciences, *77*(7), 1319–1343. https://doi.org/10.1007/s00018-019-03326-8

Yang, Q., Wang, Y., An, Q., Sa, R., Zhang, D., & Xu, R. (2021). Research on the role of LuxS/AI-2 quorum sensing in biofilm of *Leuconostoc citreum* 37 based on complete genome sequencing. *3 Biotech*, *11*(4). https://doi.org/10.1007/s13205-021-02747-2

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., O'toole, P. W., Pot, B., Vandamme, P., Walter, J., Watanabe, K., Wuyts, S., Felis, G. E., Gänzle, M. G., & Lebeer, S. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *International Journal of Systematic and Evolutionary Microbiology*, *70*(4), 2782–2858. https://doi.org/10.1099/ijsem.0.004107

Table 1. Primers sequences for qRT-PCR reactions.

Primer designation	Sequence 5'3'	Target	Reference
Luxplan1F	ACGTGATCGTATGGATGG	*Lp. plantarum luxS*	Almeida et al. (2021)
Luxplan1R	CCCTTGTACGTCTTCCCA		
PlanF	TTACATTTGAGTGAGTGGCGAACT	*Lp. plantarum 16S rRNA*	Schwendimann et al. (2015)
PlanR	AGGTGTTATCCCCCGCTTCT		
LFermF	GCACCTGATTGATTTTGGTCG	*Lm. fermentum 16S rRNA*	
LFermR	GGTATTAGCATCTGTTTCCAAATG		
lux forward	AAACGGTGAGCAGTCAATCA	*Lm. fermentum luxS*	Gu et al. (2018)
lux reverse	AAGGTCCTAAGGGTGACAAGA		

Table 2. Detailing of volatile metabolic compounds detected in experimental fermentations and sensorial characteristics.

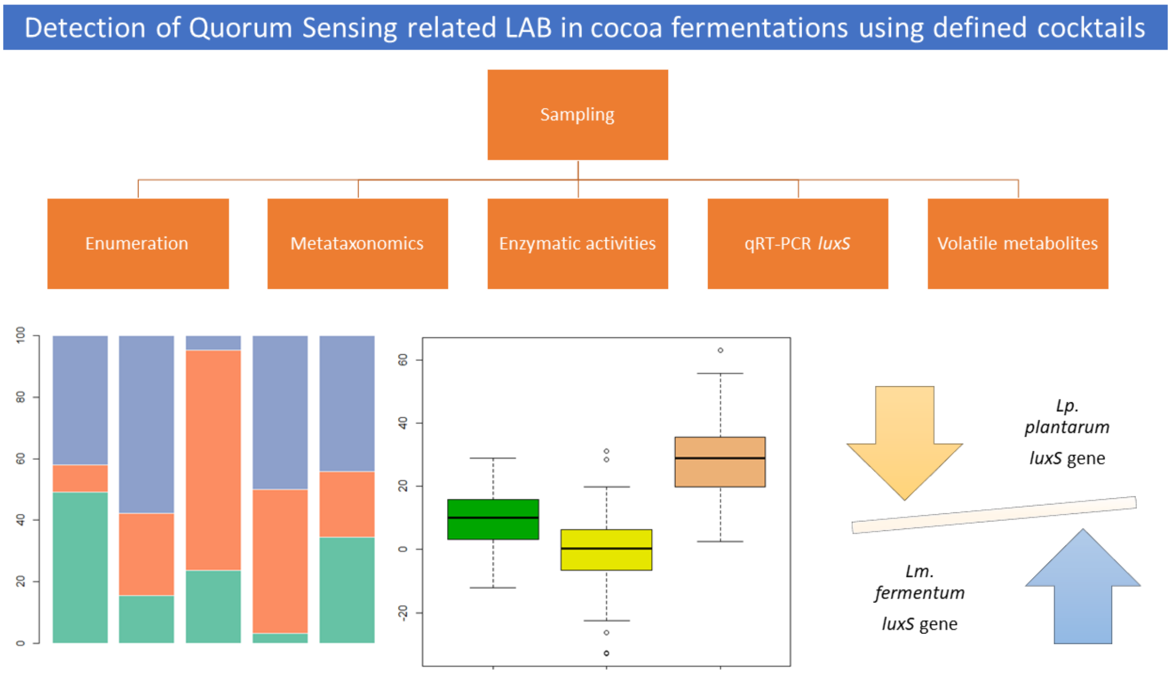
Fermentation	Compounds class	Metabolite	Sensory	Odor
F0 – 0h				
	Alcohols and phenols	(S)-(+)-2-Pentanol	Vegetal	Green, mild green
	-	3-propoxy-1-propene	-	-
	Aldehydes and ketones	2-methylbutanal	Sweet chocolate	Chocolate
	Aldehydes and ketones	2-methylpropanal	Sweet chocolate	Chocolate
	Esters	2-Pentanol, acetate	Vegetal	Green, mild green
	Aldehydes and ketones	2-Pentanone	Fruity	Fruity
F0 – 96h				
	Alcohols and phenols	Ethanol	-	-
	Alcohols and phenols	(S)-(+)-2-Pentanol	Vegetal	Green, mild green
	Alcohols and phenols	2-methyl-1-butanol	Fruity	Fruity, grape
	Alcohols and phenols	3-methyl-1-butanol	Undesirable	Malty
	Esters	3-methyl-1-butanol, acetate	Undesirable	Malty
	Alcohols and phenols	1-Propanol	Sweet chocolate	Sweet, candy
	Alcohols and phenols	2-methyl-1-propanol	Undesirable	Wine-like
	Aldehydes and ketones	2-Pentanone	Fruity	Fruity
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate
	Sulfur compounds	Dimethyl sulfide	Undesirable	Sulfurous

Fermentation	Compounds class	Metabolite	Sensory	Odor
	Alcohols and phenols	Ethanol	-	-
	Esters	Ethyl Acetate	Fruity	Fruity
F1 – 0h				
	Alcohols and phenols	Ethanol	-	-
	Sulfur compounds	Dimethyl sulfide	Undesirable	Sulfurous
	Aldehydes and ketones	2-methylpropanal	Sweet chocolate	Chocolate
	-	2-methyl-3-Buten-2-ol	-	-
	Aldehydes and ketones	2-methylbutanal	Sweet chocolate	Chocolate
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate
	Aldehydes and ketones	2-pentanone	Fruity	Fruity
	Alcohols and phenols	(S)-2-pentanol	Vegetal	Green, mild green
	Alcohols and phenols	(R)-2-pentanol	Vegetal	Green, mild green
	Alcohols and phenols	3-methyl-2-butanol		
	Esters	2-Pentanol, acetate	Vegetal	Green, mild green
	-	Ethanone, 1-[4-[4-(2-hydroxyethyl)-1-piperazinylsulfonyl]phenyl]-	-	-
F1 – 96h				
	Alcohols and phenols	(R)-2-pentanol	Vegetal	Green, mild green
	Alcohols and phenols	(S)-2-pentanol	Vegetal	Green, mild green

Fermentation	Compounds class	Metabolite	Sensory	Odor
	Esters	2-methyl-1-butanol, acetate	Fruity	Fruity, grape
	Esters	3-methyl-1-butanol, acetate	Undesirable	Malty
	-	2-methyl-3-Buten-2-ol	-	-
	Aldehydes and ketones	2-methylbutanal	Sweet chocolate	Chocolate
	Aldehydes and ketones	2-pentanone	Fruity	Fruity
	Alcohols and phenols	3-methyl-1-butanol	Undesirable	Malty
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate
	Esters	Acetic acid, 2-methylpropyl ester	Vinegar-like	Vinegar
	Esters	Acetic acid, methyl ester	Fruity	Fruity
	Sulfur compounds	Dimethyl sulfide	Undesirable	Sulfurous
	Alcohols and phenols	Ethanol	-	-
	Esters	Ethyl Acetate	Fruity	Fruity
	Alcohols and phenols	2-methyl-1-propanol	Undesirable	Wine-like
F2 – 0h				
	-	2-methyl-3-buten-2-ol	-	-
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate
	-	Acetone	-	-
	Alcohols and phenols	Ethanol	-	-
	-	Hexane	-	-
	-	Methylene Chloride	-	-
	-	3-methylpentane	-	-
F2 – 96h				

Fermentation	Compounds class	Metabolite	Sensory	Odor
	Alcohols and phenols	3-methyl-1-butanol	Undesirable	Malty
	Esters	3-methyl-1-butanol, acetate	Undesirable	Malty
	Alcohols and phenols	2-methyl-1-propanol	Undesirable	Wine-like
	Alcohols and phenols	2-heptanol	Fruity	Fruity
	Aldehydes and ketones	2-Pentanone	Fruity	Fruity
	-	2-methyl-3-buten-2-ol	-	-
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate
	Aldehydes and ketones	2-methylbutanal	Sweet chocolate	Chocolate
	Alcohols and phenols	Ethanol	-	-
	Esters	Ethyl Acetate	Fruity	Fruity
	-	Hexane	-	-
	-	Methylene Chloride	-	-
	Aldehydes and ketones	2-methylpropanal	Sweet chocolate	Chocolate
F3 – 0h				
	Aldehydes and ketones	2-methylbutanal	Sweet chocolate	Chocolate
	Alcohols and phenols	2-Pentanol	Vegetal	Green, mild green
	Esters	2-pentanol, acetate	Vegetal	Green, mild green
	Aldehydes and ketones	2-Pentanone	Fruity	Fruity
	-	2-methyl-3-buten-2-ol	-	-
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate

Fermentation	Compounds class	Metabolite	Sensory	Odor
	Alcohols and phenols	Ethanol	-	-
	-	Hexane	-	-
	-	Methylene Chloride	-	-
	-	trimethyloxirane	-	-
	Aldehydes and ketones	2-methylpropanal	Sweet chocolate	Chocolate
F3 – 96h				
	Alcohols and phenols	3-methyl-1-butanol	Undesirable	Malty
	Alcohols and phenols	2-methyl-1-propanol	Undesirable	Wine-like
	Alcohols and phenols	(S)-2-heptanol	Fruity	Fruity
	Aldehydes and ketones	2-Heptanone	Green Flowery	Fruity
	Aldehydes and ketones	2-methylbutanal	Sweet chocolate	Chocolate
	Aldehydes and ketones	2-Pentanone	Fruity	Fruity
	-	2-methyl-3-buten-2-ol	-	-
	Aldehydes and ketones	3-methylbutanal	Sweet chocolate	Chocolate
	Alcohols and phenols	Ethanol	-	-
	Esters	Ethyl Acetate	Fruity	Fruity
	-	Hexane	-	-
	-	Methyl vinyl ketone	-	-
	-	Methylene Chloride	-	-



Detection of Quorum Sensing related LAB in cocoa fermentations using defined cocktails

Sampling

Enumeration

Metataxonomics

Enzymatic activities

qRT-PCR *luxS*

Volatile metabolites

Legend: Graphical abstract.

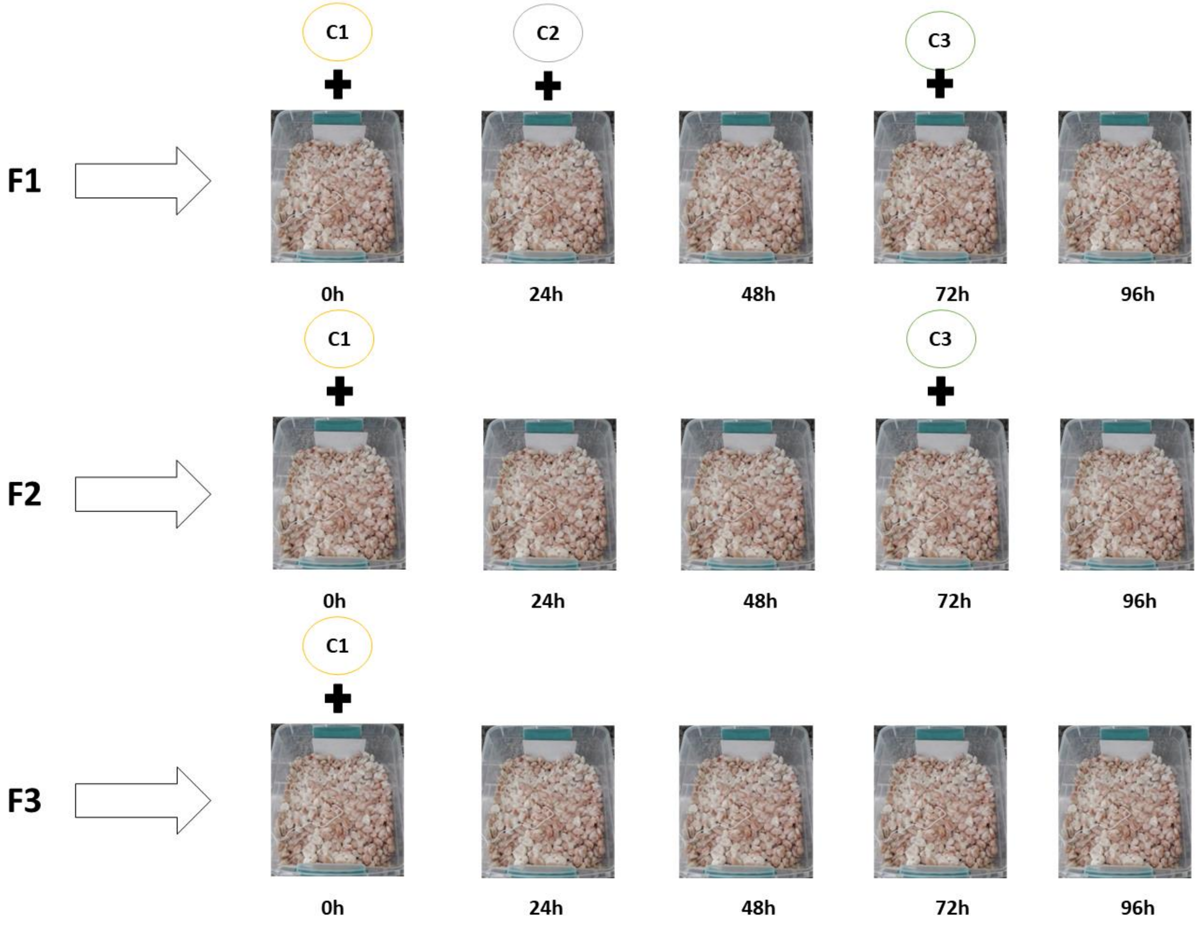


Fig. 1. Experimental design. Schema of inoculation of the defined cocktails C1 (*Pichia kudriavzevii* PCA1, *Pichia kluyveri* PCA4, *Saccharomyces cerevisiae* strains F3, F6, F11, and F12), C2 (*Lactiplantibacillus plantarum* Lb2 and *Lm. fermentum* Lb1), and

C3 (*Acetobacter senegalensis* strains MRS7, GYC10, GYC12, GYC19 and GYC27) in F1, F2 and F3 fermentations.

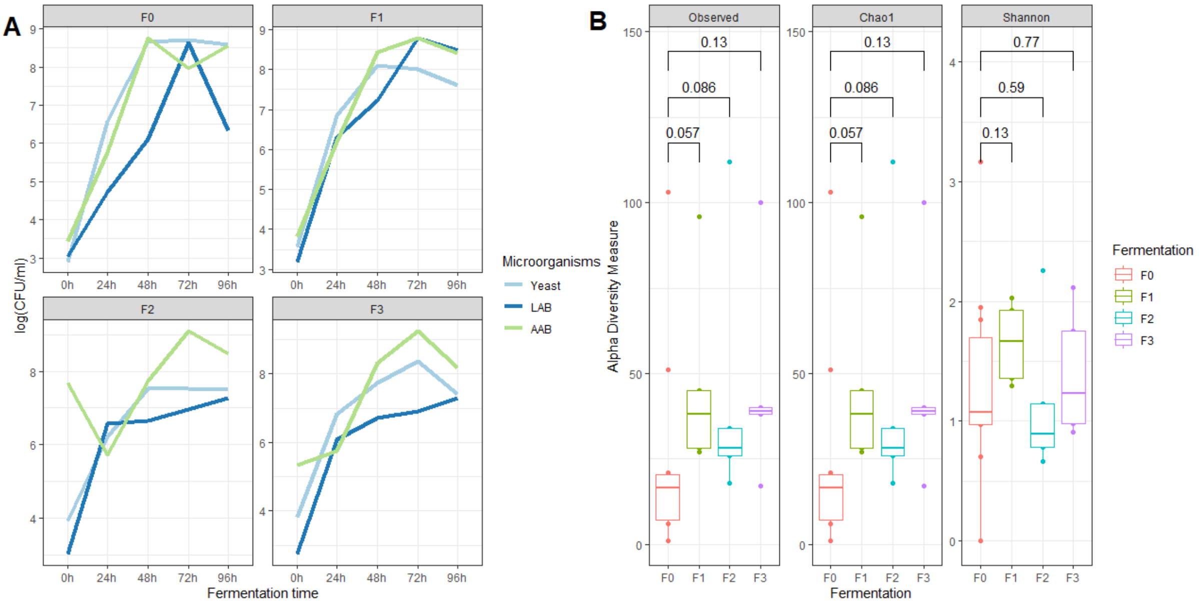


Fig. 2. (A) Enumeration of culturable microbial groups in fermentations and (B) representation of uncultured microbial diversity generated by NGS. The comparison lines above box plots and respective numbers stands for p-values calculations of Wilcoxon test from the comparison of F0 and F1, F2, and F3 fermentations.

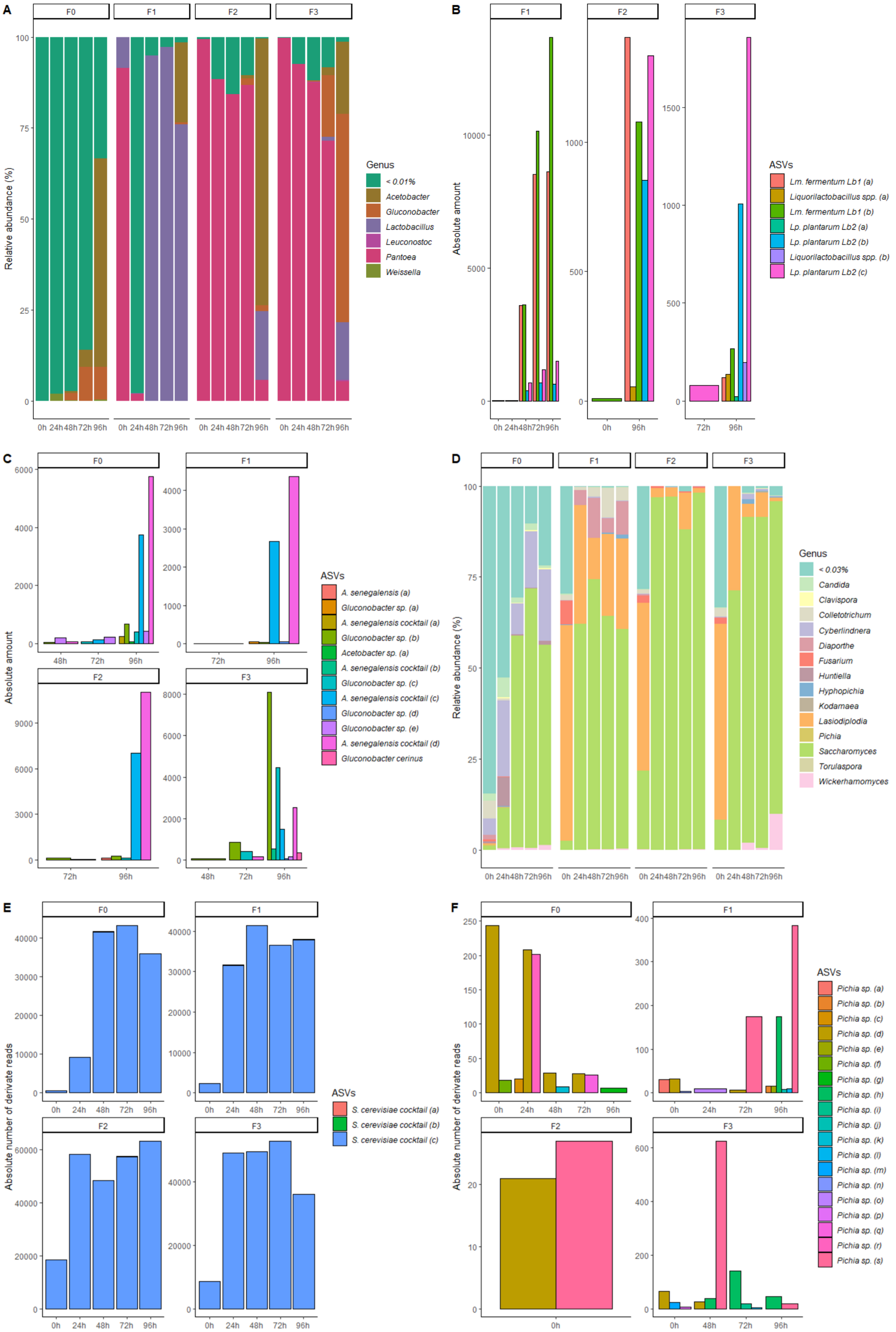


Fig. 3. Microbial composition along cocoa fermentation. (A) Bacterial composition of main genera detected along fermentation. (B) Scrutinization of ASVs identified as lactobacilli and association with inoculated strains (determined by blasting ASVs

against strains' *16S rRNA* gene) and autochthonous ones (determined by blasting ASVs against the NCBI database). (C) Depiction of ASVs related to *Acetobacter* sp. and association with inoculated strains (determined by blasting ASVs against strains' *16S rRNA* gene) and autochthonous ones (determined by blasting ASVs against the NCBI database). (D) Fungal composition of main genera detected along fermentation. (E) Representation of *S. cerevisiae* ASVs diversity in fermentation and association with inoculated strains (determined by blasting ASVs against strains' *16S rRNA* gene) and autochthonous ones (determined by blasting ASVs against the NCBI database). (F) Depiction of ASVs related to *Pichia* genus and association with inoculated strains (determined by blasting ASVs against strains' *16S rRNA* gene) and autochthonous ones (determined by blasting ASVs against the NCBI database).

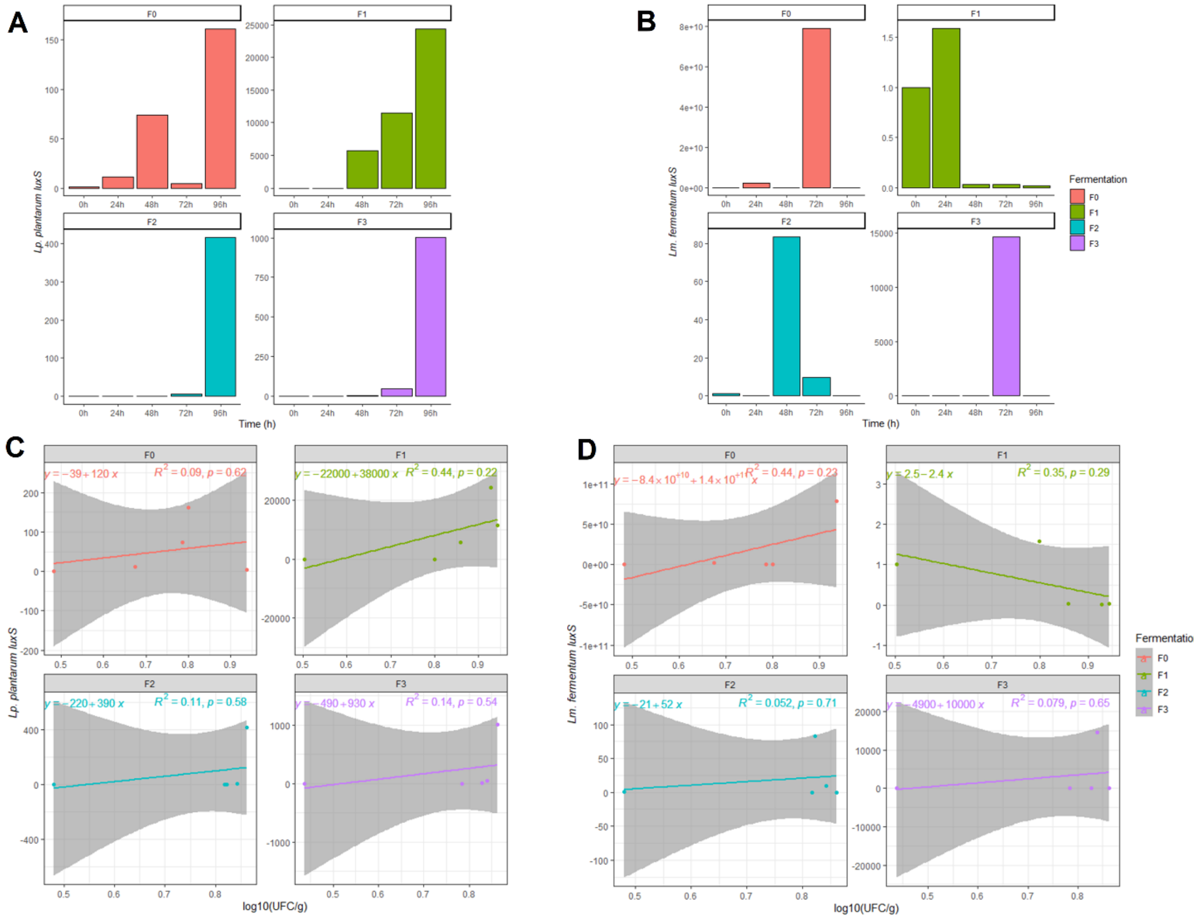


Fig. 4. *luxS* gene expression in lab-scale cocoa fermentations. (A) Variation of $2^{-\Delta\Delta Ct}$ values of *Lp. plantarum luxS* gene expression along fermentation. (B) Variation of $2^{-\Delta\Delta Ct}$ values of *Lm. fermentum luxS* gene expression along fermentation; (C) Linear regression of *Lp. plantarum luxS* gene expression versus LAB enumeration (cellular density) along fermentation; (D) Linear regression of *Lm. fermentum luxS* gene expression versus LAB enumeration (cellular density) along fermentation.

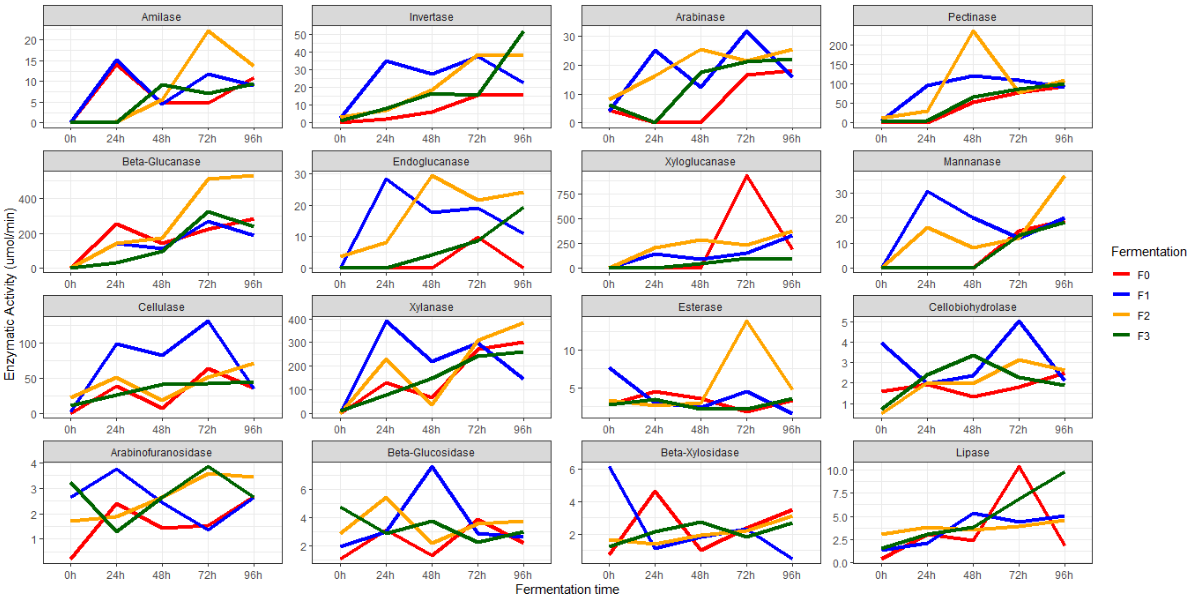


Fig. 5. Enzymatic profiles measured in surveyed fermentations related to complex carbohydrates and polysaccharides breakdown.

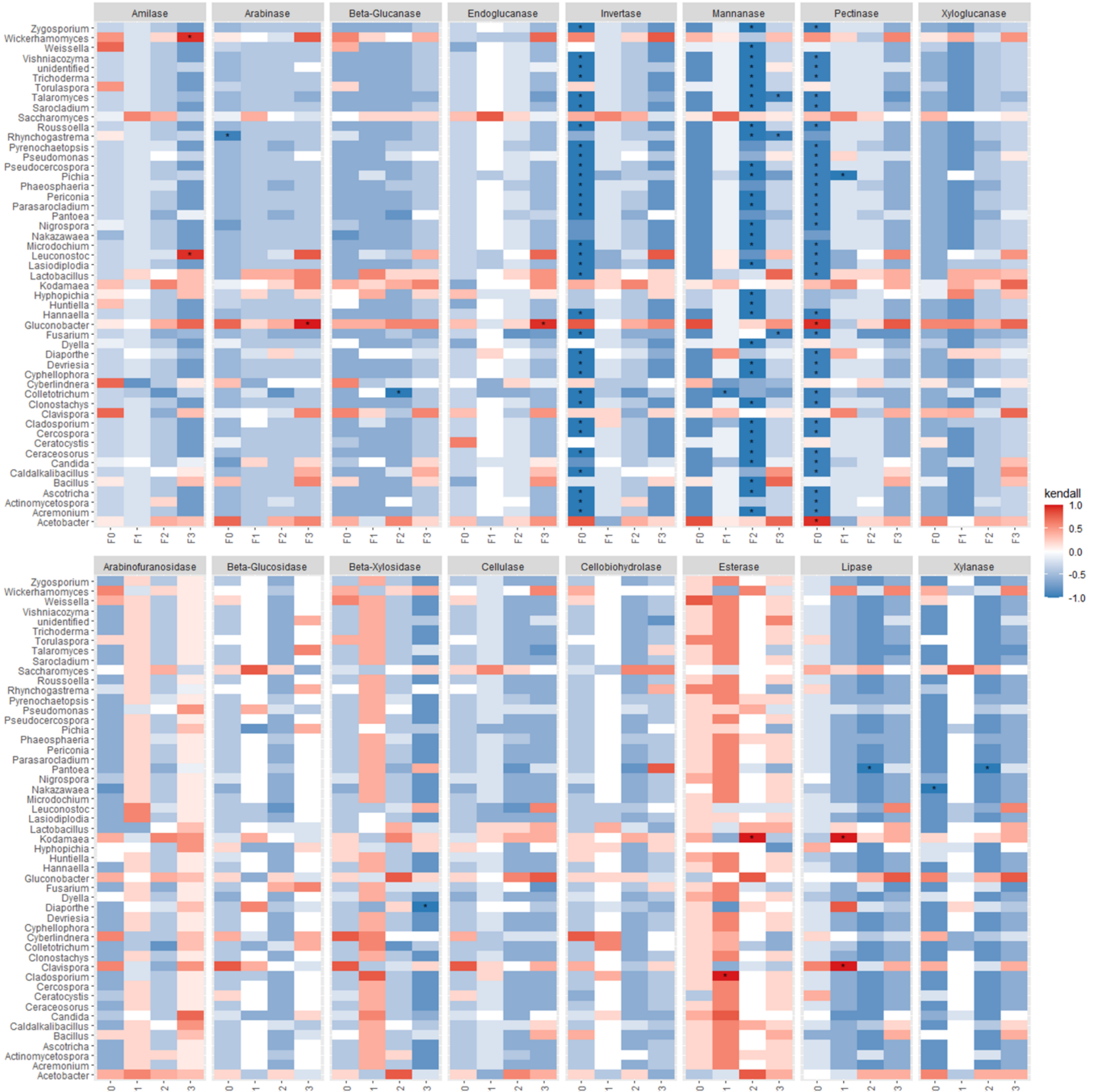


Fig. 6. Correlation of enzymatic profiles and microbial composition determined by the Kendall correlation test.

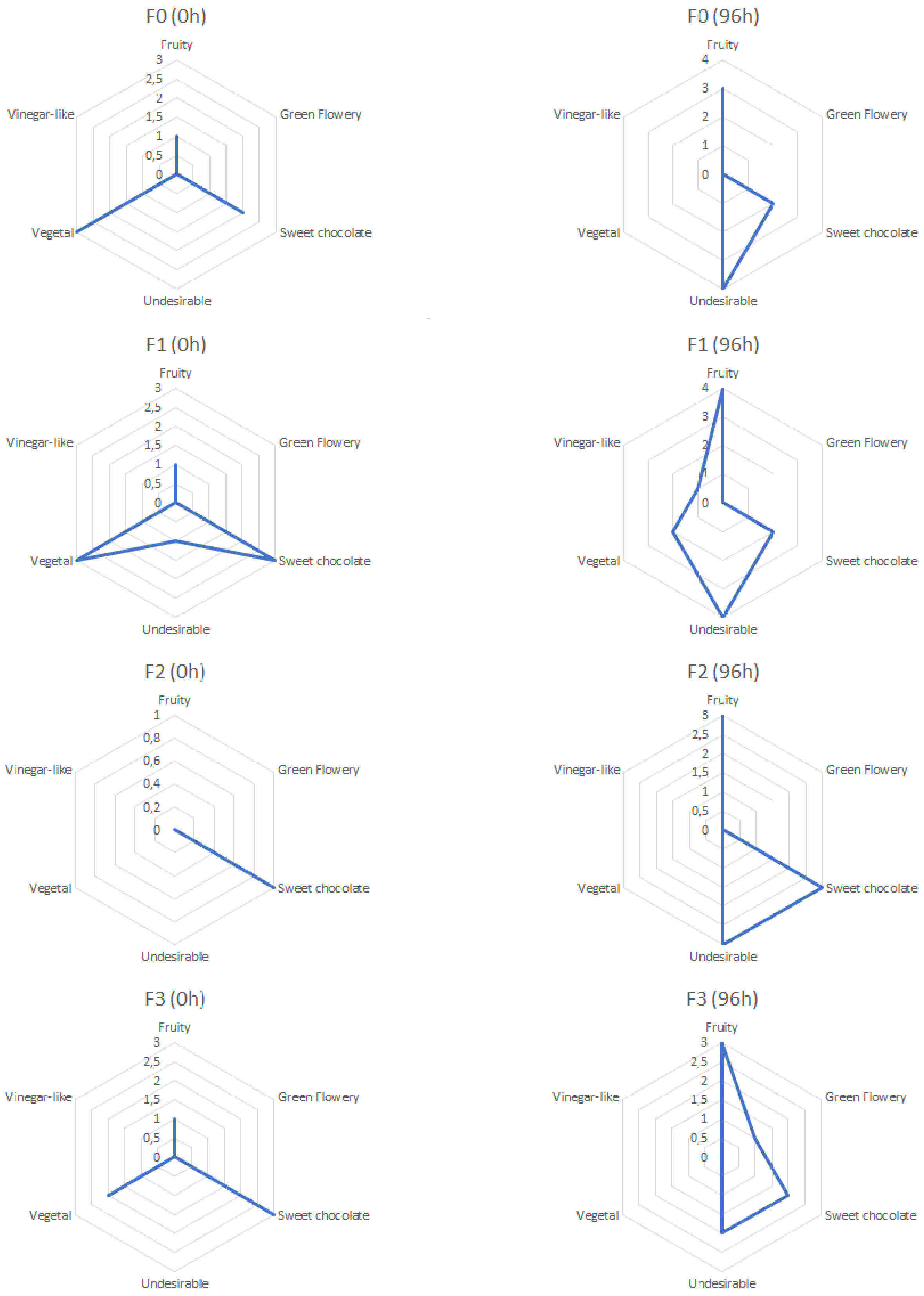


Fig. 7. Radar plots representing the scores of sensorial traits present at the beginning (0h) and at the end (96h) of fermentations.

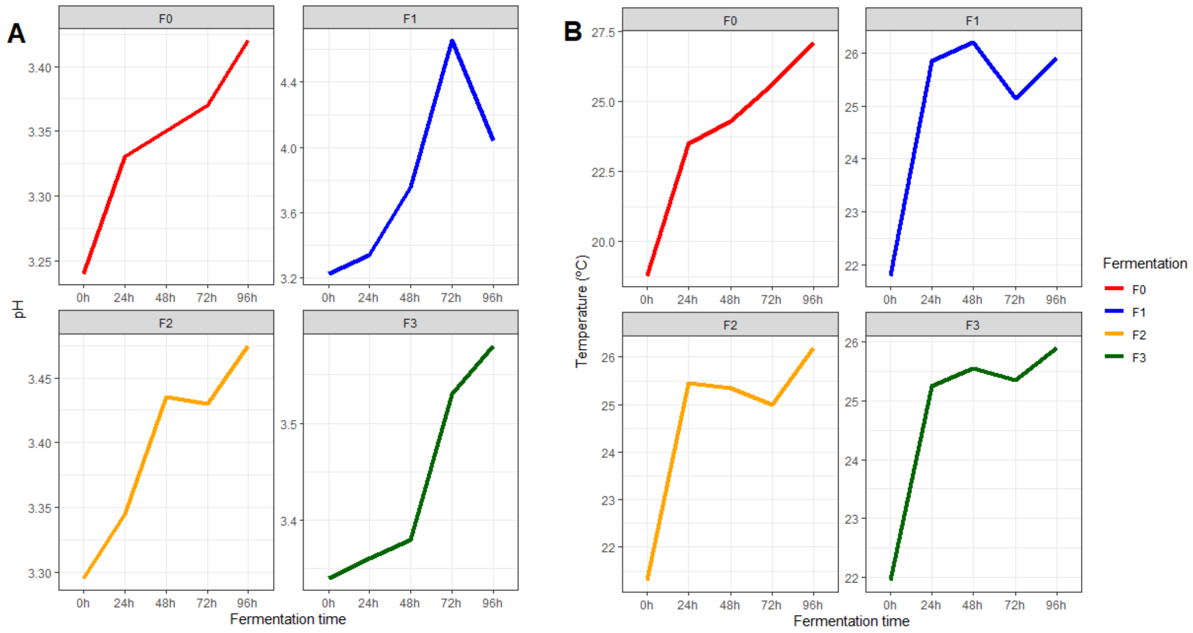


Fig. A.1. pH and Temperature monitoring measured in experimental fermentations.

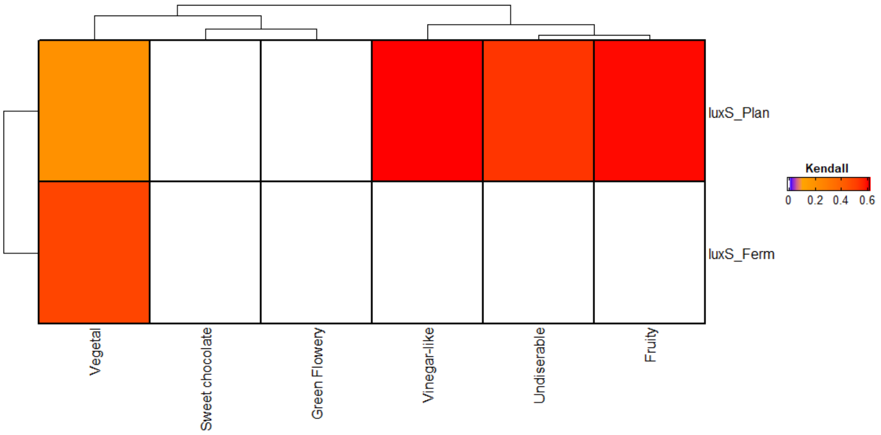


Fig. A.2. Kendall correlation among sensorial profiles and luxS gene expression of Lp. plantarum (luxS_Plan) and Lm. fermentum (luxS_Ferm) species.

Conclusions

7. Conclusions

The process of cocoa fermentation has a huge variability in microbial composition but some microorganisms such as *Saccharomyces, Pichia, Pantoea*, *Lactiplantibacillus plantarum, Limosilactobacillus fermentum, Acetobacter,* and *Gluconobacter* were observed in Brazilian cocoa fermentations, as shown in the first Chapter and fourth Chapter, independently if the fermentation was carried out in field or in lab conditions. The repetition of detection of these members in cocoa fermentation performed in Brazil shows the prevalence of these microbial entities in Brazilian producing regions. On the other hand, it was demonstrated the fungi are present in fermentation ranges, not only during the first 24h of fermentation, which deserves special attention to investigate their interaction with other microbes. Regarding QS potential, it was entirely detected almost linearly during the first 72h of fermentation, a period which coincided with LAB abundances in the fermentative scenario. Comparative genomic analysis revealed *Lp. plantarum* and *Lm. fermentum* carry *luxS* gene distributed in different clusters as some of these genes are paralogous and present different genomic organisations. The horizontal acquisition of *luxS* by some of *Lm. fermentum* strains was an interesting result, pointing out the possible importance of this gene in higher copies and distributed in different genomic arrangements, which may have relationship with the modular expression of *luxS* in consonance with other genes and activated in different conditions as these regions may be regulated by different promoters. However, no investigation of this matter was still performed and could be done in a future to disclose these novel hypotheses arisen from this thesis. Another interesting point of this study was the presence of quorum quenching and intraspecific QS effectors in *A. senegalensis*. The strains recovered from the spontaneous cocoa beans fermentation did not present *luxS* genes, which exclude the possibility of influence of the expression of this gene in *Acetobacter* species along fermentation. However, the detection of heat shock proteins, alcohol and aldehyde dehydrogenases enzymes indicates a potential use of these strains as starters in fermentative processes, such as vinegar fermentation. Besides, the data generated in the fourth Chapter corroborate their use in industrial fermentations, as some *A. senegalensis* from defined cocktails were viable in lab scale cocoa fermentations. Finally, even with the difficult to reproduce field conditions, the experimental fermentations with inoculated cocoa-related microbiota have shown the *luxS* gene of *Lp. plantarum* was expressed in all fermentation time points, while the *luxS* gene of *Lm. fermentum* was expressed in the middle to the end of fermentations, evidencing the central question haphazard in this thesis was completely answered: *Lp. plantarum* is strongly related to the *luxS* activity as initially supposed and considering the QS could be related to bacterial adaptations to harsh conditions and confer fitness for bacteria involved with this process, the *luxS* expression in higher levels by *Lp. plantarum* may indicate it could be an important determinant for this bacterium dominance in cocoa fermentation. However, this raises other hypothesis that are out of the scope of this single work. In conclusion, this set of studies scrutinized several implications of QS and other determinants for bacterial adaptation in cocoa fermentation as well as the cocoa-related biodiversity by the assessment of microbial diversity and function.

References

8. References

GUDA, Prasanna; GADHE, Shruthi. PRIMARY PROCESSING OF COCOA. **International Journal of Agricultural, Science and Research**, *[S. l]*, v. 7, 2017. Available in: http://www.tjprc.org/publishpapers/2-50-1490778585-57.IJASRAPR201757.pdf.

ALVAREZ-VILLAGOMEZ, K. G.; LEDESMA-ESCOBAR, C. A.; PRIEGO-CAPOTE, F.; ROBLES-OLVERA, V. J.; GARCÍA-ALAMILLA, P. Influence of the starter culture on the volatile profile of processed cocoa beans by gas chromatography–mass spectrometry in high resolution mode. **Food Bioscience**, *[S. l.]*, v. 47, n. March, p. 101669, 2022. DOI: 10.1016/j.fbio.2022.101669. Available in: https://doi.org/10.1016/j.fbio.2022.101669.

BATISTA, Nádia Nara; RAMOS, Cíntia Lacerda; DIAS, Disney Ribeiro; PINHEIRO, Ana Carla Marques; SCHWAN, Rosane Freitas. The impact of yeast starter cultures on the microbial communities and volatile compounds in cocoa fermentation and the resulting sensory attributes of chocolate. **Journal of Food Science and Technology**, *[S. l.]*, v. 53, n. 2, p. 1101–1110, 2016. DOI: 10.1007/s13197-015-2132-5.

CAMU, Nicholas; DE WINTER, Tom; VERBRUGGHE, Kristof; CLEENWERCK, Ilse; VANDAMME, Peter; TAKRAMA, Jemmy S.; VANCANNEYT, Marc; DE VUYST, Luc. Dynamics and biodiversity of populations of lactic acid bacteria and acetic acid bacteria involved in spontaneous heap fermentation of cocoa beans in Ghana. **Applied and Environmental Microbiology**, *[S. l.]*, v. 73, n. 6, p. 1809–1824, 2007. DOI: 10.1128/AEM.02189-06.

CAMU, Nicholas; GONZÁLEZ, Ángel; DE WINTER, Tom; VAN SCHOOR, Ann; DE BRUYNE, Katrien; VANDAMME, Peter; TAKRAMA, Jemmy S.; ADDO, Solomon K.; DE VUYST, Luc. Influence of turning and environmental contamination on the dynamics of populations of lactic acid and acetic acid bacteria involved in spontaneous cocoa bean heap fermentation in Ghana. **Applied and Environmental Microbiology**, *[S. l.]*, v. 74, n. 1, p. 86–98, 2008. DOI: 10.1128/AEM.01512-07.

CASTRO-ALAYO, Efraín M.; IDROGO-VÁSQUEZ, Guillermo; SICHE, Raúl; CARDENAS-TORO, Fiorella P. Formation of aromatic compounds precursors during fermentation of Criollo and Forastero cocoa. **Heliyon**, *[S. l.]*, v. 5, n. 1, 2019. DOI: 10.1016/j.heliyon.2019.e01157.

CHAGAS JUNIOR, Gilson Celso Albuquerque; FERREIRA, Nelson Rosa; LOPES, Alessandra Santos. The microbiota diversity identified during the cocoa fermentation and the benefits of the starter cultures use: an overview. **International Journal of Food Science and Technology**, *[S. l.]*, v. 56, n. 2, p. 544–552, 2021. DOI: 10.1111/ijfs.14740.

CRAFACK, Michael et al. Influencing cocoa flavour using *Pichia kluyveri* and *Kluyveromyces marxianus* in a defined mixed starter culture for cocoa fermentation. **International Journal of Food Microbiology**, *[S. l.]*, v. 167, n. 1, p. 103–116, 2013. DOI: 10.1016/j.ijfoodmicro.2013.06.024.

DE ARAÚJO, Jean Aquino; FERREIRA, Nelson Rosa; DA SILVA, Silvia Helena Marques; OLIVEIRA, Guilherme; MONTEIRO, Ruan Campos; ALVES, Yamila Fernandes Mota; LOPES, Alessandra Santos. Filamentous fungi diversity in the natural fermentation of Amazonian cocoa beans and the microbial enzyme activities. **Annals of Microbiology**, *[S. l.]*, v. 69, n. 9, p. 975–987, 2019. DOI: 10.1007/s13213-019-01488-1.

DE VUYST, L.; WECKX, S. The cocoa bean fermentation process: from ecosystem analysis to starter culture development. **Journal of Applied Microbiology**, *[S. l.]*, v. 121, n. 1, p. 5–17, 2016. DOI: 10.1111/jam.13045.

DE VUYST, Luc; LEROY, Frederic. Functional role of yeasts, lactic acid bacteria and acetic acid bacteria in cocoa fermentation processes. **FEMS Microbiology Reviews**, *[S. l.]*, v. 44, n. 4, p. 432–453, 2020. DOI: 10.1093/femsre/fuaa014.

FARRERA, Lucie; COLAS, Alexandre; NOUE, De; STRUB, Caroline; GUIBERT, Benjamin; KOUAME, Christelle; GRABULOS, Joël; MONTET, Didier; TEYSSIER, Corinne. Towards a Starter Culture for Cocoa Fermentation by the Selection of Acetic Acid Bacteria. **Fermentation**, *[S. l.]*, v. 7, p. 1–19, 2021.

FIGUEROA-HERNÁNDEZ, Claudia; MOTA-GUTIERREZ, Jatziri; FERROCINO, Ilario; HERNÁNDEZ-ESTRADA, Zorba J.; GONZÁLEZ-RÍOS, Oscar; COCOLIN, Luca; SUÁREZ-QUIROZ, Mirna L. The challenges and perspectives of the selection of starter cultures for fermented cocoa beans. **International Journal of Food**

Microbiology, *[S. l.]*, v. 301, n. January, p. 41–50, 2019. DOI: 10.1016/j.ijfoodmicro.2019.05.002. Available in: https://doi.org/10.1016/j.ijfoodmicro.2019.05.002.

GUEHI, Tagro S.; DADIE, Adjéhi T.; KOFFI, Kouadio P. B.; DABONNE, Soumaïla; BAN-KOFFI, Louis; KEDJEBO, Kra D.; NEMLIN, Gnopo J. Performance of different fermentation methods and the effect of their duration on the quality of raw cocoa beans. **International Journal of Food Science and Technology**, *[S. l.]*, v. 45, n. 12, p. 2508–2514, 2010. DOI: 10.1111/j.1365-2621.2010.02424.x.

HAWVER, Lisa A.; JUNG, Sarah A.; NG, Wai Leung. Specificity and complexity in bacterial quorum-sensing systems. **FEMS Microbiology Reviews**, *[S. l.]*, v. 40, n. 5, p. 738–752, 2016. DOI: 10.1093/femsre/fuw014.

HENKE, Jennifer M.; BASSLER, Bonnie L. Three parallel quorum-sensing systems regulate gene expression in *Vibrio harveyi*. **Journal of Bacteriology**, *[S. l.]*, v. 186, n. 20, p. 6902–6914, 2004. DOI: 10.1128/JB.186.20.6902-6914.2004.

HONG, Xutao et al. Metagenomic sequencing reveals the relationship between microbiota composition and quality of Chinese Rice Wine. **Scientific Reports**, *[S. l.]*, v. 6, n. November 2015, p. 1–11, 2016. DOI: 10.1038/srep26621. Available in: http://dx.doi.org/10.1038/srep26621.

JOHANSEN, Pernille; JESPERSEN, Lene. Impact of quorum sensing on the quality of fermented foods. **Current Opinion in Food Science**, *[S. l.]*, v. 13, p. 16–25, 2017. DOI: 10.1016/j.cofs.2017.01.001. Available in: http://dx.doi.org/10.1016/j.cofs.2017.01.001.

LEE, Andrew H. et al. A laboratory-scale model cocoa fermentation using dried, unfermented beans and artificial pulp can simulate the microbial and chemical changes of on-farm cocoa fermentation. **European Food Research and Technology**, *[S. l.]*, v. 245, n. 2, p. 511–519, 2019. DOI: 10.1007/s00217-018-3171-8. Available in: http://dx.doi.org/10.1007/s00217-018-3171-8.

LEFEBER, Timothy; PAPALEXANDRATOU, Zoi; GOBERT, William; CAMU, Nicholas; DE VUYST, Luc. On-farm implementation of a starter culture for improved cocoa bean fermentation and its influence on the flavour of chocolates produced thereof. **Food**

Microbiology, *[S. l.]*, v. 30, n. 2, p. 379–392, 2012. DOI: 10.1016/j.fm.2011.12.021. Available in: http://dx.doi.org/10.1016/j.fm.2011.12.021.

MAGALHÃES DA VEIGA MOREIRA, Igor; DE FIGUEIREDO VILELA, Leonardo; DA CRUZ PEDROSO MIGUEL, Maria Gabriela; SANTOS, Cledir; LIMA, Nelson; FREITAS SCHWAN, Rosane. Impact of a Microbial Cocktail Used as a Starter Culture on Cocoa Fermentation and Chocolate Flavor. **Molecules (Basel, Switzerland)**, *[S. l.]*, v. 22, n. 5, 2017. DOI: 10.3390/molecules22050766.

MANDABI, Aviad; GANIN, Hadas; MEIJLER, Michael M. Synergistic activation of quorum sensing in *Vibrio harveyi*. **Bioorganic and Medicinal Chemistry Letters**, *[S. l.]*, v. 25, n. 18, p. 3966–3969, 2015. DOI: 10.1016/j.bmcl.2015.07.028. Available in: http://dx.doi.org/10.1016/j.bmcl.2015.07.028.

MOENS, Frédéric; LEFEBER, Timothy; DE VUYST, Luc. Oxidation of metabolites highlights the microbial interactions and role of *Acetobacter pasteurianus* during cocoa bean fermentation. **Applied and Environmental Microbiology**, *[S. l.]*, v. 80, n. 6, p. 1848–1857, 2014. DOI: 10.1128/AEM.03344-13.

MOTA-GUTIERREZ, Jatziri; BARBOSA-PEREIRA, Letricia; FERROCINO, Ilario; COCOLIN, Luca. Traceability of functional volatile compounds generated on inoculated cocoa fermentation and its potential health benefits. **Nutrients**, *[S. l.]*, v. 11, n. 4, 2019. DOI: 10.3390/nu11040884.

ORDOÑEZ-ARAQUE, Roberto H.; LANDINES-VERA, Edgar F.; URRESTO-VILLEGAS, Julio C.; CAICEDO-JARAMILLO, Carla F. Microorganisms during cocoa fermentation: Systematic review. **Foods and Raw Materials**, *[S. l.]*, v. 8, n. 1, p. 155–162, 2020. DOI: 10.21603/2308-4057-2020-1-155-162.

OUATTARA, Honoré G.; ELIAS, Ryan J.; DUDLEY, Edward G. Microbial synergy between *Pichia kudriazevii* YS201 and *Bacillus subtilis* BS38 improves pulp degradation and aroma production in cocoa pulp simulation medium. **Heliyon**, *[S. l.]*, v. 6, n. 1, 2020. DOI: 10.1016/j.heliyon.2020.e03269.

PAPENFORT, Kai; BASSLER, Bonnie L. Quorum sensing signal-response systems in Gram-negative bacteria. **Nature Reviews Microbiology**, *[S. l.]*, v. 14, n. 9, p. 576–588, 2016. DOI: 10.1038/nrmicro.2016.89.

PARK, Hyunjoon; SHIN, Heuynkil; LEE, Kyuyeon; HOLZAPFEL, Wilhelm. Autoinducer-2 properties of kimchi are associated with lactic acid bacteria involved in its fermentation. **International Journal of Food Microbiology**, *[S. l.]*, v. 225, p. 38–42, 2016. DOI: 10.1016/j.ijfoodmicro.2016.03.007. Available in: http://dx.doi.org/10.1016/j.ijfoodmicro.2016.03.007.

PEREIRA, Gilberto Vinícius de Melo; SOCCOL, Vanete Thomaz; SOCCOL, Carlos Ricardo. Current state of research on cocoa and coffee fermentations. **Current Opinion in Food Science**, *[S. l.]*, v. 7, p. 50–57, 2016. DOI: 10.1016/j.cofs.2015.11.001.

PINTO, Uelinton M.; PAPPAS, Katherine M.; WINANS, Stephen C. The ABCs of plasmid replication and segregation. **Nature Reviews Microbiology**, *[S. l.]*, v. 10, n. 11, p. 755–765, 2012. DOI: 10.1038/nrmicro2882.

RUL, Françoise; MONNET, Véronique. How microbes communicate in food: A review of signaling molecules and their impact on food quality. **Current Opinion in Food Science**, *[S. l.]*, v. 2, p. 100–105, 2015. DOI: 10.1016/j.cofs.2015.03.003.

SAUNSHIA, Yallappa; SANDHYA, Mudra Kola Vidya Sagar; LINGAMALLU, Jagan Mohan Rao; PADELA, Janardhan; MURTHY, Pushpa. Improved Fermentation of Cocoa Beans with Enhanced Aroma Profiles. **Food Biotechnology**, *[S. l.]*, v. 32, n. 4, p. 257–272, 2018. DOI: 10.1080/08905436.2018.1519444. Available in: https://doi.org/10.1080/08905436.2018.1519444.

SCHUSTER, Martin; JOSEPH SEXTON, D.; DIGGLE, Stephen P.; PETER GREENBERG, E. Acyl-Homoserine Lactone Quorum Sensing: From Evolution to Application. **Annual Review of Microbiology**, *[S. l.]*, v. 67, n. 1, p. 43–63, 2013. DOI: 10.1146/annurev-micro-092412-155635. Available in: https://doi.org/10.1146/annurev-micro-092412-155635.

SCHWAN, Rosane F.; PEREIRA, Gilberto Vinícius de Melo; FLEET, Graham H. Microbial activities during cocoa fermentation. **Cocoa and coffee fermentations**, *[S. l.]*, n. January 2014, p. 129–92, 2014.

SMID, Eddy J.; LACROIX, Christophe. Microbe-microbe interactions in mixed culture food fermentations. **Current Opinion in Biotechnology**, *[S. l.]*, v. 24, n. 2, p. 148–

154, 2013. DOI: 10.1016/j.copbio.2012.11.007. Available in: http://dx.doi.org/10.1016/j.copbio.2012.11.007.

TUROVSKIY, Yevgeniy; CHIKINDAS, Michael L. Autoinducer-2 bioassay is a qualitative, not quantitative method influenced by glucose. **Journal of Microbiological Methods**, *[S. l.]*, v. 66, n. 3, p. 497–503, 2006. DOI: 10.1016/j.mimet.2006.02.001.

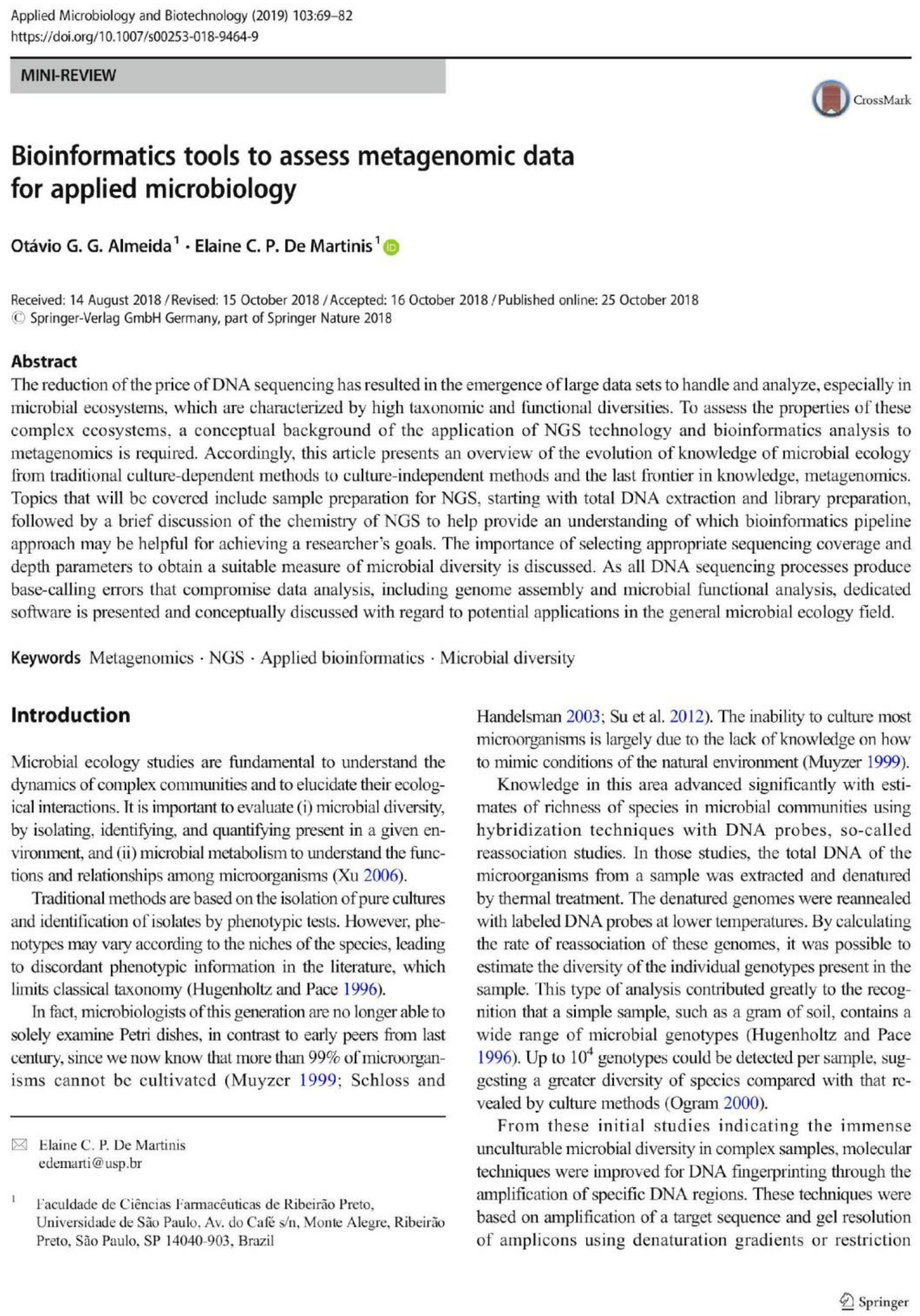
VISINTIN, Simonetta; RAMOS, Lacerda; BATISTA, Nara; DOLCI, Paola; SCHWAN, Freitas; COCOLIN, Luca. Impact of *Saccharomyces cerevisiae* and *Torulaspora delbrueckii* starter cultures on cocoa beans fermentation. **International Journal of Food Microbiology**, *[S. l.]*, v. 257, n. June, p. 31–40, 2017. DOI: 10.1016/j.ijfoodmicro.2017.06.004. Disponível em: http://dx.doi.org/10.1016/j.ijfoodmicro.2017.06.004.

WU, Liang; LUO, Yubin. Bacterial Quorum-Sensing Systems and Their Role in Intestinal Bacteria-Host Crosstalk. **Frontiers in Microbiology**, 12, 611413, 2021. DOI: 10.3389/fmicb.2021.611413.

Appendix A

Appendix A

This appendix presents a review article published on *Applied Microbiology and Biotechnology*. The importance of this article resides in the conceptual background it provided to design and choose the dedicated pipelines and strategies used in the presented research in this thesis. The Authors Rights license to reproduce this material is shown in the Attachment D.



Applied Microbiology and Biotechnology (2019) 103:69–82
https://doi.org/10.1007/s00253-018-9464-9

MINI-REVIEW

Bioinformatics tools to assess metagenomic data for applied microbiology

Otávio G. G. Almeida[1] · Elaine C. P. De Martinis[1]

Received: 14 August 2018 / Revised: 15 October 2018 / Accepted: 16 October 2018 / Published online: 25 October 2018

Abstract

The reduction of the price of DNA sequencing has resulted in the emergence of large data sets to handle and analyze, especially in microbial ecosystems, which are characterized by high taxonomic and functional diversities. To assess the properties of these complex ecosystems, a conceptual background of the application of NGS technology and bioinformatics analysis to metagenomics is required. Accordingly, this article presents an overview of the evolution of knowledge of microbial ecology from traditional culture-dependent methods to culture-independent methods and the last frontier in knowledge, metagenomics. Topics that will be covered include sample preparation for NGS, starting with total DNA extraction and library preparation, followed by a brief discussion of the chemistry of NGS to help provide an understanding of which bioinformatics pipeline approach may be helpful for achieving a researcher's goals. The importance of selecting appropriate sequencing coverage and depth parameters to obtain a suitable measure of microbial diversity is discussed. As all DNA sequencing processes produce base-calling errors that compromise data analysis, including genome assembly and microbial functional analysis, dedicated software is presented and conceptually discussed with regard to potential applications in the general microbial ecology field.

Keywords Metagenomics · NGS · Applied bioinformatics · Microbial diversity

Introduction

Microbial ecology studies are fundamental to understand the dynamics of complex communities and to elucidate their ecological interactions. It is important to evaluate (i) microbial diversity, by isolating, identifying, and quantifying present in a given environment, and (ii) microbial metabolism to understand the functions and relationships among microorganisms (Xu 2006).

Traditional methods are based on the isolation of pure cultures and identification of isolates by phenotypic tests. However, phenotypes may vary according to the niches of the species, leading to discordant phenotypic information in the literature, which limits classical taxonomy (Hugenholtz and Pace 1996).

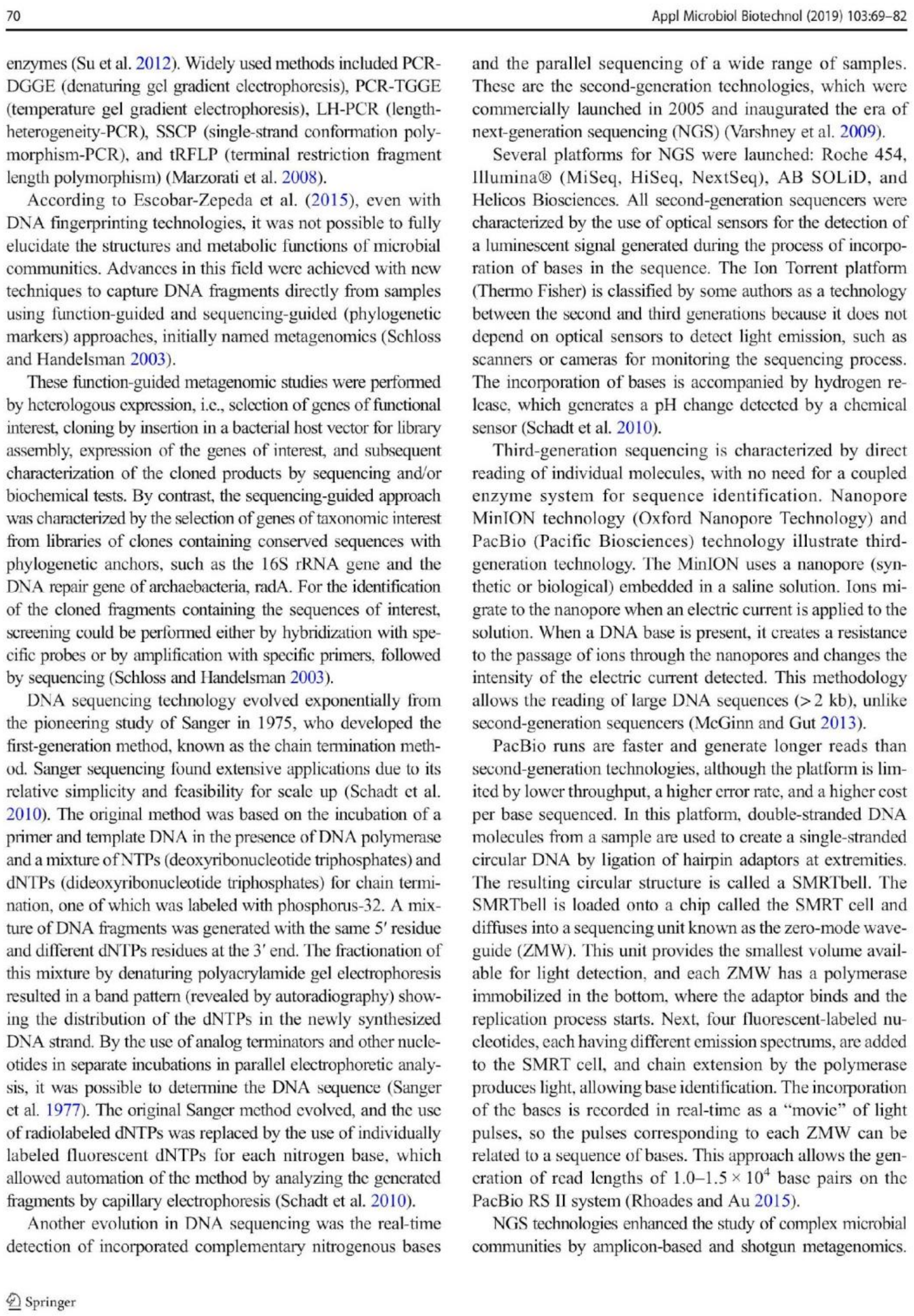
In fact, microbiologists of this generation are no longer able to solely examine Petri dishes, in contrast to early peers from last century, since we now know that more than 99% of microorganisms cannot be cultivated (Muyzer 1999; Schloss and Handelsman 2003; Su et al. 2012). The inability to culture most microorganisms is largely due to the lack of knowledge on how to mimic conditions of the natural environment (Muyzer 1999).

Knowledge in this area advanced significantly with estimates of richness of species in microbial communities using hybridization techniques with DNA probes, so-called reassociation studies. In those studies, the total DNA of the microorganisms from a sample was extracted and denatured by thermal treatment. The denatured genomes were reannealed with labeled DNA probes at lower temperatures. By calculating the rate of reassociation of these genomes, it was possible to estimate the diversity of the individual genotypes present in the sample. This type of analysis contributed greatly to the recognition that a simple sample, such as a gram of soil, contains a wide range of microbial genotypes (Hugenholtz and Pace 1996). Up to 10^4 genotypes could be detected per sample, suggesting a greater diversity of species compared with that revealed by culture methods (Ogram 2000).

From these initial studies indicating the immense unculturable microbial diversity in complex samples, molecular techniques were improved for DNA fingerprinting through the amplification of specific DNA regions. These techniques were based on amplification of a target sequence and gel resolution of amplicons using denaturation gradients or restriction

✉ Elaine C. P. De Martinis
edemarti@usp.br

[1] Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Av. do Café s/n, Monte Alegre, Ribeirão Preto, São Paulo, SP 14040-903, Brazil



enzymes (Su et al. 2012). Widely used methods included PCR-DGGE (denaturing gel gradient electrophoresis), PCR-TGGE (temperature gel gradient electrophoresis), LH-PCR (length-heterogeneity-PCR), SSCP (single-strand conformation polymorphism-PCR), and tRFLP (terminal restriction fragment length polymorphism) (Marzorati et al. 2008).

According to Escobar-Zepeda et al. (2015), even with DNA fingerprinting technologies, it was not possible to fully elucidate the structures and metabolic functions of microbial communities. Advances in this field were achieved with new techniques to capture DNA fragments directly from samples using function-guided and sequencing-guided (phylogenetic markers) approaches, initially named metagenomics (Schloss and Handelsman 2003).

These function-guided metagenomic studies were performed by heterologous expression, i.e., selection of genes of functional interest, cloning by insertion in a bacterial host vector for library assembly, expression of the genes of interest, and subsequent characterization of the cloned products by sequencing and/or biochemical tests. By contrast, the sequencing-guided approach was characterized by the selection of genes of taxonomic interest from libraries of clones containing conserved sequences with phylogenetic anchors, such as the 16S rRNA gene and the DNA repair gene of archaebacteria, radA. For the identification of the cloned fragments containing the sequences of interest, screening could be performed either by hybridization with specific probes or by amplification with specific primers, followed by sequencing (Schloss and Handelsman 2003).

DNA sequencing technology evolved exponentially from the pioneering study of Sanger in 1975, who developed the first-generation method, known as the chain termination method. Sanger sequencing found extensive applications due to its relative simplicity and feasibility for scale up (Schadt et al. 2010). The original method was based on the incubation of a primer and template DNA in the presence of DNA polymerase and a mixture of NTPs (deoxyribonucleotide triphosphates) and dNTPs (dideoxyribonucleotide triphosphates) for chain termination, one of which was labeled with phosphorus-32. A mixture of DNA fragments was generated with the same 5′ residue and different dNTPs residues at the 3′ end. The fractionation of this mixture by denaturing polyacrylamide gel electrophoresis resulted in a band pattern (revealed by autoradiography) showing the distribution of the dNTPs in the newly synthesized DNA strand. By the use of analog terminators and other nucleotides in separate incubations in parallel electrophoretic analysis, it was possible to determine the DNA sequence (Sanger et al. 1977). The original Sanger method evolved, and the use of radiolabeled dNTPs was replaced by the use of individually labeled fluorescent dNTPs for each nitrogen base, which allowed automation of the method by analyzing the generated fragments by capillary electrophoresis (Schadt et al. 2010).

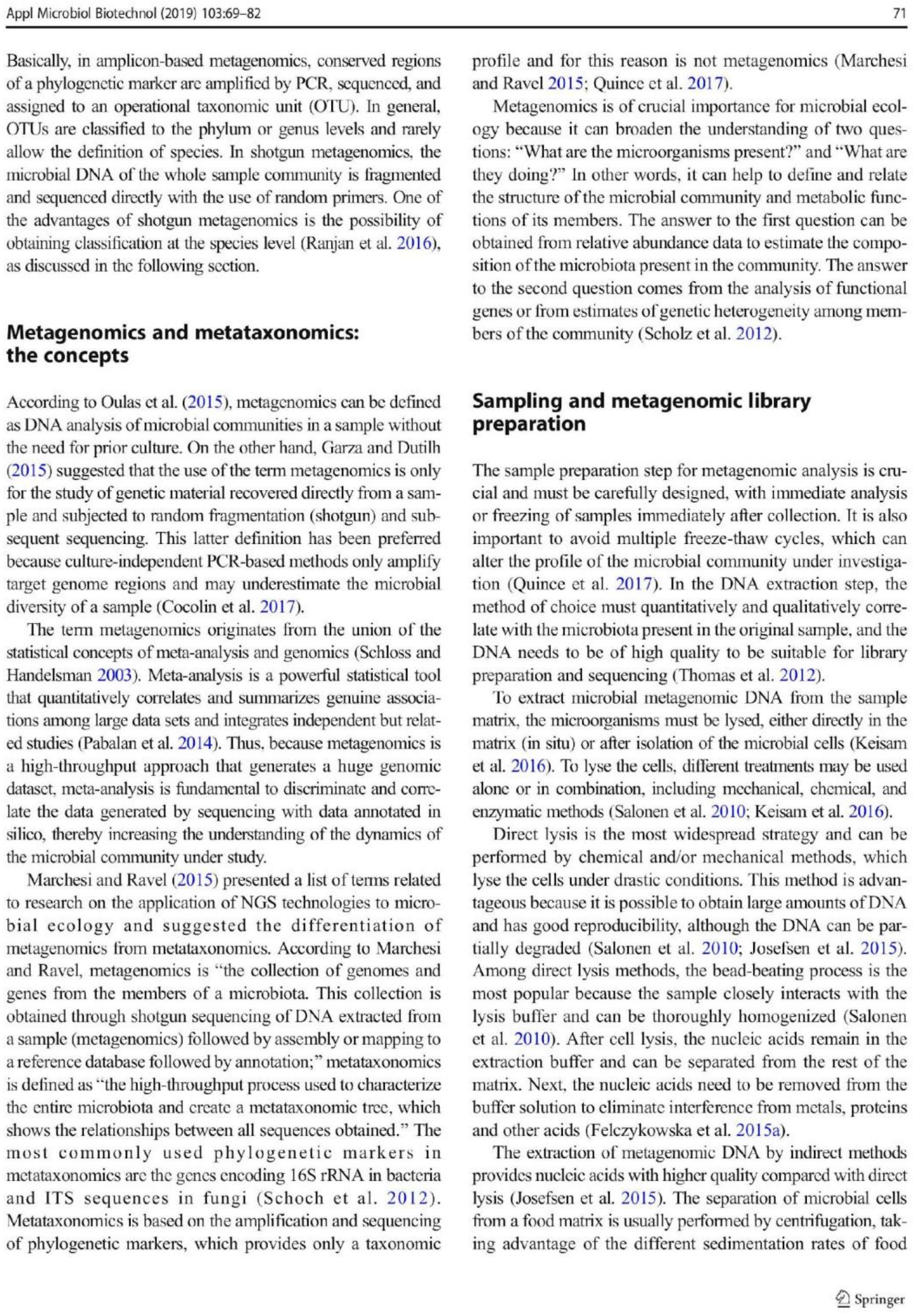
Another evolution in DNA sequencing was the real-time detection of incorporated complementary nitrogenous bases and the parallel sequencing of a wide range of samples. These are the second-generation technologies, which were commercially launched in 2005 and inaugurated the era of next-generation sequencing (NGS) (Varshney et al. 2009).

Several platforms for NGS were launched: Roche 454, Illumina® (MiSeq, HiSeq, NextSeq), AB SOLiD, and Helicos Biosciences. All second-generation sequencers were characterized by the use of optical sensors for the detection of a luminescent signal generated during the process of incorporation of bases in the sequence. The Ion Torrent platform (Thermo Fisher) is classified by some authors as a technology between the second and third generations because it does not depend on optical sensors to detect light emission, such as scanners or cameras for monitoring the sequencing process. The incorporation of bases is accompanied by hydrogen release, which generates a pH change detected by a chemical sensor (Schadt et al. 2010).

Third-generation sequencing is characterized by direct reading of individual molecules, with no need for a coupled enzyme system for sequence identification. Nanopore MinION technology (Oxford Nanopore Technology) and PacBio (Pacific Biosciences) technology illustrate third-generation technology. The MinION uses a nanopore (synthetic or biological) embedded in a saline solution. Ions migrate to the nanopore when an electric current is applied to the solution. When a DNA base is present, it creates a resistance to the passage of ions through the nanopores and changes the intensity of the electric current detected. This methodology allows the reading of large DNA sequences (> 2 kb), unlike second-generation sequencers (McGinn and Gut 2013).

PacBio runs are faster and generate longer reads than second-generation technologies, although the platform is limited by lower throughput, a higher error rate, and a higher cost per base sequenced. In this platform, double-stranded DNA molecules from a sample are used to create a single-stranded circular DNA by ligation of hairpin adaptors at extremities. The resulting circular structure is called a SMRTbell. The SMRTbell is loaded onto a chip called the SMRT cell and diffuses into a sequencing unit known as the zero-mode waveguide (ZMW). This unit provides the smallest volume available for light detection, and each ZMW has a polymerase immobilized in the bottom, where the adaptor binds and the replication process starts. Next, four fluorescent-labeled nucleotides, each having different emission spectrums, are added to the SMRT cell, and chain extension by the polymerase produces light, allowing base identification. The incorporation of the bases is recorded in real-time as a "movie" of light pulses, so the pulses corresponding to each ZMW can be related to a sequence of bases. This approach allows the generation of read lengths of $1.0–1.5 \times 10^4$ base pairs on the PacBio RS II system (Rhoades and Au 2015).

NGS technologies enhanced the study of complex microbial communities by amplicon-based and shotgun metagenomics.



Basically, in amplicon-based metagenomics, conserved regions of a phylogenetic marker are amplified by PCR, sequenced, and assigned to an operational taxonomic unit (OTU). In general, OTUs are classified to the phylum or genus levels and rarely allow the definition of species. In shotgun metagenomics, the microbial DNA of the whole sample community is fragmented and sequenced directly with the use of random primers. One of the advantages of shotgun metagenomics is the possibility of obtaining classification at the species level (Ranjan et al. 2016), as discussed in the following section.

Metagenomics and metataxonomics: the concepts

According to Oulas et al. (2015), metagenomics can be defined as DNA analysis of microbial communities in a sample without the need for prior culture. On the other hand, Garza and Dutilh (2015) suggested that the use of the term metagenomics is only for the study of genetic material recovered directly from a sample and subjected to random fragmentation (shotgun) and subsequent sequencing. This latter definition has been preferred because culture-independent PCR-based methods only amplify target genome regions and may underestimate the microbial diversity of a sample (Cocolin et al. 2017).

The term metagenomics originates from the union of the statistical concepts of meta-analysis and genomics (Schloss and Handelsman 2003). Meta-analysis is a powerful statistical tool that quantitatively correlates and summarizes genuine associations among large data sets and integrates independent but related studies (Pabalan et al. 2014). Thus, because metagenomics is a high-throughput approach that generates a huge genomic dataset, meta-analysis is fundamental to discriminate and correlate the data generated by sequencing with data annotated in silico, thereby increasing the understanding of the dynamics of the microbial community under study.

Marchesi and Ravel (2015) presented a list of terms related to research on the application of NGS technologies to microbial ecology and suggested the differentiation of metagenomics from metataxonomics. According to Marchesi and Ravel, metagenomics is "the collection of genomes and genes from the members of a microbiota. This collection is obtained through shotgun sequencing of DNA extracted from a sample (metagenomics) followed by assembly or mapping to a reference database followed by annotation;" metataxonomics is defined as "the high-throughput process used to characterize the entire microbiota and create a metataxonomic tree, which shows the relationships between all sequences obtained." The most commonly used phylogenetic markers in metataxonomics are the genes encoding 16S rRNA in bacteria and ITS sequences in fungi (Schoch et al. 2012). Metataxonomics is based on the amplification and sequencing of phylogenetic markers, which provides only a taxonomic profile and for this reason is not metagenomics (Marchesi and Ravel 2015; Quince et al. 2017).

Metagenomics is of crucial importance for microbial ecology because it can broaden the understanding of two questions: "What are the microorganisms present?" and "What are they doing?" In other words, it can help to define and relate the structure of the microbial community and metabolic functions of its members. The answer to the first question can be obtained from relative abundance data to estimate the composition of the microbiota present in the community. The answer to the second question comes from the analysis of functional genes or from estimates of genetic heterogeneity among members of the community (Scholz et al. 2012).

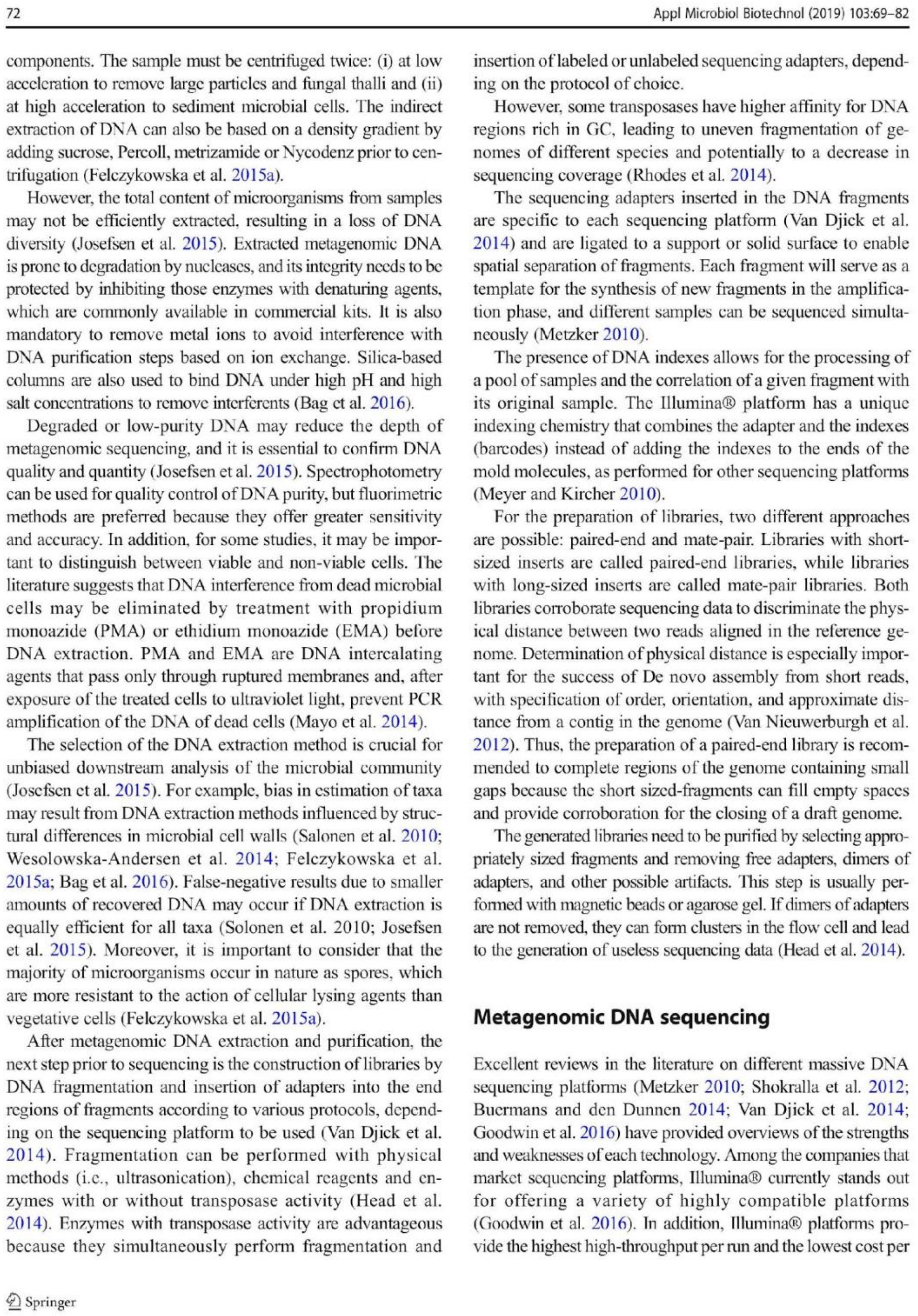
Sampling and metagenomic library preparation

The sample preparation step for metagenomic analysis is crucial and must be carefully designed, with immediate analysis or freezing of samples immediately after collection. It is also important to avoid multiple freeze-thaw cycles, which can alter the profile of the microbial community under investigation (Quince et al. 2017). In the DNA extraction step, the method of choice must quantitatively and qualitatively correlate with the microbiota present in the original sample, and the DNA needs to be of high quality to be suitable for library preparation and sequencing (Thomas et al. 2012).

To extract microbial metagenomic DNA from the sample matrix, the microorganisms must be lysed, either directly in the matrix (in situ) or after isolation of the microbial cells (Keisam et al. 2016). To lyse the cells, different treatments may be used alone or in combination, including mechanical, chemical, and enzymatic methods (Salonen et al. 2010; Keisam et al. 2016).

Direct lysis is the most widespread strategy and can be performed by chemical and/or mechanical methods, which lyse the cells under drastic conditions. This method is advantageous because it is possible to obtain large amounts of DNA and has good reproducibility, although the DNA can be partially degraded (Salonen et al. 2010; Josefsen et al. 2015). Among direct lysis methods, the bead-beating process is the most popular because the sample closely interacts with the lysis buffer and can be thoroughly homogenized (Salonen et al. 2010). After cell lysis, the nucleic acids remain in the extraction buffer and can be separated from the rest of the matrix. Next, the nucleic acids need to be removed from the buffer solution to eliminate interference from metals, proteins and other acids (Felczykowska et al. 2015a).

The extraction of metagenomic DNA by indirect methods provides nucleic acids with higher quality compared with direct lysis (Josefsen et al. 2015). The separation of microbial cells from a food matrix is usually performed by centrifugation, taking advantage of the different sedimentation rates of food



components. The sample must be centrifuged twice: (i) at low acceleration to remove large particles and fungal thalli and (ii) at high acceleration to sediment microbial cells. The indirect extraction of DNA can also be based on a density gradient by adding sucrose, Percoll, metrizamide or Nycodenz prior to centrifugation (Felczykowska et al. 2015a).

However, the total content of microorganisms from samples may not be efficiently extracted, resulting in a loss of DNA diversity (Josefsen et al. 2015). Extracted metagenomic DNA is prone to degradation by nucleases, and its integrity needs to be protected by inhibiting those enzymes with denaturing agents, which are commonly available in commercial kits. It is also mandatory to remove metal ions to avoid interference with DNA purification steps based on ion exchange. Silica-based columns are also used to bind DNA under high pH and high salt concentrations to remove interferents (Bag et al. 2016).

Degraded or low-purity DNA may reduce the depth of metagenomic sequencing, and it is essential to confirm DNA quality and quantity (Josefsen et al. 2015). Spectrophotometry can be used for quality control of DNA purity, but fluorimetric methods are preferred because they offer greater sensitivity and accuracy. In addition, for some studies, it may be important to distinguish between viable and non-viable cells. The literature suggests that DNA interference from dead microbial cells may be eliminated by treatment with propidium monoazide (PMA) or ethidium monoazide (EMA) before DNA extraction. PMA and EMA are DNA intercalating agents that pass only through ruptured membranes and, after exposure of the treated cells to ultraviolet light, prevent PCR amplification of the DNA of dead cells (Mayo et al. 2014).

The selection of the DNA extraction method is crucial for unbiased downstream analysis of the microbial community (Josefsen et al. 2015). For example, bias in estimation of taxa may result from DNA extraction methods influenced by structural differences in microbial cell walls (Salonen et al. 2010; Wesolowska-Andersen et al. 2014; Felczykowska et al. 2015a; Bag et al. 2016). False-negative results due to smaller amounts of recovered DNA may occur if DNA extraction is equally efficient for all taxa (Solonen et al. 2010; Josefsen et al. 2015). Moreover, it is important to consider that the majority of microorganisms occur in nature as spores, which are more resistant to the action of cellular lysing agents than vegetative cells (Felczykowska et al. 2015a).

After metagenomic DNA extraction and purification, the next step prior to sequencing is the construction of libraries by DNA fragmentation and insertion of adapters into the end regions of fragments according to various protocols, depending on the sequencing platform to be used (Van Djick et al. 2014). Fragmentation can be performed with physical methods (i.e., ultrasonication), chemical reagents and enzymes with or without transposase activity (Head et al. 2014). Enzymes with transposase activity are advantageous because they simultaneously perform fragmentation and insertion of labeled or unlabeled sequencing adapters, depending on the protocol of choice.

However, some transposases have higher affinity for DNA regions rich in GC, leading to uneven fragmentation of genomes of different species and potentially to a decrease in sequencing coverage (Rhodes et al. 2014).

The sequencing adapters inserted in the DNA fragments are specific to each sequencing platform (Van Djick et al. 2014) and are ligated to a support or solid surface to enable spatial separation of fragments. Each fragment will serve as a template for the synthesis of new fragments in the amplification phase, and different samples can be sequenced simultaneously (Metzker 2010).

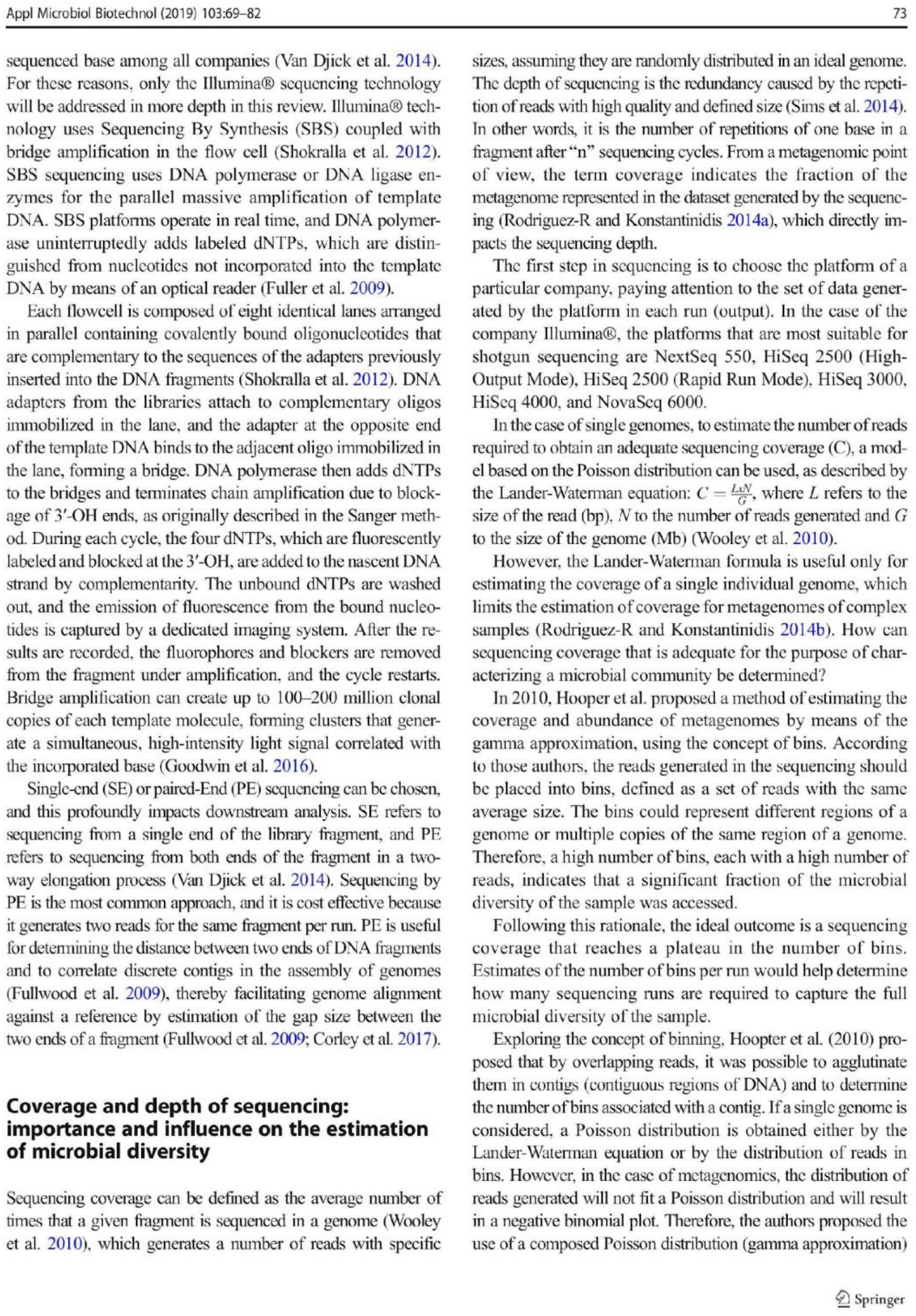
The presence of DNA indexes allows for the processing of a pool of samples and the correlation of a given fragment with its original sample. The Illumina® platform has a unique indexing chemistry that combines the adapter and the indexes (barcodes) instead of adding the indexes to the ends of the mold molecules, as performed for other sequencing platforms (Meyer and Kircher 2010).

For the preparation of libraries, two different approaches are possible: paired-end and mate-pair. Libraries with short-sized inserts are called paired-end libraries, while libraries with long-sized inserts are called mate-pair libraries. Both libraries corroborate sequencing data to discriminate the physical distance between two reads aligned in the reference genome. Determination of physical distance is especially important for the success of De novo assembly from short reads, with specification of order, orientation, and approximate distance from a contig in the genome (Van Nieuwerburgh et al. 2012). Thus, the preparation of a paired-end library is recommended to complete regions of the genome containing small gaps because the short sized-fragments can fill empty spaces and provide corroboration for the closing of a draft genome.

The generated libraries need to be purified by selecting appropriately sized fragments and removing free adapters, dimers of adapters, and other possible artifacts. This step is usually performed with magnetic beads or agarose gel. If dimers of adapters are not removed, they can form clusters in the flow cell and lead to the generation of useless sequencing data (Head et al. 2014).

Metagenomic DNA sequencing

Excellent reviews in the literature on different massive DNA sequencing platforms (Metzker 2010; Shokralla et al. 2012; Buermans and den Dunnen 2014; Van Djick et al. 2014; Goodwin et al. 2016) have provided overviews of the strengths and weaknesses of each technology. Among the companies that market sequencing platforms, Illumina® currently stands out for offering a variety of highly compatible platforms (Goodwin et al. 2016). In addition, Illumina® platforms provide the highest high-throughput per run and the lowest cost per



sequenced base among all companies (Van Djick et al. 2014). For these reasons, only the Illumina® sequencing technology will be addressed in more depth in this review. Illumina® technology uses Sequencing By Synthesis (SBS) coupled with bridge amplification in the flow cell (Shokralla et al. 2012). SBS sequencing uses DNA polymerase or DNA ligase enzymes for the parallel massive amplification of template DNA. SBS platforms operate in real time, and DNA polymerase uninterruptedly adds labeled dNTPs, which are distinguished from nucleotides not incorporated into the template DNA by means of an optical reader (Fuller et al. 2009).

Each flowcell is composed of eight identical lanes arranged in parallel containing covalently bound oligonucleotides that are complementary to the sequences of the adapters previously inserted into the DNA fragments (Shokralla et al. 2012). DNA adapters from the libraries attach to complementary oligos immobilized in the lane, and the adapter at the opposite end of the template DNA binds to the adjacent oligo immobilized in the lane, forming a bridge. DNA polymerase then adds dNTPs to the bridges and terminates chain amplification due to blockage of 3′-OH ends, as originally described in the Sanger method. During each cycle, the four dNTPs, which are fluorescently labeled and blocked at the 3′-OH, are added to the nascent DNA strand by complementarity. The unbound dNTPs are washed out, and the emission of fluorescence from the bound nucleotides is captured by a dedicated imaging system. After the results are recorded, the fluorophores and blockers are removed from the fragment under amplification, and the cycle restarts. Bridge amplification can create up to 100–200 million clonal copies of each template molecule, forming clusters that generate a simultaneous, high-intensity light signal correlated with the incorporated base (Goodwin et al. 2016).

Single-end (SE) or paired-End (PE) sequencing can be chosen, and this profoundly impacts downstream analysis. SE refers to sequencing from a single end of the library fragment, and PE refers to sequencing from both ends of the fragment in a two-way elongation process (Van Djick et al. 2014). Sequencing by PE is the most common approach, and it is cost effective because it generates two reads for the same fragment per run. PE is useful for determining the distance between two ends of DNA fragments and to correlate discrete contigs in the assembly of genomes (Fullwood et al. 2009), thereby facilitating genome alignment against a reference by estimation of the gap size between the two ends of a fragment (Fullwood et al. 2009; Corley et al. 2017).

Coverage and depth of sequencing: importance and influence on the estimation of microbial diversity

Sequencing coverage can be defined as the average number of times that a given fragment is sequenced in a genome (Wooley et al. 2010), which generates a number of reads with specific sizes, assuming they are randomly distributed in an ideal genome. The depth of sequencing is the redundancy caused by the repetition of reads with high quality and defined size (Sims et al. 2014). In other words, it is the number of repetitions of one base in a fragment after "n" sequencing cycles. From a metagenomic point of view, the term coverage indicates the fraction of the metagenome represented in the dataset generated by the sequencing (Rodriguez-R and Konstantinidis 2014a), which directly impacts the sequencing depth.

The first step in sequencing is to choose the platform of a particular company, paying attention to the set of data generated by the platform in each run (output). In the case of the company Illumina®, the platforms that are most suitable for shotgun sequencing are NextSeq 550, HiSeq 2500 (High-Output Mode), HiSeq 2500 (Rapid Run Mode), HiSeq 3000, HiSeq 4000, and NovaSeq 6000.

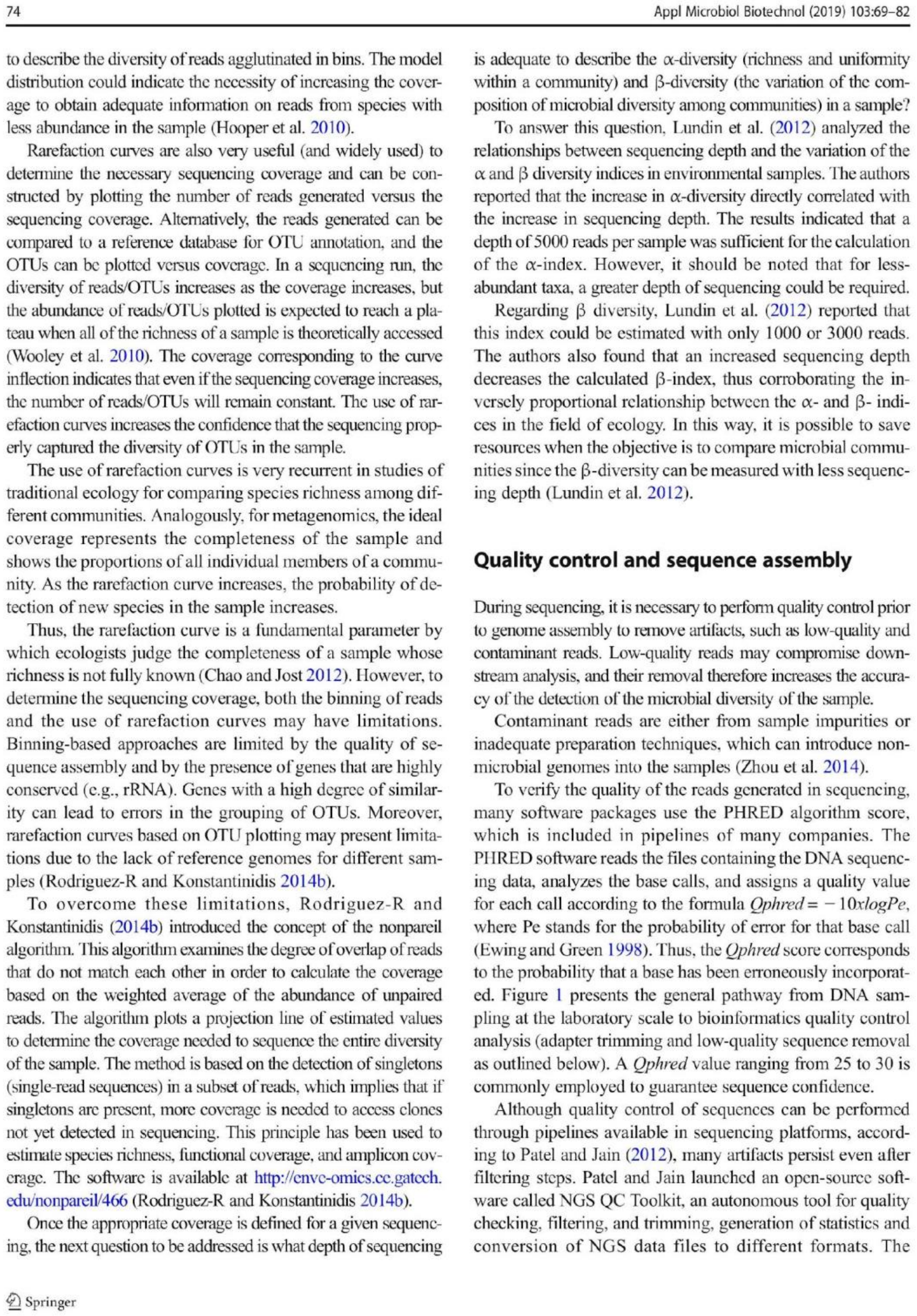
In the case of single genomes, to estimate the number of reads required to obtain an adequate sequencing coverage (C), a model based on the Poisson distribution can be used, as described by the Lander-Waterman equation: $C = \frac{LxN}{G}$, where L refers to the size of the read (bp), N to the number of reads generated and G to the size of the genome (Mb) (Wooley et al. 2010).

However, the Lander-Waterman formula is useful only for estimating the coverage of a single individual genome, which limits the estimation of coverage for metagenomes of complex samples (Rodriguez-R and Konstantinidis 2014b). How can sequencing coverage that is adequate for the purpose of characterizing a microbial community be determined?

In 2010, Hooper et al. proposed a method of estimating the coverage and abundance of metagenomes by means of the gamma approximation, using the concept of bins. According to those authors, the reads generated in the sequencing should be placed into bins, defined as a set of reads with the same average size. The bins could represent different regions of a genome or multiple copies of the same region of a genome. Therefore, a high number of bins, each with a high number of reads, indicates that a significant fraction of the microbial diversity of the sample was accessed.

Following this rationale, the ideal outcome is a sequencing coverage that reaches a plateau in the number of bins. Estimates of the number of bins per run would help determine how many sequencing runs are required to capture the full microbial diversity of the sample.

Exploring the concept of binning, Hoopter et al. (2010) proposed that by overlapping reads, it was possible to agglutinate them in contigs (contiguous regions of DNA) and to determine the number of bins associated with a contig. If a single genome is considered, a Poisson distribution is obtained either by the Lander-Waterman equation or by the distribution of reads in bins. However, in the case of metagenomics, the distribution of reads generated will not fit a Poisson distribution and will result in a negative binomial plot. Therefore, the authors proposed the use of a composed Poisson distribution (gamma approximation)



to describe the diversity of reads agglutinated in bins. The model distribution could indicate the necessity of increasing the coverage to obtain adequate information on reads from species with less abundance in the sample (Hooper et al. 2010).

Rarefaction curves are also very useful (and widely used) to determine the necessary sequencing coverage and can be constructed by plotting the number of reads generated versus the sequencing coverage. Alternatively, the reads generated can be compared to a reference database for OTU annotation, and the OTUs can be plotted versus coverage. In a sequencing run, the diversity of reads/OTUs increases as the coverage increases, but the abundance of reads/OTUs plotted is expected to reach a plateau when all of the richness of a sample is theoretically accessed (Wooley et al. 2010). The coverage corresponding to the curve inflection indicates that even if the sequencing coverage increases, the number of reads/OTUs will remain constant. The use of rarefaction curves increases the confidence that the sequencing properly captured the diversity of OTUs in the sample.

The use of rarefaction curves is very recurrent in studies of traditional ecology for comparing species richness among different communities. Analogously, for metagenomics, the ideal coverage represents the completeness of the sample and shows the proportions of all individual members of a community. As the rarefaction curve increases, the probability of detection of new species in the sample increases.

Thus, the rarefaction curve is a fundamental parameter by which ecologists judge the completeness of a sample whose richness is not fully known (Chao and Jost 2012). However, to determine the sequencing coverage, both the binning of reads and the use of rarefaction curves may have limitations. Binning-based approaches are limited by the quality of sequence assembly and by the presence of genes that are highly conserved (e.g., rRNA). Genes with a high degree of similarity can lead to errors in the grouping of OTUs. Moreover, rarefaction curves based on OTU plotting may present limitations due to the lack of reference genomes for different samples (Rodriguez-R and Konstantinidis 2014b).

To overcome these limitations, Rodriguez-R and Konstantinidis (2014b) introduced the concept of the nonpareil algorithm. This algorithm examines the degree of overlap of reads that do not match each other in order to calculate the coverage based on the weighted average of the abundance of unpaired reads. The algorithm plots a projection line of estimated values to determine the coverage needed to sequence the entire diversity of the sample. The method is based on the detection of singletons (single-read sequences) in a subset of reads, which implies that if singletons are present, more coverage is needed to access clones not yet detected in sequencing. This principle has been used to estimate species richness, functional coverage, and amplicon coverage. The software is available at http://enve-omics.ce.gatech.edu/nonpareil/466 (Rodriguez-R and Konstantinidis 2014b).

Once the appropriate coverage is defined for a given sequencing, the next question to be addressed is what depth of sequencing is adequate to describe the α-diversity (richness and uniformity within a community) and β-diversity (the variation of the composition of microbial diversity among communities) in a sample?

To answer this question, Lundin et al. (2012) analyzed the relationships between sequencing depth and the variation of the α and β diversity indices in environmental samples. The authors reported that the increase in α-diversity directly correlated with the increase in sequencing depth. The results indicated that a depth of 5000 reads per sample was sufficient for the calculation of the α-index. However, it should be noted that for less-abundant taxa, a greater depth of sequencing could be required.

Regarding β diversity, Lundin et al. (2012) reported that this index could be estimated with only 1000 or 3000 reads. The authors also found that an increased sequencing depth decreases the calculated β-index, thus corroborating the inversely proportional relationship between the α- and β- indices in the field of ecology. In this way, it is possible to save resources when the objective is to compare microbial communities since the β-diversity can be measured with less sequencing depth (Lundin et al. 2012).

Quality control and sequence assembly

During sequencing, it is necessary to perform quality control prior to genome assembly to remove artifacts, such as low-quality and contaminant reads. Low-quality reads may compromise downstream analysis, and their removal therefore increases the accuracy of the detection of the microbial diversity of the sample.

Contaminant reads are either from sample impurities or inadequate preparation techniques, which can introduce non-microbial genomes into the samples (Zhou et al. 2014).

To verify the quality of the reads generated in sequencing, many software packages use the PHRED algorithm score, which is included in pipelines of many companies. The PHRED software reads the files containing the DNA sequencing data, analyzes the base calls, and assigns a quality value for each call according to the formula $Qphred = -10xlogPe$, where Pe stands for the probability of error for that base call (Ewing and Green 1998). Thus, the *Qphred* score corresponds to the probability that a base has been erroneously incorporated. Figure 1 presents the general pathway from DNA sampling at the laboratory scale to bioinformatics quality control analysis (adapter trimming and low-quality sequence removal as outlined below). A *Qphred* value ranging from 25 to 30 is commonly employed to guarantee sequence confidence.

Although quality control of sequences can be performed through pipelines available in sequencing platforms, according to Patel and Jain (2012), many artifacts persist even after filtering steps. Patel and Jain launched an open-source software called NGS QC Toolkit, an autonomous tool for quality checking, filtering, and trimming, generation of statistics and conversion of NGS data files to different formats. The

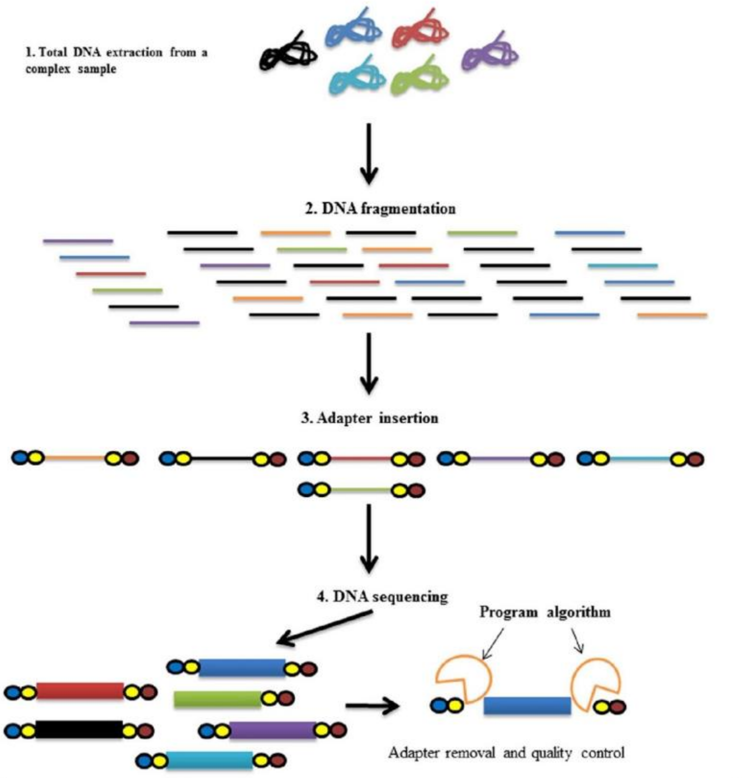

1. Total DNA extraction from a complex sample
2. DNA fragmentation
3. Adapter insertion
4. DNA sequencing
Program algorithm
Adapter removal and quality control

Fig. 1 Data acquisition. (1) First, the total DNA extracted directly from a complex sample (2) must be fragmented and processed in a step called library preparation. This step is crucial since the libraries are sequencing-platform specific and allow several different samples derived from distinct sampling points to be pooled due to the insertion of adapters with barcodes (short sequences represented in red and yellow, respectively). After library preparation, (3) next-generation DNA sequencing (NGS) can be performed, and (4) the generated data can be parsed to unravel information regarding the taxonomic and functional content of the microbial community. The bioinformatics analysis starts successfully after the sequencing data pass through rigorous quality control, which is characterized by the trimming of adapters inserted during library preparation and the removal of reads possessing low-quality base calls identified according to the *Qphred* score value for each sequenced base. The *Qphred* score is very useful because higher values indicate lower error probability of random base calling

Quality control

```
> COMMAND LINE

> Trimming adapter (i.e. Cutadapt)
> Remove all base calling in which Q < 30 (i.e. NGS QC Toolkit)
```

Qphred values and base calling confidence

Q	Pe	Confidence (%)
30	0,001	99,999
20	0,01	99,99
10	0,1	99,9

software is user-friendly and permits the processing of a large number of data sequences in parallel using the quality criteria of the PHRED score (Patel and Jain 2012).

Sequences from adapters must also be removed, and this step can be performed with the Cutadapt software, which has a user-friendly command line to perform adapter trimming. The software is written in the Python language and was developed in the Ubuntu Linux operating system, which is also compatible with the Windows and Mac OS X operating systems. The concept of this algorithm is the alignment of the reads with all adapter sequences, depending on the sequencing platform. The algorithm penalizes alignments in which adapter sequences are aligned with the 3′ region of reads, and thus all sequences from the adapters are removed. On the other hand, if the adapter sequence is overlapped at the beginning of the read, the sequences prior to the overlap are removed (Martin 2011).

The resulting high-quality reads must be agglutinated in large contiguous segments with the aid of genome assembly tools. Most of these tools are based on De novo assembly (van der Walt et al. 2017). De novo assembly is defined as the reconstruction of genomes in pure form without consulting databases (Miller et al. 2010). Another strategy for genome assembly is based on a reference genome for annotation of microorganisms already deposited in databases.

In general, in De novo assembly, the sequences are divided into pre-defined segments of size k (*k-mers*), which are overlapped to form a network of overlapping paths that interactively form the contigs (van der Walt et al. 2017). This is the basis of *de*

Bruijn graphs for the assembly of genomes from short reads (Miller et al. 2010). The advantage of De novo assembly is the discovery and reconstruction of new genomes, which make gene prediction more reliable (van der Walt et al. 2017).

The assembly of genomes is a hierarchical structuring that maps the data sequences into a putative reconstruction of the target. Genome assembly groups reads into contigs and contigs into scaffolds. Contigs allow for multiple sequence alignment of reads relative to a consensus sequence, whereas scaffolds list the order and orientation of the contigs and the size of the gaps between contigs (Fig. 2). These sequence clusters are important metrics for the quality of genome assembly because the assemblies are measured according to the size of the contigs and scaffolds generated. An important parameter for the comparison of genome assemblies is the N50, which refers to the size of the smallest contig in a set of contigs that represents at least 50% of the assembly (Miller et al. 2010).

The assembly of metagenomes should be performed with proper software since assemblers of unique genomes (i.e., SOAPdenovo, Velvet, EULER, and abyss) are not suitable for this purpose. Two factors limit the use of single genome assemblers: (i) the presence of polymorphisms among subspecies and of genomic regions conserved in different species and (ii) the uneven abundance of species in the sample. The presence of

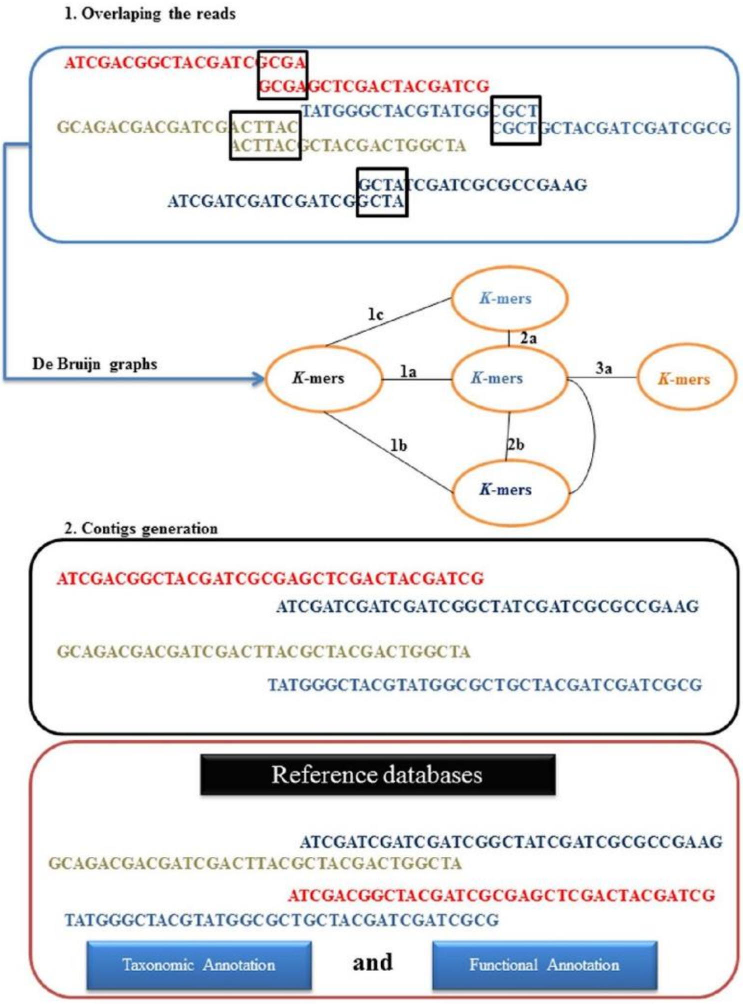
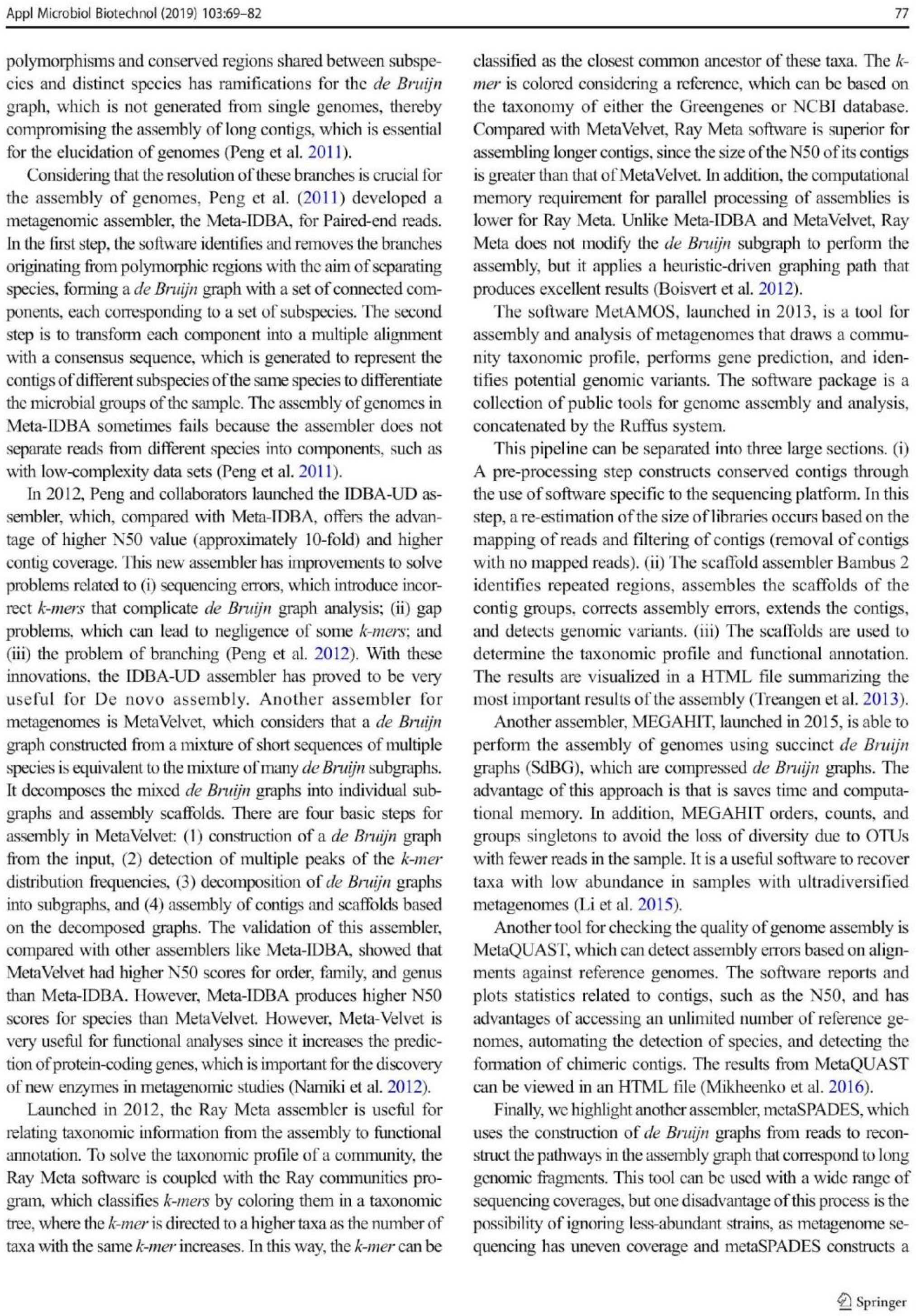


Fig. 2 Genome assembly and annotation. Algorithms based on *de Bruijn* graphs are applied to overlapping reads with similarity to achieve a contiguous configuration. As many sequences reads are derived from a complex sample, these overlaps among complementary reads are represented by *de Bruijn* graph pathways. (1) Overlapping reads: a magnified view of the read overlapping process, in which similar sequences of reads are o-verlapped until assembly into a higher sequence (contigs). As several genomes are present, many branches are expected in a *de Bruijn* graph, and following the pathway, the assembler identifies a large sequence. For instance, imagine that sequence 1a overlaps with sequence 2a, which overlaps with 2b, which overlaps again with 2a. Finally, this large overlapped sequence overlaps with sequence 3a (other scenarios may be possible if all pathways involved are considered). Therefore, the assembler identifies a contiguous sequence: 1a → 2a → 2b → 2a → 3a. This is the rationale of genome assembly. (2) The results of the overlapped reads can be aligned against a dedicated database for taxonomic and functional annotation



polymorphisms and conserved regions shared between subspecies and distinct species has ramifications for the *de Bruijn* graph, which is not generated from single genomes, thereby compromising the assembly of long contigs, which is essential for the elucidation of genomes (Peng et al. 2011).

Considering that the resolution of these branches is crucial for the assembly of genomes, Peng et al. (2011) developed a metagenomic assembler, the Meta-IDBA, for Paired-end reads. In the first step, the software identifies and removes the branches originating from polymorphic regions with the aim of separating species, forming a *de Bruijn* graph with a set of connected components, each corresponding to a set of subspecies. The second step is to transform each component into a multiple alignment with a consensus sequence, which is generated to represent the contigs of different subspecies of the same species to differentiate the microbial groups of the sample. The assembly of genomes in Meta-IDBA sometimes fails because the assembler does not separate reads from different species into components, such as with low-complexity data sets (Peng et al. 2011).

In 2012, Peng and collaborators launched the IDBA-UD assembler, which, compared with Meta-IDBA, offers the advantage of higher N50 value (approximately 10-fold) and higher contig coverage. This new assembler has improvements to solve problems related to (i) sequencing errors, which introduce incorrect *k-mers* that complicate *de Bruijn* graph analysis; (ii) gap problems, which can lead to negligence of some *k-mers*; and (iii) the problem of branching (Peng et al. 2012). With these innovations, the IDBA-UD assembler has proved to be very useful for De novo assembly. Another assembler for metagenomes is MetaVelvet, which considers that a *de Bruijn* graph constructed from a mixture of short sequences of multiple species is equivalent to the mixture of many *de Bruijn* subgraphs. It decomposes the mixed *de Bruijn* graphs into individual subgraphs and assembly scaffolds. There are four basic steps for assembly in MetaVelvet: (1) construction of a *de Bruijn* graph from the input, (2) detection of multiple peaks of the *k-mer* distribution frequencies, (3) decomposition of *de Bruijn* graphs into subgraphs, and (4) assembly of contigs and scaffolds based on the decomposed graphs. The validation of this assembler, compared with other assemblers like Meta-IDBA, showed that MetaVelvet had higher N50 scores for order, family, and genus than Meta-IDBA. However, Meta-IDBA produces higher N50 scores for species than MetaVelvet. However, Meta-Velvet is very useful for functional analyses since it increases the prediction of protein-coding genes, which is important for the discovery of new enzymes in metagenomic studies (Namiki et al. 2012).

Launched in 2012, the Ray Meta assembler is useful for relating taxonomic information from the assembly to functional annotation. To solve the taxonomic profile of a community, the Ray Meta software is coupled with the Ray communities program, which classifies *k-mers* by coloring them in a taxonomic tree, where the *k-mer* is directed to a higher taxa as the number of taxa with the same *k-mer* increases. In this way, the *k-mer* can be classified as the closest common ancestor of these taxa. The *k-mer* is colored considering a reference, which can be based on the taxonomy of either the Greengenes or NCBI database. Compared with MetaVelvet, Ray Meta software is superior for assembling longer contigs, since the size of the N50 of its contigs is greater than that of MetaVelvet. In addition, the computational memory requirement for parallel processing of assemblies is lower for Ray Meta. Unlike Meta-IDBA and MetaVelvet, Ray Meta does not modify the *de Bruijn* subgraph to perform the assembly, but it applies a heuristic-driven graphing path that produces excellent results (Boisvert et al. 2012).

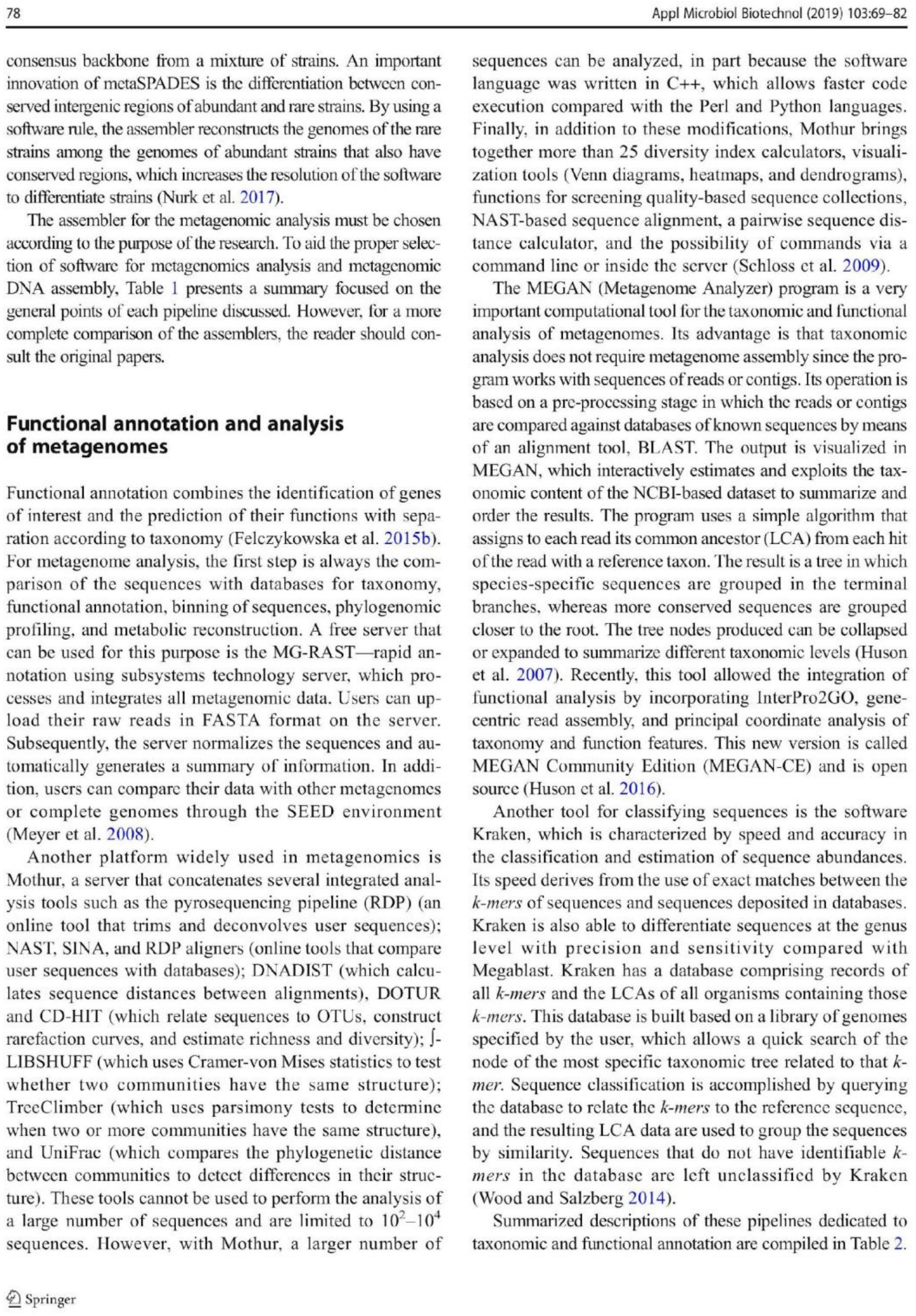
The software MetAMOS, launched in 2013, is a tool for assembly and analysis of metagenomes that draws a community taxonomic profile, performs gene prediction, and identifies potential genomic variants. The software package is a collection of public tools for genome assembly and analysis, concatenated by the Ruffus system.

This pipeline can be separated into three large sections. (i) A pre-processing step constructs conserved contigs through the use of software specific to the sequencing platform. In this step, a re-estimation of the size of libraries occurs based on the mapping of reads and filtering of contigs (removal of contigs with no mapped reads). (ii) The scaffold assembler Bambus 2 identifies repeated regions, assembles the scaffolds of the contig groups, corrects assembly errors, extends the contigs, and detects genomic variants. (iii) The scaffolds are used to determine the taxonomic profile and functional annotation. The results are visualized in a HTML file summarizing the most important results of the assembly (Treangen et al. 2013).

Another assembler, MEGAHIT, launched in 2015, is able to perform the assembly of genomes using succinct *de Bruijn* graphs (SdBG), which are compressed *de Bruijn* graphs. The advantage of this approach is that is saves time and computational memory. In addition, MEGAHIT orders, counts, and groups singletons to avoid the loss of diversity due to OTUs with fewer reads in the sample. It is a useful software to recover taxa with low abundance in samples with ultradiversified metagenomes (Li et al. 2015).

Another tool for checking the quality of genome assembly is MetaQUAST, which can detect assembly errors based on alignments against reference genomes. The software reports and plots statistics related to contigs, such as the N50, and has advantages of accessing an unlimited number of reference genomes, automating the detection of species, and detecting the formation of chimeric contigs. The results from MetaQUAST can be viewed in an HTML file (Mikheenko et al. 2016).

Finally, we highlight another assembler, metaSPADES, which uses the construction of *de Bruijn* graphs from reads to reconstruct the pathways in the assembly graph that correspond to long genomic fragments. This tool can be used with a wide range of sequencing coverages, but one disadvantage of this process is the possibility of ignoring less-abundant strains, as metagenome sequencing has uneven coverage and metaSPADES constructs a



consensus backbone from a mixture of strains. An important innovation of metaSPADES is the differentiation between conserved intergenic regions of abundant and rare strains. By using a software rule, the assembler reconstructs the genomes of the rare strains among the genomes of abundant strains that also have conserved regions, which increases the resolution of the software to differentiate strains (Nurk et al. 2017).

The assembler for the metagenomic analysis must be chosen according to the purpose of the research. To aid the proper selection of software for metagenomics analysis and metagenomic DNA assembly, Table 1 presents a summary focused on the general points of each pipeline discussed. However, for a more complete comparison of the assemblers, the reader should consult the original papers.

Functional annotation and analysis of metagenomes

Functional annotation combines the identification of genes of interest and the prediction of their functions with separation according to taxonomy (Felczykowska et al. 2015b). For metagenome analysis, the first step is always the comparison of the sequences with databases for taxonomy, functional annotation, binning of sequences, phylogenomic profiling, and metabolic reconstruction. A free server that can be used for this purpose is the MG-RAST—rapid annotation using subsystems technology server, which processes and integrates all metagenomic data. Users can upload their raw reads in FASTA format on the server. Subsequently, the server normalizes the sequences and automatically generates a summary of information. In addition, users can compare their data with other metagenomes or complete genomes through the SEED environment (Meyer et al. 2008).

Another platform widely used in metagenomics is Mothur, a server that concatenates several integrated analysis tools such as the pyrosequencing pipeline (RDP) (an online tool that trims and deconvolves user sequences); NAST, SINA, and RDP aligners (online tools that compare user sequences with databases); DNADIST (which calculates sequence distances between alignments), DOTUR and CD-HIT (which relate sequences to OTUs, construct rarefaction curves, and estimate richness and diversity); ∫-LIBSHUFF (which uses Cramer-von Mises statistics to test whether two communities have the same structure); TreeClimber (which uses parsimony tests to determine when two or more communities have the same structure), and UniFrac (which compares the phylogenetic distance between communities to detect differences in their structure). These tools cannot be used to perform the analysis of a large number of sequences and are limited to 10^2–10^4 sequences. However, with Mothur, a larger number of sequences can be analyzed, in part because the software language was written in C++, which allows faster code execution compared with the Perl and Python languages. Finally, in addition to these modifications, Mothur brings together more than 25 diversity index calculators, visualization tools (Venn diagrams, heatmaps, and dendrograms), functions for screening quality-based sequence collections, NAST-based sequence alignment, a pairwise sequence distance calculator, and the possibility of commands via a command line or inside the server (Schloss et al. 2009).

The MEGAN (Metagenome Analyzer) program is a very important computational tool for the taxonomic and functional analysis of metagenomes. Its advantage is that taxonomic analysis does not require metagenome assembly since the program works with sequences of reads or contigs. Its operation is based on a pre-processing stage in which the reads or contigs are compared against databases of known sequences by means of an alignment tool, BLAST. The output is visualized in MEGAN, which interactively estimates and exploits the taxonomic content of the NCBI-based dataset to summarize and order the results. The program uses a simple algorithm that assigns to each read its common ancestor (LCA) from each hit of the read with a reference taxon. The result is a tree in which species-specific sequences are grouped in the terminal branches, whereas more conserved sequences are grouped closer to the root. The tree nodes produced can be collapsed or expanded to summarize different taxonomic levels (Huson et al. 2007). Recently, this tool allowed the integration of functional analysis by incorporating InterPro2GO, gene-centric read assembly, and principal coordinate analysis of taxonomy and function features. This new version is called MEGAN Community Edition (MEGAN-CE) and is open source (Huson et al. 2016).

Another tool for classifying sequences is the software Kraken, which is characterized by speed and accuracy in the classification and estimation of sequence abundances. Its speed derives from the use of exact matches between the *k-mers* of sequences and sequences deposited in databases. Kraken is also able to differentiate sequences at the genus level with precision and sensitivity compared with Megablast. Kraken has a database comprising records of all *k-mers* and the LCAs of all organisms containing those *k-mers*. This database is built based on a library of genomes specified by the user, which allows a quick search of the node of the most specific taxonomic tree related to that *k-mer*. Sequence classification is accomplished by querying the database to relate the *k-mers* to the reference sequence, and the resulting LCA data are used to group the sequences by similarity. Sequences that do not have identifiable *k-mers* in the database are left unclassified by Kraken (Wood and Salzberg 2014).

Summarized descriptions of these pipelines dedicated to taxonomic and functional annotation are compiled in Table 2.

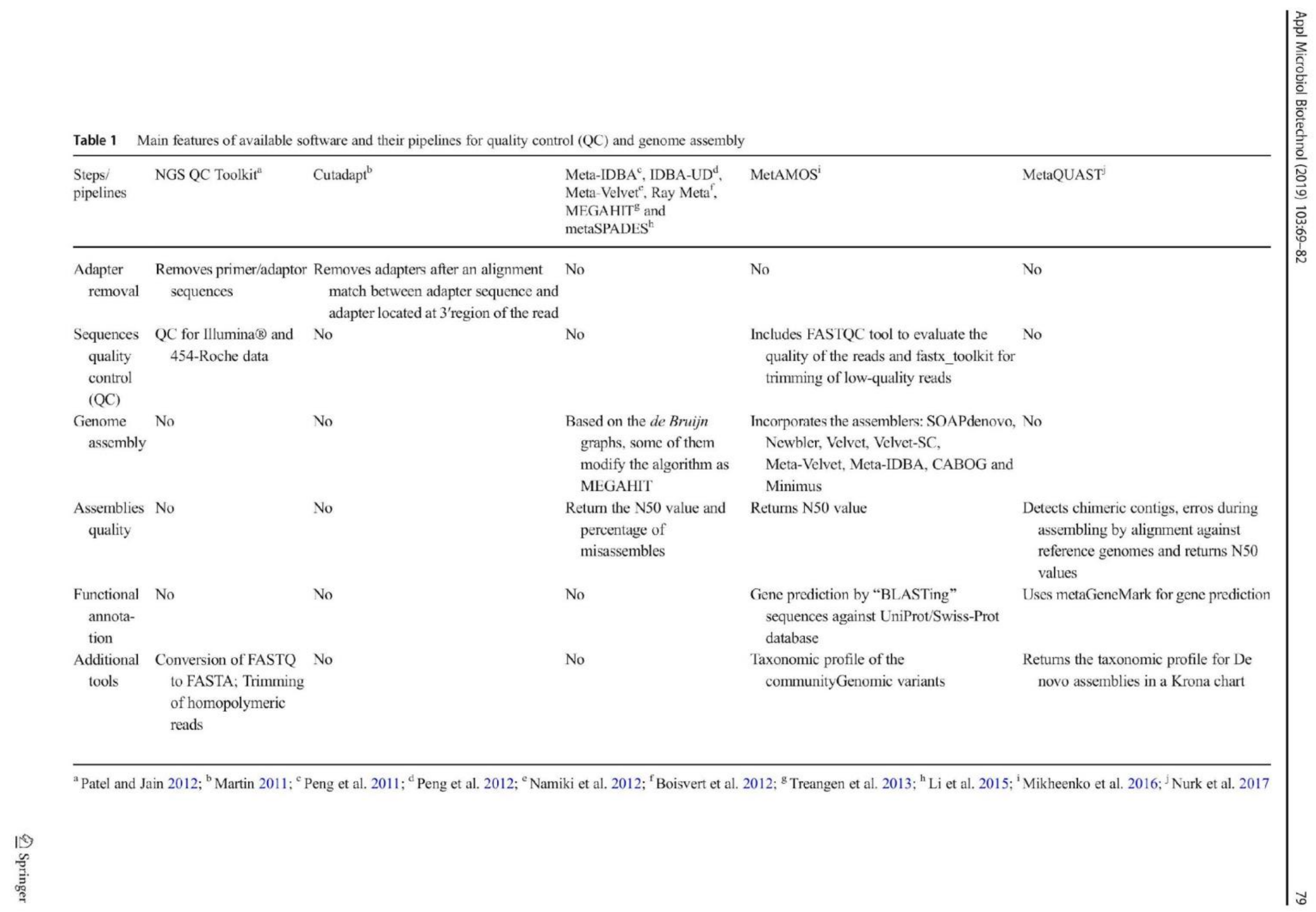


Table 1 Main features of available software and their pipelines for quality control (QC) and genome assembly

Steps/ pipelines	NGS QC Toolkit[a]	Cutadapt[b]	Meta-IDBA[c], IDBA-UD[d], Meta-Velvet[e], Ray Meta[f], MEGAHIT[g] and metaSPADES[h]	MetAMOS[i]	MetaQUAST[j]
Adapter removal	Removes primer/adaptor sequences	Removes adapters after an alignment match between adapter sequence and adapter located at 3′region of the read	No	No	No
Sequences quality control (QC)	QC for Illumina® and 454-Roche data	No	No	Includes FASTQC tool to evaluate the quality of the reads and fastx_toolkit for trimming of low-quality reads	No
Genome assembly	No	No	Based on the *de Bruijn* graphs, some of them modify the algorithm as MEGAHIT	Incorporates the assemblers: SOAPdenovo, Newbler, Velvet, Velvet-SC, Meta-Velvet, Meta-IDBA, CABOG and Minimus	No
Assemblies quality	No	No	Return the N50 value and percentage of misassembles	Returns N50 value	Detects chimeric contigs, erros during assembling by alignment against reference genomes and returns N50 values
Functional annota-tion	No	No	No	Gene prediction by "BLASTing" sequences against UniProt/Swiss-Prot database	Uses metaGeneMark for gene prediction
Additional tools	Conversion of FASTQ to FASTA; Trimming of homopolymeric reads	No	No	Taxonomic profile of the communityGenomic variants	Returns the taxonomic profile for De novo assemblies in a Krona chart

[a] Patel and Jain 2012; [b] Martin 2011; [c] Peng et al. 2011; [d] Peng et al. 2012; [e] Namiki et al. 2012; [f] Boisvert et al. 2012; [g] Treangen et al. 2013; [h] Li et al. 2015; [i] Mikheenko et al. 2016; [j] Nurk et al. 2017

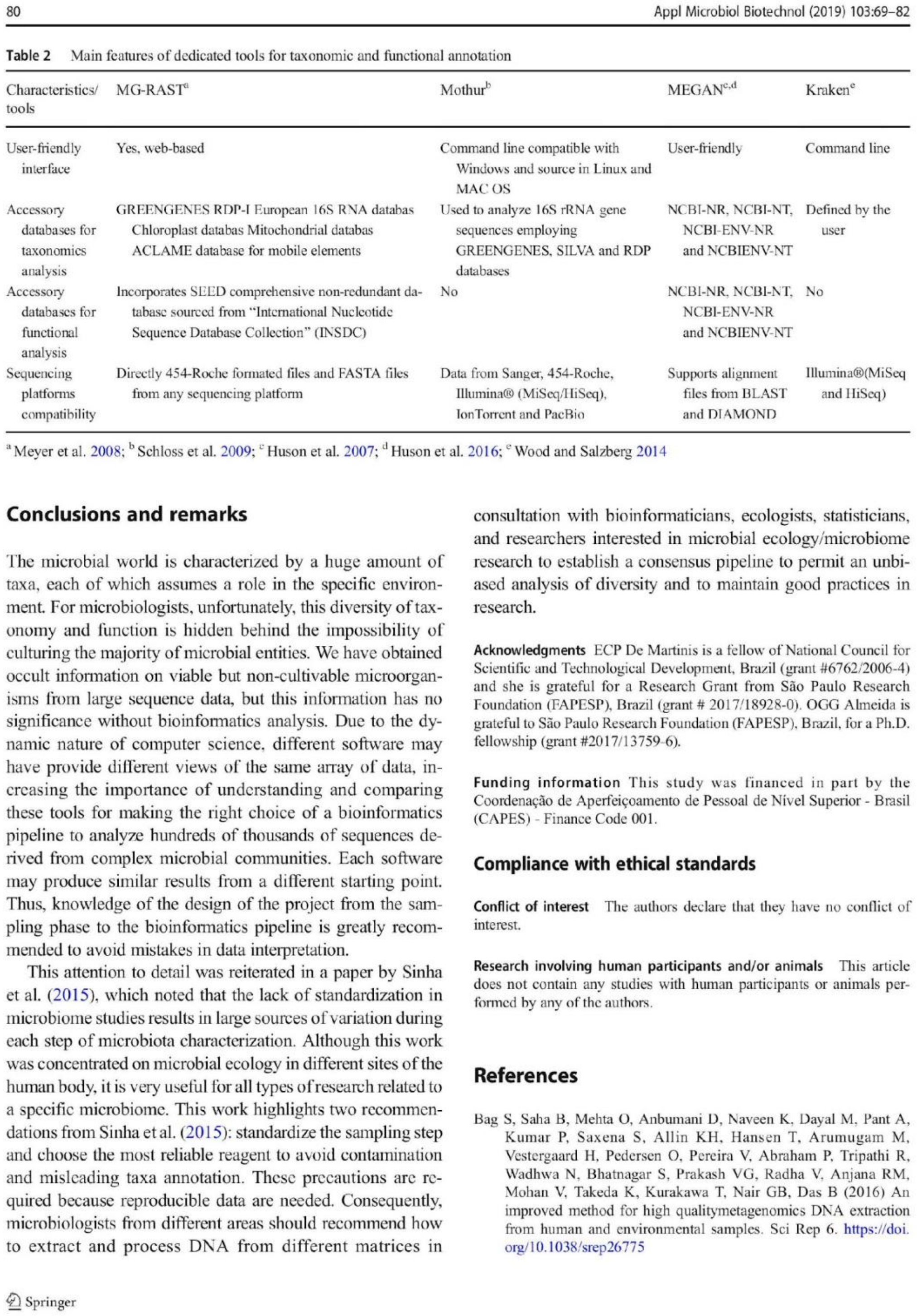


Table 2 Main features of dedicated tools for taxonomic and functional annotation

Characteristics/ tools	MG-RAST[a]	Mothur[b]	MEGAN[c,d]	Kraken[e]
User-friendly interface	Yes, web-based	Command line compatible with Windows and source in Linux and MAC OS	User-friendly	Command line
Accessory databases for taxonomics analysis	GREENGENES RDP-I European 16S RNA databas Chloroplast databas Mitochondrial databas ACLAME database for mobile elements	Used to analyze 16S rRNA gene sequences employing GREENGENES, SILVA and RDP databases	NCBI-NR, NCBI-NT, NCBI-ENV-NR and NCBIENV-NT	Defined by the user
Accessory databases for functional analysis	Incorporates SEED comprehensive non-redundant database sourced from "International Nucleotide Sequence Database Collection" (INSDC)	No	NCBI-NR, NCBI-NT, NCBI-ENV-NR and NCBIENV-NT	No
Sequencing platforms compatibility	Directly 454-Roche formated files and FASTA files from any sequencing platform	Data from Sanger, 454-Roche, Illumina® (MiSeq/HiSeq), IonTorrent and PacBio	Supports alignment files from BLAST and DIAMOND	Illumina®(MiSeq and HiSeq)

[a] Meyer et al. 2008; [b] Schloss et al. 2009; [c] Huson et al. 2007; [d] Huson et al. 2016; [e] Wood and Salzberg 2014

Conclusions and remarks

The microbial world is characterized by a huge amount of taxa, each of which assumes a role in the specific environment. For microbiologists, unfortunately, this diversity of taxonomy and function is hidden behind the impossibility of culturing the majority of microbial entities. We have obtained occult information on viable but non-cultivable microorganisms from large sequence data, but this information has no significance without bioinformatics analysis. Due to the dynamic nature of computer science, different software may have provide different views of the same array of data, increasing the importance of understanding and comparing these tools for making the right choice of a bioinformatics pipeline to analyze hundreds of thousands of sequences derived from complex microbial communities. Each software may produce similar results from a different starting point. Thus, knowledge of the design of the project from the sampling phase to the bioinformatics pipeline is greatly recommended to avoid mistakes in data interpretation.

This attention to detail was reiterated in a paper by Sinha et al. (2015), which noted that the lack of standardization in microbiome studies results in large sources of variation during each step of microbiota characterization. Although this work was concentrated on microbial ecology in different sites of the human body, it is very useful for all types of research related to a specific microbiome. This work highlights two recommendations from Sinha et al. (2015): standardize the sampling step and choose the most reliable reagent to avoid contamination and misleading taxa annotation. These precautions are required because reproducible data are needed. Consequently, microbiologists from different areas should recommend how to extract and process DNA from different matrices in consultation with bioinformaticians, ecologists, statisticians, and researchers interested in microbial ecology/microbiome research to establish a consensus pipeline to permit an unbiased analysis of diversity and to maintain good practices in research.

Acknowledgments ECP De Martinis is a fellow of National Council for Scientific and Technological Development, Brazil (grant #6762/2006-4) and she is grateful for a Research Grant from São Paulo Research Foundation (FAPESP), Brazil (grant # 2017/18928-0). OGG Almeida is grateful to São Paulo Research Foundation (FAPESP), Brazil, for a Ph.D. fellowship (grant #2017/13759-6).

Funding information This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

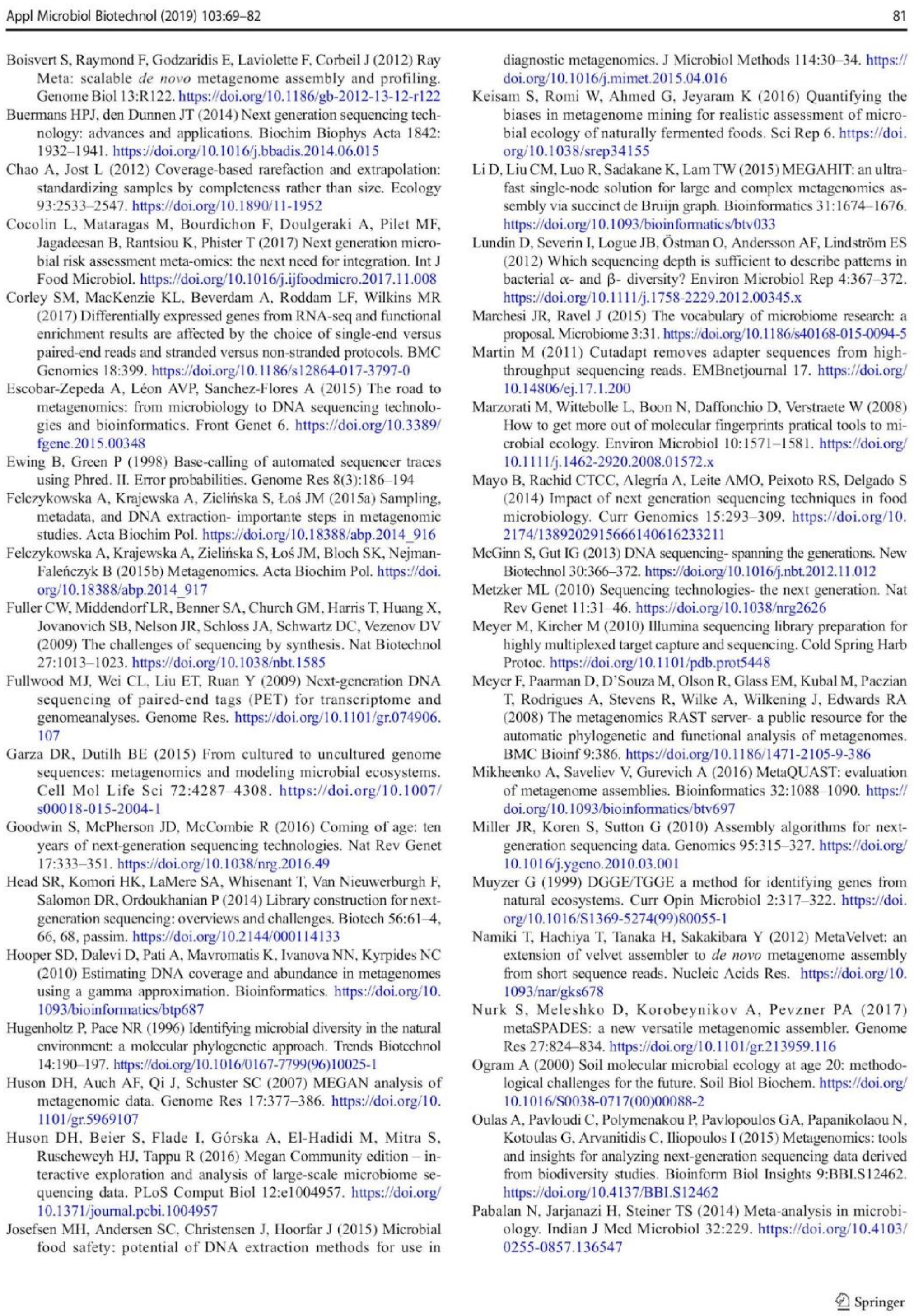
Compliance with ethical standards

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

Research involving human participants and/or animals This article does not contain any studies with human participants or animals performed by any of the authors.

References

Bag S, Saha B, Mehta O, Anbumani D, Naveen K, Dayal M, Pant A, Kumar P, Saxena S, Allin KH, Hansen T, Arumugam M, Vestergaard H, Pedersen O, Pereira V, Abraham P, Tripathi R, Wadhwa N, Bhatnagar S, Prakash VG, Radha V, Anjana RM, Mohan V, Takeda K, Kurakawa T, Nair GB, Das B (2016) An improved method for high qualitymetagenomics DNA extraction from human and environmental samples. Sci Rep 6. https://doi.org/10.1038/srep26775



Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray Meta: scalable *de novo* metagenome assembly and profiling. Genome Biol 13:R122. https://doi.org/10.1186/gb-2012-13-12-r122

Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta 1842:1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Chao A, Jost L (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93:2533–2547. https://doi.org/10.1890/11-1952

Cocolin L, Mataragas M, Bourdichon F, Doulgeraki A, Pilet MF, Jagadeesan B, Rantsiou K, Phister T (2017) Next generation microbial risk assessment meta-omics: the next need for integration. Int J Food Microbiol. https://doi.org/10.1016/j.ijfoodmicro.2017.11.008

Corley SM, MacKenzie KL, Beverdam A, Roddam LF, Wilkins MR (2017) Differentially expressed genes from RNA-seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. BMC Genomics 18:399. https://doi.org/10.1186/s12864-017-3797-0

Escobar-Zepeda A, Léon AVP, Sanchez-Flores A (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet 6. https://doi.org/10.3389/fgene.2015.00348

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res 8(3):186–194

Felczykowska A, Krajewska A, Zielińska S, Łoś JM (2015a) Sampling, metadata, and DNA extraction- importante steps in metagenomic studies. Acta Biochim Pol. https://doi.org/10.18388/abp.2014_916

Felczykowska A, Krajewska A, Zielińska S, Łoś JM, Bloch SK, Nejman-Faleńczyk B (2015b) Metagenomics. Acta Biochim Pol. https://doi.org/10.18388/abp.2014_917

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27:1013–1023. https://doi.org/10.1038/nbt.1585

Fullwood MJ, Wei CL, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genomeanalyses. Genome Res. https://doi.org/10.1101/gr.074906.107

Garza DR, Dutilh BE (2015) From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell Mol Life Sci 72:4287–4308. https://doi.org/10.1007/s00018-015-2004-1

Goodwin S, McPherson JD, McCombie R (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. Biotech 56:61–4, 66, 68, passim. https://doi.org/10.2144/000114133

Hooper SD, Dalevi D, Pati A, Mavromatis K, Ivanova NN, Kyrpides NC (2010) Estimating DNA coverage and abundance in metagenomes using a gamma approximation. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp687

Hugenholtz P, Pace NR (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol 14:190–197. https://doi.org/10.1016/0167-7799(96)10025-1

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17:377–386. https://doi.org/10.1101/gr.5969107

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R (2016) Megan Community edition – interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol 12:e1004957. https://doi.org/10.1371/journal.pcbi.1004957

Josefsen MH, Andersen SC, Christensen J, Hoorfar J (2015) Microbial food safety: potential of DNA extraction methods for use in diagnostic metagenomics. J Microbiol Methods 114:30–34. https://doi.org/10.1016/j.mimet.2015.04.016

Keisam S, Romi W, Ahmed G, Jeyaram K (2016) Quantifying the biases in metagenome mining for realistic assessment of microbial ecology of naturally fermented foods. Sci Rep 6. https://doi.org/10.1038/srep34155

Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Lundin D, Severin I, Logue JB, Östman O, Andersson AF, Lindström ES (2012) Which sequencing depth is sufficient to describe patterns in bacterial α- and β- diversity? Environ Microbiol Rep 4:367–372. https://doi.org/10.1111/j.1758-2229.2012.00345.x

Marchesi JR, Ravel J (2015) The vocabulary of microbiome research: a proposal. Microbiome 3:31. https://doi.org/10.1186/s40168-015-0094-5

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17. https://doi.org/10.14806/ej.17.1.200

Marzorati M, Wittebolle L, Boon N, Daffonchio D, Verstraete W (2008) How to get more out of molecular fingerprints pratical tools to microbial ecology. Environ Microbiol 10:1571–1581. https://doi.org/10.1111/j.1462-2920.2008.01572.x

Mayo B, Rachid CTCC, Alegría A, Leite AMO, Peixoto RS, Delgado S (2014) Impact of next generation sequencing techniques in food microbiology. Curr Genomics 15:293–309. https://doi.org/10.2174/1389202915666140616233211

McGinn S, Gut IG (2013) DNA sequencing- spanning the generations. New Biotechnol 30:366–372. https://doi.org/10.1016/j.nbt.2012.11.012

Metzker ML (2010) Sequencing technologies- the next generation. Nat Rev Genet 11:31–46. https://doi.org/10.1038/nrg2626

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. https://doi.org/10.1101/pdb.prot5448

Meyer F, Paarman D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodrigues A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server- a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinf 9:386. https://doi.org/10.1186/1471-2105-9-386

Mikheenko A, Saveliev V, Gurevich A (2016) MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32:1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. Curr Opin Microbiol 2:317–322. https://doi.org/10.1016/S1369-5274(99)80055-1

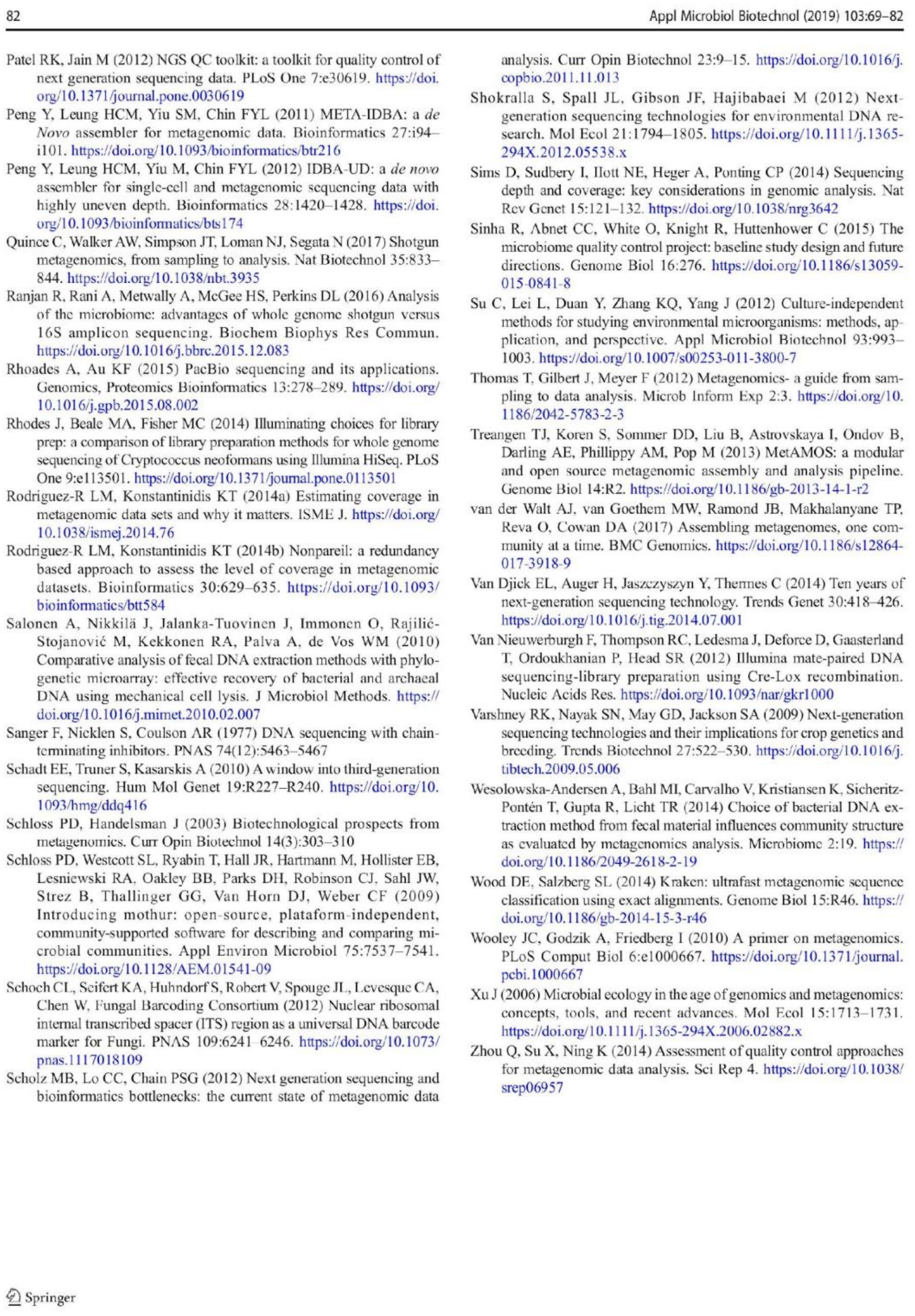
Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of velvet assembler to *de novo* metagenome assembly from short sequence reads. Nucleic Acids Res. https://doi.org/10.1093/nar/gks678

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPADES: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116

Ogram A (2000) Soil molecular microbial ecology at age 20: methodological challenges for the future. Soil Biol Biochem. https://doi.org/10.1016/S0038-0717(00)00088-2

Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights 9:BBI.S12462. https://doi.org/10.4137/BBI.S12462

Pabalan N, Jarjanazi H, Steiner TS (2014) Meta-analysis in microbiology. Indian J Med Microbiol 32:229. https://doi.org/10.4103/0255-0857.136547



Patel RK, Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619. https://doi.org/10.1371/journal.pone.0030619

Peng Y, Leung HCM, Yiu SM, Chin FYL (2011) META-IDBA: a *de Novo* assembler for metagenomic data. Bioinformatics 27:i94–i101. https://doi.org/10.1093/bioinformatics/btr216

Peng Y, Leung HCM, Yiu M, Chin FYL (2012) IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844. https://doi.org/10.1038/nbt.3935

Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL (2016) Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. https://doi.org/10.1016/j.bbrc.2015.12.083

Rhoades A, Au KF (2015) PacBio sequencing and its applications. Genomics, Proteomics Bioinformatics 13:278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rhodes J, Beale MA, Fisher MC (2014) Illuminating choices for library prep: a comparison of library preparation methods for whole genome sequencing of Cryptococcus neoformans using Illumina HiSeq. PLoS One 9:e113501. https://doi.org/10.1371/journal.pone.0113501

Rodriguez-R LM, Konstantinidis KT (2014a) Estimating coverage in metagenomic data sets and why it matters. ISME J. https://doi.org/10.1038/ismej.2014.76

Rodriguez-R LM, Konstantinidis KT (2014b) Nonpareil: a redundancy based approach to assess the level of coverage in metagenomic datasets. Bioinformatics 30:629–635. https://doi.org/10.1093/bioinformatics/btt584

Salonen A, Nikkilä J, Jalanka-Tuovinen J, Immonen O, Rajilić-Stojanović M, Kekkonen RA, Palva A, de Vos WM (2010) Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods. https://doi.org/10.1016/j.mimet.2010.02.007

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. PNAS 74(12):5463–5467

Schadt EE, Truner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240. https://doi.org/10.1093/hmg/ddq416

Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14(3):303–310

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Strez B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, plataform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS 109:6241–6246. https://doi.org/10.1073/pnas.1117018109

Scholz MB, Lo CC, Chain PSG (2012) Next generation sequencing and bioinformatics bottlenecks: the current state of metagenomic data analysis. Curr Opin Biotechnol 23:9–15. https://doi.org/10.1016/j.copbio.2011.11.013

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Mol Ecol 21:1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analysis. Nat Rev Genet 15:121–132. https://doi.org/10.1038/nrg3642

Sinha R, Abnet CC, White O, Knight R, Huttenhower C (2015) The microbiome quality control project: baseline study design and future directions. Genome Biol 16:276. https://doi.org/10.1186/s13059-015-0841-8

Su C, Lei L, Duan Y, Zhang KQ, Yang J (2012) Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. Appl Microbiol Biotechnol 93:993–1003. https://doi.org/10.1007/s00253-011-3800-7

Thomas T, Gilbert J, Meyer F (2012) Metagenomics- a guide from sampling to data analysis. Microb Inform Exp 2:3. https://doi.org/10.1186/2042-5783-2-3

Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M (2013) MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol 14:R2. https://doi.org/10.1186/gb-2013-14-1-r2

van der Walt AJ, van Goethem MW, Ramond JB, Makhalanyane TP, Reva O, Cowan DA (2017) Assembling metagenomes, one community at a time. BMC Genomics. https://doi.org/10.1186/s12864-017-3918-9

Van Djick EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30:418–426. https://doi.org/10.1016/j.tig.2014.07.001

Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR (2012) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr1000

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530. https://doi.org/10.1016/j.tibtech.2009.05.006

Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, Licht TR (2014) Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomics analysis. Microbiome 2:19. https://doi.org/10.1186/2049-2618-2-19

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6:e1000667. https://doi.org/10.1371/journal.pcbi.1000667

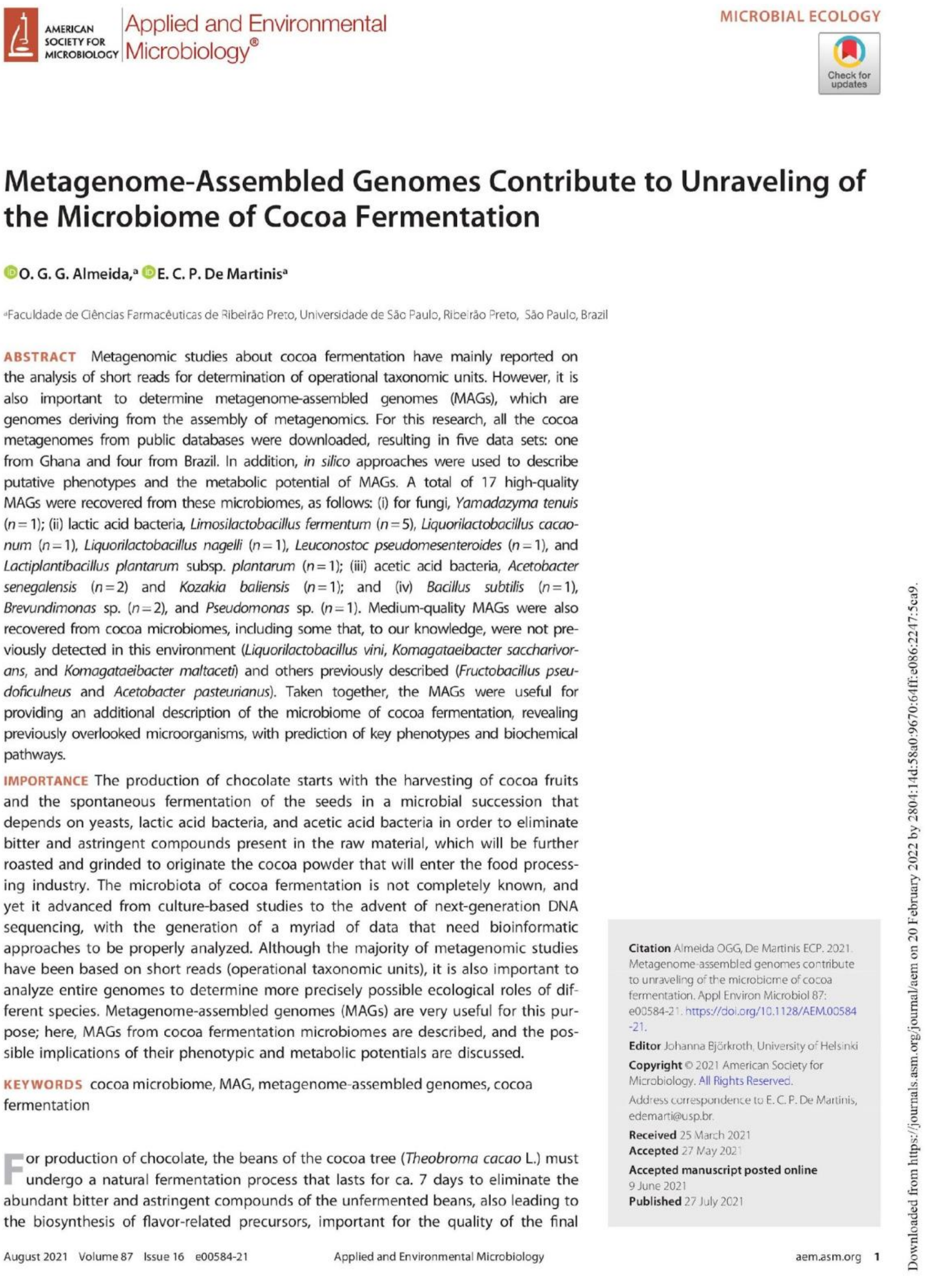
Xu J (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Mol Ecol 15:1713–1731. https://doi.org/10.1111/j.1365-294X.2006.02882.x

Zhou Q, Su X, Ning K (2014) Assessment of quality control approaches for metagenomic data analysis. Sci Rep 4. https://doi.org/10.1038/srep06957

Appendix B

Appendix B

The following article is an extension of this thesis as it presents a comparative study in which all publicly available cocoa microbiomes were surveyed for entire genomes reconstruction. The importance of that paper was to evidence novel species previously overlooked in cocoa fermentation. Besides, high-quality genomes of *A. senegalensis* were recovered and used in comparative genome analysis in the article presented in the third Chapter. The Authors Rights license is presented in the Attachment E.



AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

MICROBIAL ECOLOGY

Metagenome-Assembled Genomes Contribute to Unraveling of the Microbiome of Cocoa Fermentation

O. G. G. Almeida,[a] E. C. P. De Martinis[a]

[a]Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil

ABSTRACT Metagenomic studies about cocoa fermentation have mainly reported on the analysis of short reads for determination of operational taxonomic units. However, it is also important to determine metagenome-assembled genomes (MAGs), which are genomes deriving from the assembly of metagenomics. For this research, all the cocoa metagenomes from public databases were downloaded, resulting in five data sets: one from Ghana and four from Brazil. In addition, *in silico* approaches were used to describe putative phenotypes and the metabolic potential of MAGs. A total of 17 high-quality MAGs were recovered from these microbiomes, as follows: (i) for fungi, *Yamadazyma tenuis* (*n* = 1); (ii) lactic acid bacteria, *Limosilactobacillus fermentum* (*n* = 5), *Liquorilactobacillus cacaonum* (*n* = 1), *Liquorilactobacillus nagelli* (*n* = 1), *Leuconostoc pseudomesenteroides* (*n* = 1), and *Lactiplantibacillus plantarum* subsp. *plantarum* (*n* = 1); (iii) acetic acid bacteria, *Acetobacter senegalensis* (*n* = 2) and *Kozakia baliensis* (*n* = 1); and (iv) *Bacillus subtilis* (*n* = 1), *Brevundimonas* sp. (*n* = 2), and *Pseudomonas* sp. (*n* = 1). Medium-quality MAGs were also recovered from cocoa microbiomes, including some that, to our knowledge, were not previously detected in this environment (*Liquorilactobacillus vini*, *Komagataeibacter saccharivorans*, and *Komagataeibacter maltaceti*) and others previously described (*Fructobacillus pseudoficulneus* and *Acetobacter pasteurianus*). Taken together, the MAGs were useful for providing an additional description of the microbiome of cocoa fermentation, revealing previously overlooked microorganisms, with prediction of key phenotypes and biochemical pathways.

IMPORTANCE The production of chocolate starts with the harvesting of cocoa fruits and the spontaneous fermentation of the seeds in a microbial succession that depends on yeasts, lactic acid bacteria, and acetic acid bacteria in order to eliminate bitter and astringent compounds present in the raw material, which will be further roasted and grinded to originate the cocoa powder that will enter the food processing industry. The microbiota of cocoa fermentation is not completely known, and yet it advanced from culture-based studies to the advent of next-generation DNA sequencing, with the generation of a myriad of data that need bioinformatic approaches to be properly analyzed. Although the majority of metagenomic studies have been based on short reads (operational taxonomic units), it is also important to analyze entire genomes to determine more precisely possible ecological roles of different species. Metagenome-assembled genomes (MAGs) are very useful for this purpose; here, MAGs from cocoa fermentation microbiomes are described, and the possible implications of their phenotypic and metabolic potentials are discussed.

KEYWORDS cocoa microbiome, MAG, metagenome-assembled genomes, cocoa fermentation

For production of chocolate, the beans of the cocoa tree (*Theobroma cacao* L.) must undergo a natural fermentation process that lasts for ca. 7 days to eliminate the abundant bitter and astringent compounds of the unfermented beans, also leading to the biosynthesis of flavor-related precursors, important for the quality of the final

Citation Almeida OGG, De Martinis ECP. 2021. Metagenome-assembled genomes contribute to unraveling of the microbiome of cocoa fermentation. Appl Environ Microbiol 87: e00584-21. https://doi.org/10.1128/AEM.00584-21.

Editor Johanna Björkroth, University of Helsinki

Address correspondence to E. C. P. De Martinis, edemarti@usp.br.

Received 25 March 2021
Accepted 27 May 2021
Accepted manuscript posted online 9 June 2021
Published 27 July 2021

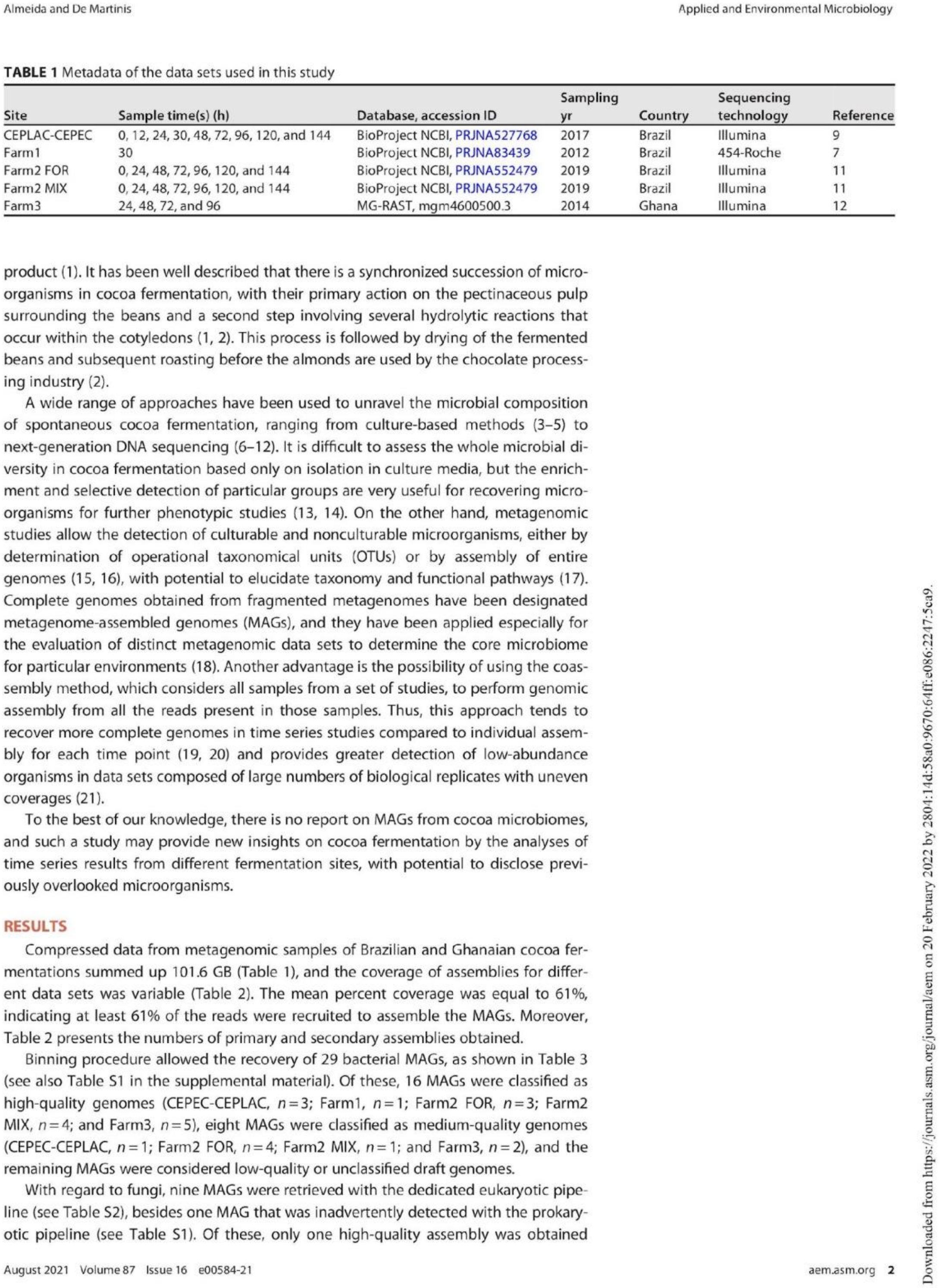


TABLE 1 Metadata of the data sets used in this study

Site	Sample time(s) (h)	Database, accession ID	Sampling yr	Country	Sequencing technology	Reference
CEPLAC-CEPEC	0, 12, 24, 30, 48, 72, 96, 120, and 144	BioProject NCBI, PRJNA527768	2017	Brazil	Illumina	9
Farm1	30	BioProject NCBI, PRJNA83439	2012	Brazil	454-Roche	7
Farm2 FOR	0, 24, 48, 72, 96, 120, and 144	BioProject NCBI, PRJNA552479	2019	Brazil	Illumina	11
Farm2 MIX	0, 24, 48, 72, 96, 120, and 144	BioProject NCBI, PRJNA552479	2019	Brazil	Illumina	11
Farm3	24, 48, 72, and 96	MG-RAST, mgm4600500.3	2014	Ghana	Illumina	12

product (1). It has been well described that there is a synchronized succession of microorganisms in cocoa fermentation, with their primary action on the pectinaceous pulp surrounding the beans and a second step involving several hydrolytic reactions that occur within the cotyledons (1, 2). This process is followed by drying of the fermented beans and subsequent roasting before the almonds are used by the chocolate processing industry (2).

A wide range of approaches have been used to unravel the microbial composition of spontaneous cocoa fermentation, ranging from culture-based methods (3–5) to next-generation DNA sequencing (6–12). It is difficult to assess the whole microbial diversity in cocoa fermentation based only on isolation in culture media, but the enrichment and selective detection of particular groups are very useful for recovering microorganisms for further phenotypic studies (13, 14). On the other hand, metagenomic studies allow the detection of culturable and nonculturable microorganisms, either by determination of operational taxonomical units (OTUs) or by assembly of entire genomes (15, 16), with potential to elucidate taxonomy and functional pathways (17). Complete genomes obtained from fragmented metagenomes have been designated metagenome-assembled genomes (MAGs), and they have been applied especially for the evaluation of distinct metagenomic data sets to determine the core microbiome for particular environments (18). Another advantage is the possibility of using the coassembly method, which considers all samples from a set of studies, to perform genomic assembly from all the reads present in those samples. Thus, this approach tends to recover more complete genomes in time series studies compared to individual assembly for each time point (19, 20) and provides greater detection of low-abundance organisms in data sets composed of large numbers of biological replicates with uneven coverages (21).

To the best of our knowledge, there is no report on MAGs from cocoa microbiomes, and such a study may provide new insights on cocoa fermentation by the analyses of time series results from different fermentation sites, with potential to disclose previously overlooked microorganisms.

RESULTS

Compressed data from metagenomic samples of Brazilian and Ghanaian cocoa fermentations summed up 101.6 GB (Table 1), and the coverage of assemblies for different data sets was variable (Table 2). The mean percent coverage was equal to 61%, indicating at least 61% of the reads were recruited to assemble the MAGs. Moreover, Table 2 presents the numbers of primary and secondary assemblies obtained.

Binning procedure allowed the recovery of 29 bacterial MAGs, as shown in Table 3 (see also Table S1 in the supplemental material). Of these, 16 MAGs were classified as high-quality genomes (CEPEC-CEPLAC, *n* = 3; Farm1, *n* = 1; Farm2 FOR, *n* = 3; Farm2 MIX, *n* = 4; and Farm3, *n* = 5), eight MAGs were classified as medium-quality genomes (CEPEC-CEPLAC, *n* = 1; Farm2 FOR, *n* = 4; Farm2 MIX, *n* = 1; and Farm3, *n* = 2), and the remaining MAGs were considered low-quality or unclassified draft genomes.

With regard to fungi, nine MAGs were retrieved with the dedicated eukaryotic pipeline (see Table S2), besides one MAG that was inadvertently detected with the prokaryotic pipeline (see Table S1). Of these, only one high-quality assembly was obtained

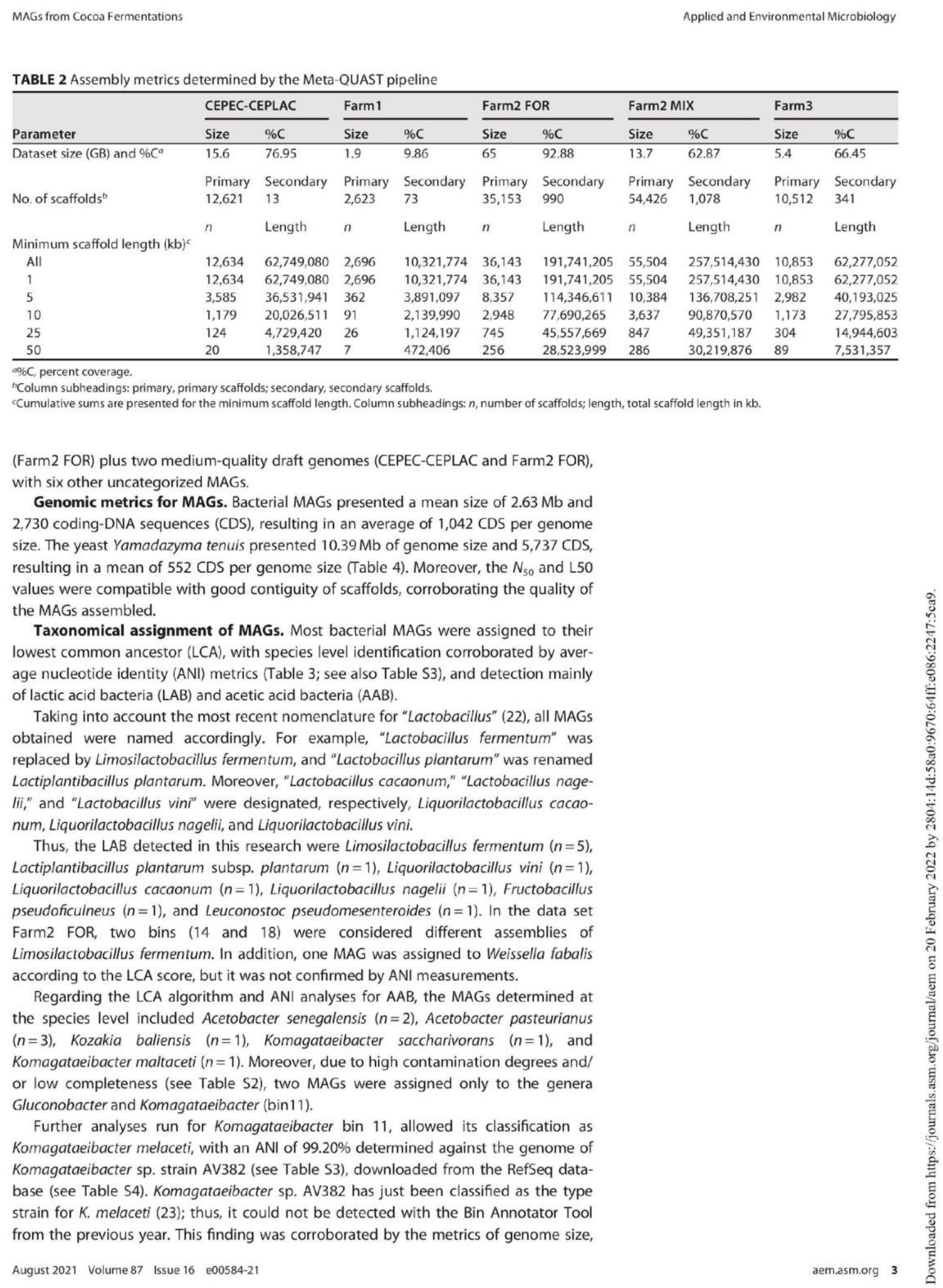


TABLE 2 Assembly metrics determined by the Meta-QUAST pipeline

	CEPEC-CEPLAC		Farm1		Farm2 FOR		Farm2 MIX		Farm3	
Parameter	Size	%C	Size	%C	Size	%C	Size	%C	Size	%C
Dataset size (GB) and %C[a]	15.6	76.95	1.9	9.86	65	92.88	13.7	62.87	5.4	66.45
	Primary	Secondary	Primary	Secondary	Primary	Secondary	Primary	Secondary	Primary	Secondary
No. of scaffolds[b]	12,621	13	2,623	73	35,153	990	54,426	1,078	10,512	341
	n	Length	*n*	Length	*n*	Length	*n*	Length	*n*	Length
Minimum scaffold length (kb)[c]										
All	12,634	62,749,080	2,696	10,321,774	36,143	191,741,205	55,504	257,514,430	10,853	62,277,052
1	12,634	62,749,080	2,696	10,321,774	36,143	191,741,205	55,504	257,514,430	10,853	62,277,052
5	3,585	36,531,941	362	3,891,097	8,357	114,346,611	10,384	136,708,251	2,982	40,193,025
10	1,179	20,026,511	91	2,139,990	2,948	77,690,265	3,637	90,870,570	1,173	27,795,853
25	124	4,729,420	26	1,124,197	745	45,557,669	847	49,351,187	304	14,944,603
50	20	1,358,747	7	472,406	256	28,523,999	286	30,219,876	89	7,531,357

[a] %C, percent coverage.

[b] Column subheadings: primary, primary scaffolds; secondary, secondary scaffolds.

[c] Cumulative sums are presented for the minimum scaffold length. Column subheadings: *n*, number of scaffolds; length, total scaffold length in kb.

(Farm2 FOR) plus two medium-quality draft genomes (CEPEC-CEPLAC and Farm2 FOR), with six other uncategorized MAGs.

Genomic metrics for MAGs. Bacterial MAGs presented a mean size of 2.63 Mb and 2,730 coding-DNA sequences (CDS), resulting in an average of 1,042 CDS per genome size. The yeast *Yamadazyma tenuis* presented 10.39 Mb of genome size and 5,737 CDS, resulting in a mean of 552 CDS per genome size (Table 4). Moreover, the N_{50} and L50 values were compatible with good contiguity of scaffolds, corroborating the quality of the MAGs assembled.

Taxonomical assignment of MAGs. Most bacterial MAGs were assigned to their lowest common ancestor (LCA), with species level identification corroborated by average nucleotide identity (ANI) metrics (Table 3; see also Table S3), and detection mainly of lactic acid bacteria (LAB) and acetic acid bacteria (AAB).

Taking into account the most recent nomenclature for "*Lactobacillus*" (22), all MAGs obtained were named accordingly. For example, "*Lactobacillus fermentum*" was replaced by *Limosilactobacillus fermentum*, and "*Lactobacillus plantarum*" was renamed *Lactiplantibacillus plantarum*. Moreover, "*Lactobacillus cacaonum*," "*Lactobacillus nagelii*," and "*Lactobacillus vini*" were designated, respectively, *Liquorilactobacillus cacaonum*, *Liquorilactobacillus nagelii*, and *Liquorilactobacillus vini*.

Thus, the LAB detected in this research were *Limosilactobacillus fermentum* ($n = 5$), *Lactiplantibacillus plantarum* subsp. *plantarum* ($n = 1$), *Liquorilactobacillus vini* ($n = 1$), *Liquorilactobacillus cacaonum* ($n = 1$), *Liquorilactobacillus nagelii* ($n = 1$), *Fructobacillus pseudoficulneus* ($n = 1$), and *Leuconostoc pseudomesenteroides* ($n = 1$). In the data set Farm2 FOR, two bins (14 and 18) were considered different assemblies of *Limosilactobacillus fermentum*. In addition, one MAG was assigned to *Weissella fabalis* according to the LCA score, but it was not confirmed by ANI measurements.

Regarding the LCA algorithm and ANI analyses for AAB, the MAGs determined at the species level included *Acetobacter senegalensis* ($n = 2$), *Acetobacter pasteurianus* ($n = 3$), *Kozakia baliensis* ($n = 1$), *Komagataeibacter saccharivorans* ($n = 1$), and *Komagataeibacter maltaceti* ($n = 1$). Moreover, due to high contamination degrees and/or low completeness (see Table S2), two MAGs were assigned only to the genera *Gluconobacter* and *Komagataeibacter* (bin11).

Further analyses run for *Komagataeibacter* bin 11, allowed its classification as *Komagataeibacter melaceti*, with an ANI of 99.20% determined against the genome of *Komagataeibacter* sp. strain AV382 (see Table S3), downloaded from the RefSeq database (see Table S4). *Komagataeibacter* sp. AV382 has just been classified as the type strain for *K. melaceti* (23); thus, it could not be detected with the Bin Annotator Tool from the previous year. This finding was corroborated by the metrics of genome size,

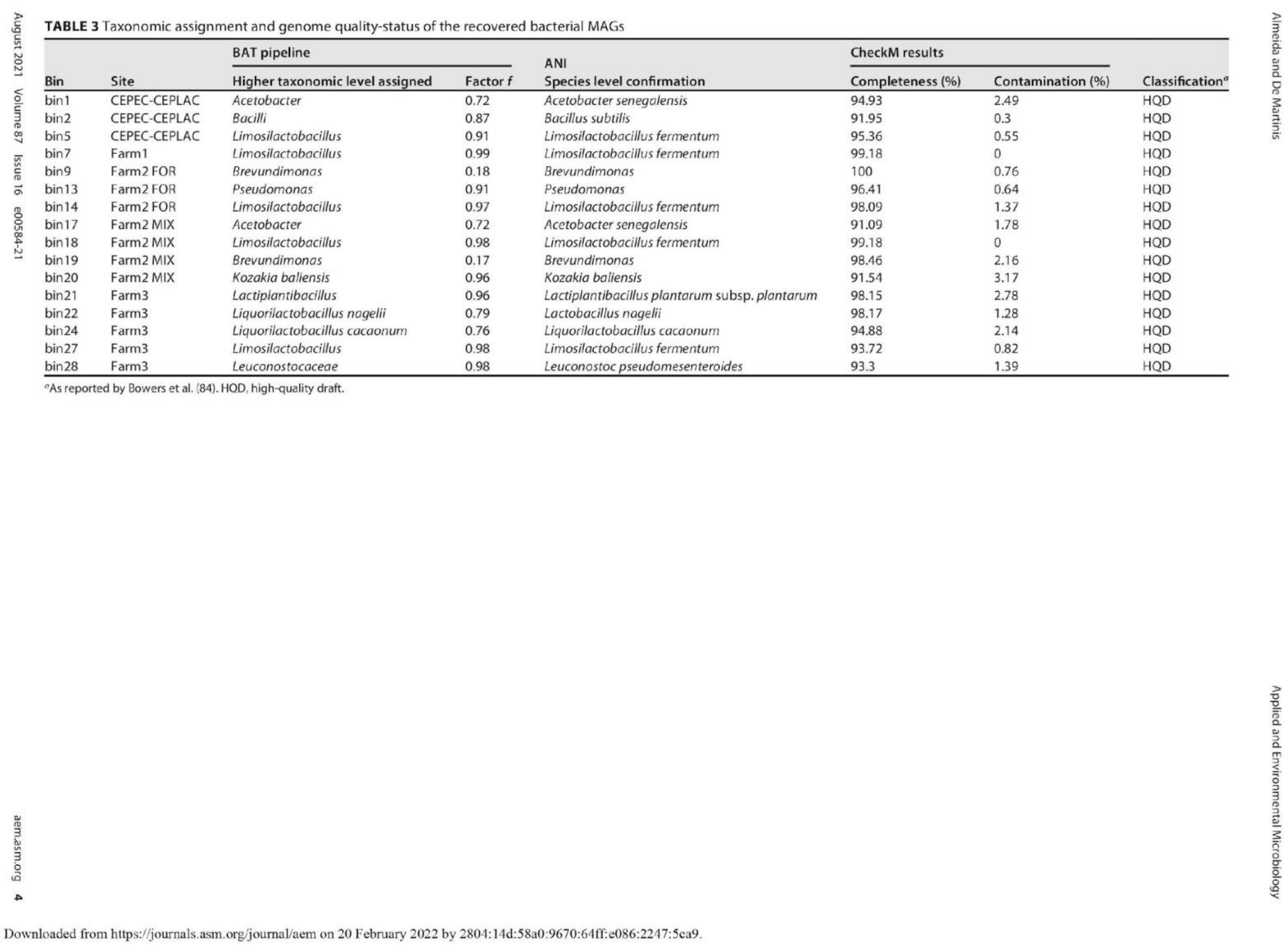


TABLE 3 Taxonomic assignment and genome quality-status of the recovered bacterial MAGs

		BAT pipeline		ANI	CheckM results		
Bin	Site	Higher taxonomic level assigned	Factor *f*	Species level confirmation	Completeness (%)	Contamination (%)	Classification[a]
bin1	CEPEC-CEPLAC	*Acetobacter*	0.72	*Acetobacter senegalensis*	94.93	2.49	HQD
bin2	CEPEC-CEPLAC	*Bacilli*	0.87	*Bacillus subtilis*	91.95	0.3	HQD
bin5	CEPEC-CEPLAC	*Limosilactobacillus*	0.91	*Limosilactobacillus fermentum*	95.36	0.55	HQD
bin7	Farm1	*Limosilactobacillus*	0.99	*Limosilactobacillus fermentum*	99.18	0	HQD
bin9	Farm2 FOR	*Brevundimonas*	0.18	*Brevundimonas*	100	0.76	HQD
bin13	Farm2 FOR	*Pseudomonas*	0.91	*Pseudomonas*	96.41	0.64	HQD
bin14	Farm2 FOR	*Limosilactobacillus*	0.97	*Limosilactobacillus fermentum*	98.09	1.37	HQD
bin17	Farm2 MIX	*Acetobacter*	0.72	*Acetobacter senegalensis*	91.09	1.78	HQD
bin18	Farm2 MIX	*Limosilactobacillus*	0.98	*Limosilactobacillus fermentum*	99.18	0	HQD
bin19	Farm2 MIX	*Brevundimonas*	0.17	*Brevundimonas*	98.46	2.16	HQD
bin20	Farm2 MIX	*Kozakia baliensis*	0.96	*Kozakia baliensis*	91.54	3.17	HQD
bin21	Farm3	*Lactiplantibacillus*	0.96	*Lactiplantibacillus plantarum* subsp. *plantarum*	98.15	2.78	HQD
bin22	Farm3	*Liquorilactobacillus nagelii*	0.79	*Lactobacillus nagelii*	98.17	1.28	HQD
bin24	Farm3	*Liquorilactobacillus cacaonum*	0.76	*Liquorilactobacillus cacaonum*	94.88	2.14	HQD
bin27	Farm3	*Limosilactobacillus*	0.98	*Limosilactobacillus fermentum*	93.72	0.82	HQD
bin28	Farm3	*Leuconostocaceae*	0.98	*Leuconostoc pseudomesenteroides*	93.3	1.39	HQD

[a] As reported by Bowers et al. (84). HQD, high-quality draft.

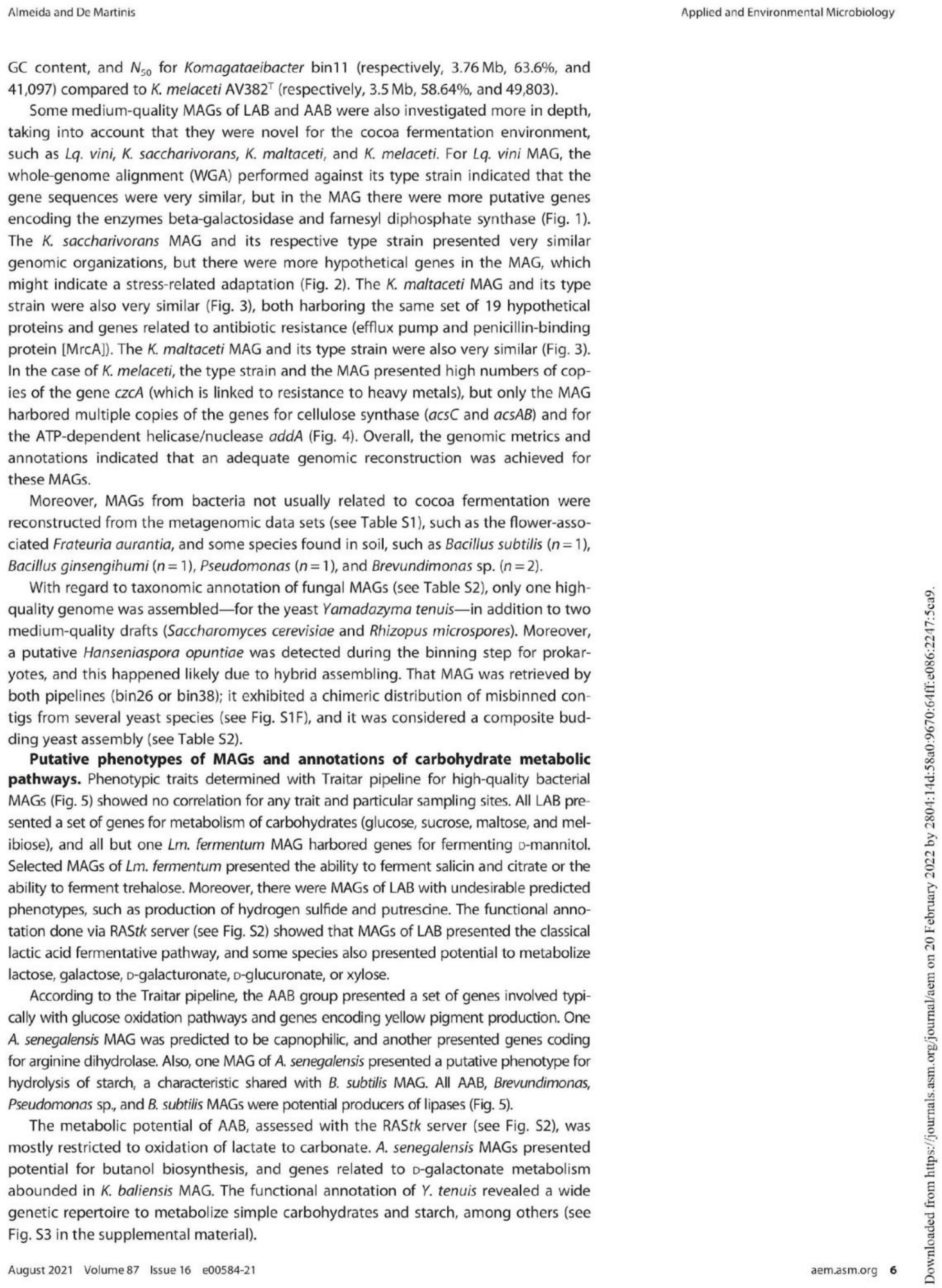
TABLE 4 Estimates of MAG features and gene prediction measurements

	RastK subsystem annotation pipeline								
MAG	Genome size (Mb)	GC percent	N_{50}	L50	No. of contigs with PEGs[a]	No. of subsystems	No. of CDS	No. of RNAs	CDS/genome size
Organism (bin)									
Acetobacter senegalensis (bin1)	3	56.1	11,432	87	356	267	3,148	40	1,049.33
Bacillus subtilis (bin2)	3.5	44.2	9,589	111	474	304	3,849	19	1,099.71
Limosilactobacillus fermentum (bin5)	1.7	52.9	33,780	18	101	206	1,756	59	1,032.94
Limosilactobacillus fermentum (bin7)	1.87	52.8	27,976	20	98	213	1,930	49	1,032.09
Brevundimonas sp. (bin9)	3.5	66	156,646	7	33	287	3,396	46	970.29
Pseudomonas sp. (bin13)	4.56	64.1	32,646	44	229	332	4,394	36	963.6
Limosilactobacillus fermentum (bin14)	2	52.1	38,130	18	107	224	2,168	46	1,084
Acetobacter senegalensis (bin17)	3.15	55.9	14,836	62	286	261	3,371	41	1,070.16
Limosilactobacillus fermentum (bin18)	1.9	52.5	41,742	13	67	215	1,930	46	1,015.79
Brevundimonas sp. (bin19)	3.6	65.9	19,394	59	267	288	3,594	46	998.33
Kozakia baliensis (bin20)	2.6	57.7	8,394	93	394	244	2,832	34	1,089.23
Lactiplantibacillus plantarum (bin21)	3	44.7	33,534	28	123	231	3,054	44	1,018
Liquorilactobacillus nagelii (bin22)	2.29	36.7	21,463	28	186	233	2,406	18	1,050.66
Liquorilactobacillus cacaonum (bin24)	1.7	33.8	49,371	8	100	181	1,763	28	1,037.06
Limosilactobacillus fermentum (bin27)	1.85	52.6	7,652	66	300	205	2,115	34	1,143.24
Leuconostoc pseudomesenteroides (bin28)	1.93	39	47,642	10	99	190	1,983	32	1,027.46
Yeast annotation									
Yamadazyma tenuis	10.39	34.06	47,674	68	345	3,384[b]	5,737	ND[c]	552.16

[a] PEGs, protein-encoding genes.

[b] No. of KEGG functional categories.

[c] ND, not determined.



GC content, and N_{50} for *Komagataeibacter* bin11 (respectively, 3.76 Mb, 63.6%, and 41,097) compared to *K. melaceti* AV382^T (respectively, 3.5 Mb, 58.64%, and 49,803).

Some medium-quality MAGs of LAB and AAB were also investigated more in depth, taking into account that they were novel for the cocoa fermentation environment, such as *Lq. vini*, *K. saccharivorans*, *K. maltaceti*, and *K. melaceti*. For *Lq. vini* MAG, the whole-genome alignment (WGA) performed against its type strain indicated that the gene sequences were very similar, but in the MAG there were more putative genes encoding the enzymes beta-galactosidase and farnesyl diphosphate synthase (Fig. 1). The *K. saccharivorans* MAG and its respective type strain presented very similar genomic organizations, but there were more hypothetical genes in the MAG, which might indicate a stress-related adaptation (Fig. 2). The *K. maltaceti* MAG and its type strain were also very similar (Fig. 3), both harboring the same set of 19 hypothetical proteins and genes related to antibiotic resistance (efflux pump and penicillin-binding protein [MrcA]). The *K. maltaceti* MAG and its type strain were also very similar (Fig. 3). In the case of *K. melaceti*, the type strain and the MAG presented high numbers of copies of the gene *czcA* (which is linked to resistance to heavy metals), but only the MAG harbored multiple copies of the genes for cellulose synthase (*acsC* and *acsAB*) and for the ATP-dependent helicase/nuclease *addA* (Fig. 4). Overall, the genomic metrics and annotations indicated that an adequate genomic reconstruction was achieved for these MAGs.

Moreover, MAGs from bacteria not usually related to cocoa fermentation were reconstructed from the metagenomic data sets (see Table S1), such as the flower-associated *Frateuria aurantia*, and some species found in soil, such as *Bacillus subtilis* ($n = 1$), *Bacillus ginsengihumi* ($n = 1$), *Pseudomonas* ($n = 1$), and *Brevundimonas* sp. ($n = 2$).

With regard to taxonomic annotation of fungal MAGs (see Table S2), only one high-quality genome was assembled—for the yeast *Yamadazyma tenuis*—in addition to two medium-quality drafts (*Saccharomyces cerevisiae* and *Rhizopus microspores*). Moreover, a putative *Hanseniaspora opuntiae* was detected during the binning step for prokaryotes, and this happened likely due to hybrid assembling. That MAG was retrieved by both pipelines (bin26 or bin38); it exhibited a chimeric distribution of misbinned contigs from several yeast species (see Fig. S1F), and it was considered a composite budding yeast assembly (see Table S2).

Putative phenotypes of MAGs and annotations of carbohydrate metabolic pathways. Phenotypic traits determined with Traitar pipeline for high-quality bacterial MAGs (Fig. 5) showed no correlation for any trait and particular sampling sites. All LAB presented a set of genes for metabolism of carbohydrates (glucose, sucrose, maltose, and melibiose), and all but one *Lm. fermentum* MAG harbored genes for fermenting D-mannitol. Selected MAGs of *Lm. fermentum* presented the ability to ferment salicin and citrate or the ability to ferment trehalose. Moreover, there were MAGs of LAB with undesirable predicted phenotypes, such as production of hydrogen sulfide and putrescine. The functional annotation done via RAS*tk* server (see Fig. S2) showed that MAGs of LAB presented the classical lactic acid fermentative pathway, and some species also presented potential to metabolize lactose, galactose, D-galacturonate, D-glucuronate, or xylose.

According to the Traitar pipeline, the AAB group presented a set of genes involved typically with glucose oxidation pathways and genes encoding yellow pigment production. One *A. senegalensis* MAG was predicted to be capnophilic, and another presented genes coding for arginine dihydrolase. Also, one MAG of *A. senegalensis* presented a putative phenotype for hydrolysis of starch, a characteristic shared with *B. subtilis* MAG. All AAB, *Brevundimonas*, *Pseudomonas* sp., and *B. subtilis* MAGs were potential producers of lipases (Fig. 5).

The metabolic potential of AAB, assessed with the RAS*tk* server (see Fig. S2), was mostly restricted to oxidation of lactate to carbonate. *A. senegalensis* MAGs presented potential for butanol biosynthesis, and genes related to D-galactonate metabolism abounded in *K. baliensis* MAG. The functional annotation of *Y. tenuis* revealed a wide genetic repertoire to metabolize simple carbohydrates and starch, among others (see Fig. S3 in the supplemental material).

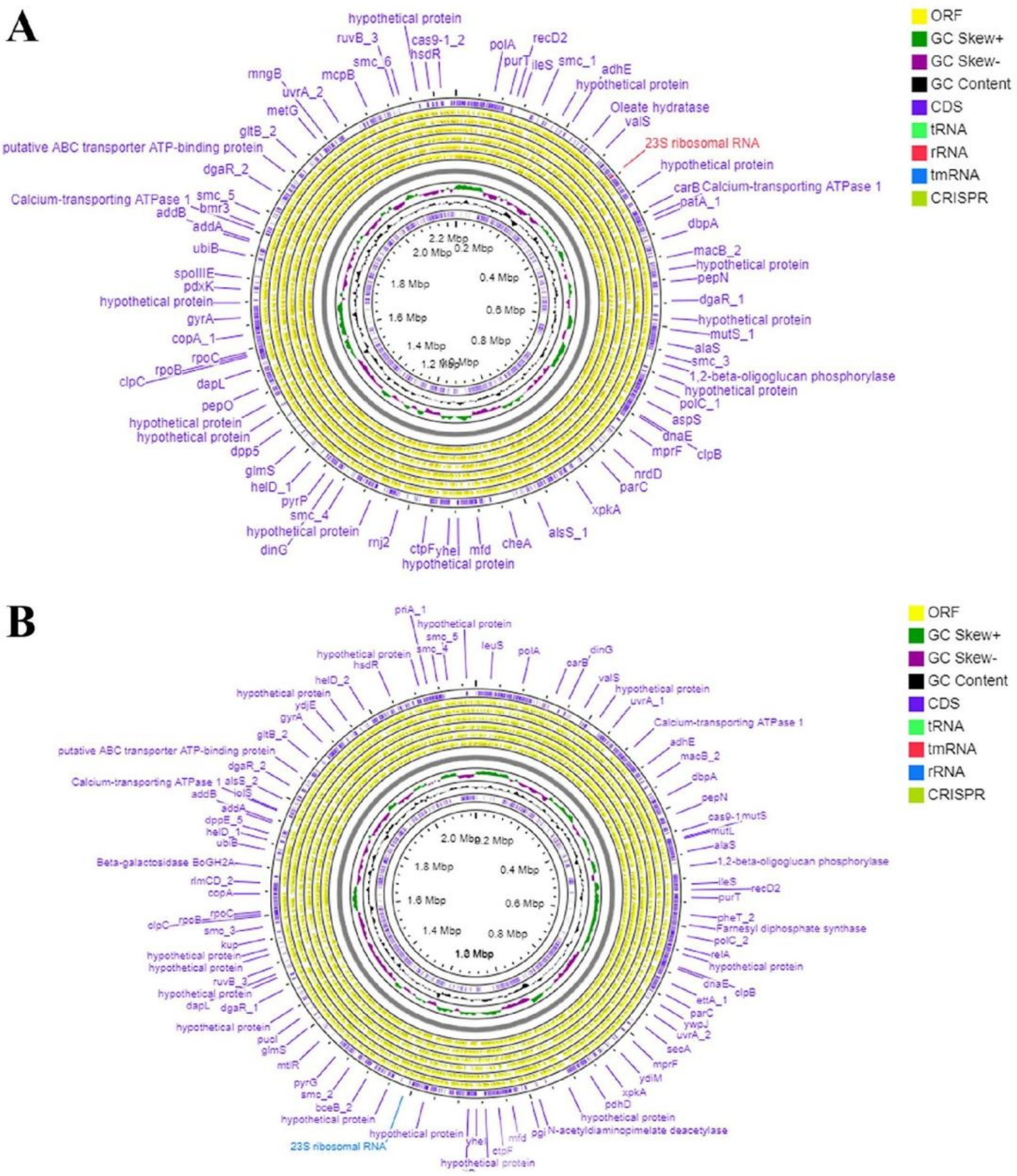


FIG 1 Genomic organizations of *Liquorilactobacillus vini*. (A) *Lq. vini* type; (B) *Lq. vini* MAG.

DISCUSSION

In this study, a total of 17 high-quality MAGs were assembled from all publicly available metagenomes from cocoa fermentations (7, 9, 11, 12), indicating coassembly was able to reconstruct draft genomes from five independent fermentations performed by distinct

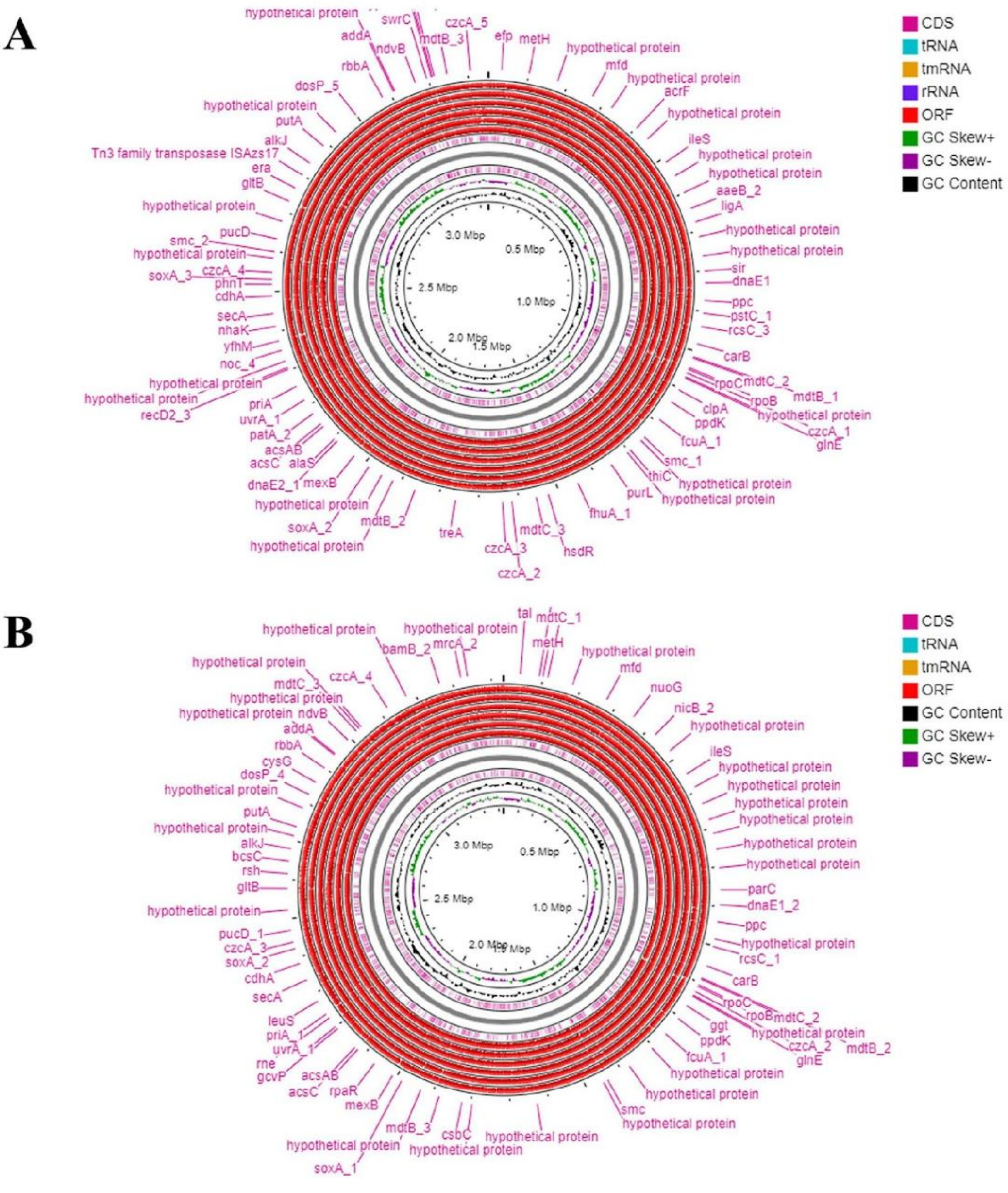


FIG 2 Genomic organizations of *Komagataeibacter saccharivorans*. (A) *K. saccharivorans* type; (B) *K. saccharivorans* MAG.

research groups from Brazil and Ghana. Using this bioinformatic strategy, the input files were constituted by reads from multiple samples, allowing for increased read depth and optimized recovery of entire genomes from metagenomes. This offered the advantage of comparing data from fermentations of diverse geographical regions, performed at different

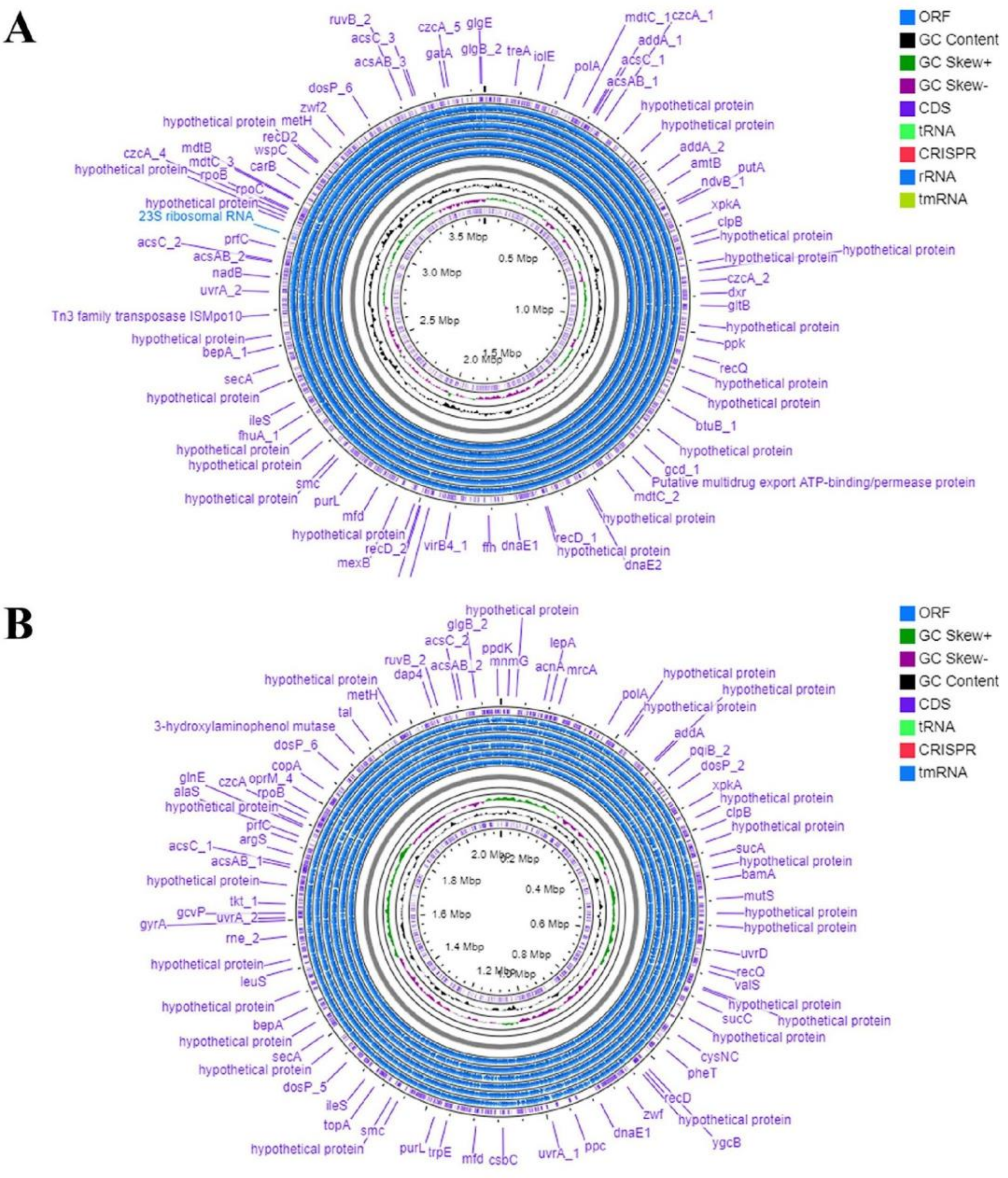


FIG 3 Genomic organizations of *Komagataeibacter maltaceti*. (A) *K. maltaceti* type; (B) *K. maltaceti* MAG.

times and with distinct farmer practices, in order to provide a broad view of the most important microbial players in action during cocoa fermentation (24).

First, the data obtained (Table 4) showed that bacterial MAGs presented higher proportions of CDS by genome size in comparison to the yeast *Y. tenuis*, a finding in accordance with

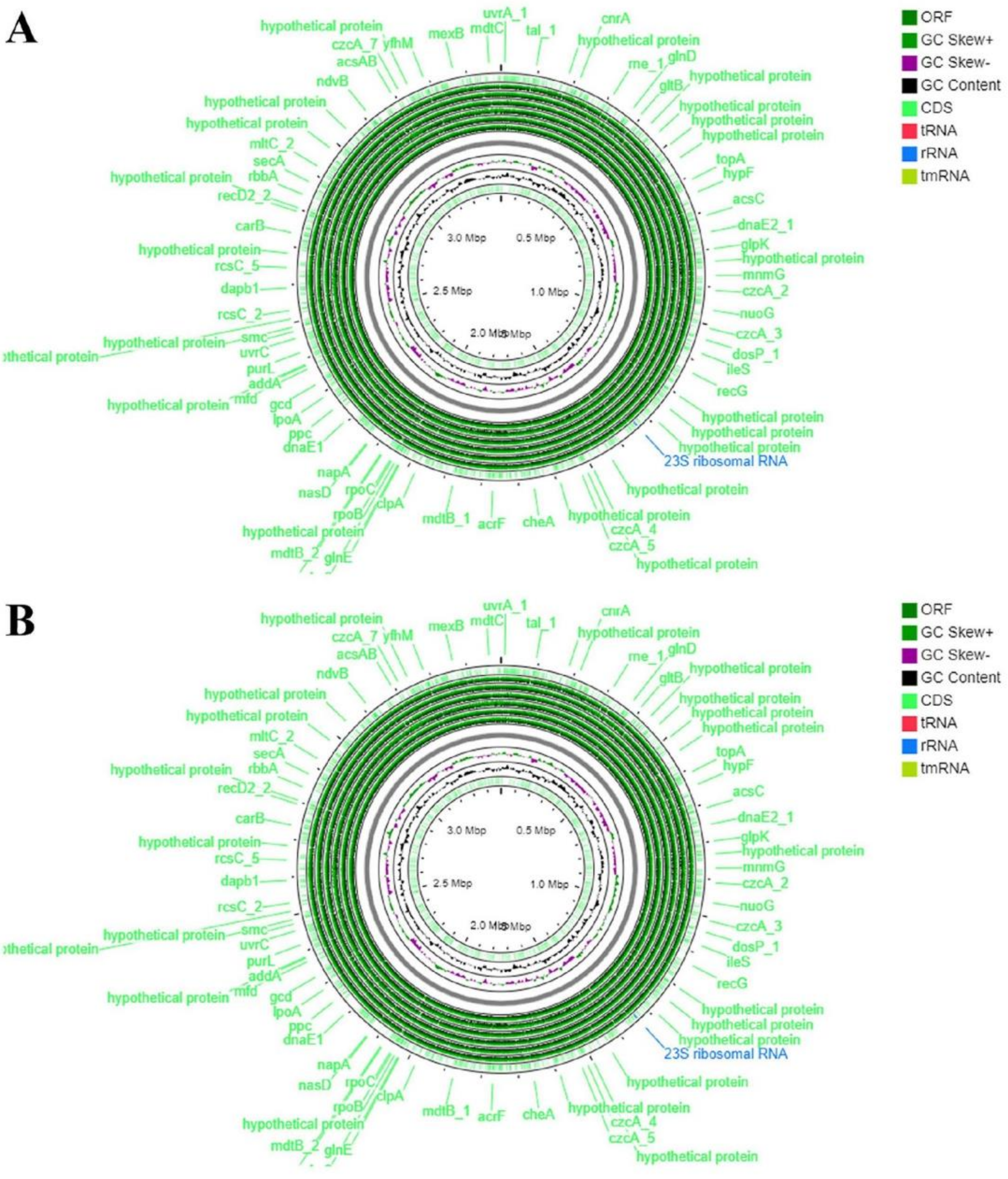


FIG 4 Genomic organizations of *Komagataeibacter maltaceti*. (A) *K. melaceti* type; (B) *K. melaceti* MAG.

literature that reports bacterial genomes are composed mostly of protein-coding sequences, in contrast to the genome of eukaryotes that harbor also nonfunctional repetitive sequences (25).

Moreover, the high-quality MAGs recovered from the CEPEC-CEPLAC data set (Brazil) were in accordance with a previous study reporting on the presence of *B.*

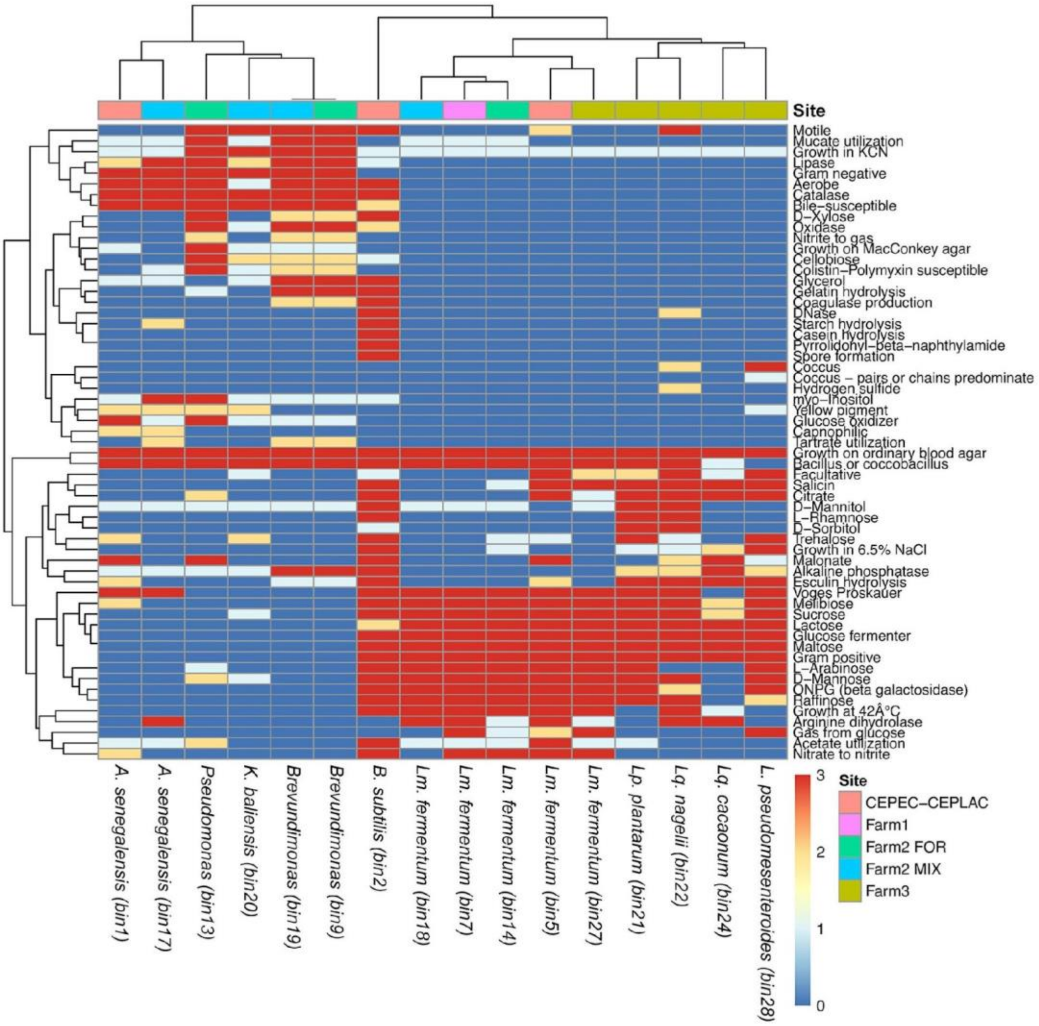


FIG 5 MAG phenotypes. The scale (0, 1, 2, and 3) refers to the number of consensual refinements to attribute a given trait to the draft. The scale is in the range 0, 1, 2, and 3, following the default criteria for trait refinement for a given genomes or MAG. A negative prediction returns zero, a unique prediction made only by the pure phyletic (phypat) classifier returns 1, a prediction made only by the phylogeny aware (PGL) classifier returns 2, and when both models are concordant the program returns the value 3.

subtilis, *Lm. fermentum*, and *Acetobacter senegalensis* (9). In addition, MAG reconstruction was important for species level taxonomy of reads previously assigned only at the genus level (*Acetobacter*), with an abundance of *A. senegalensis* (Table 3).

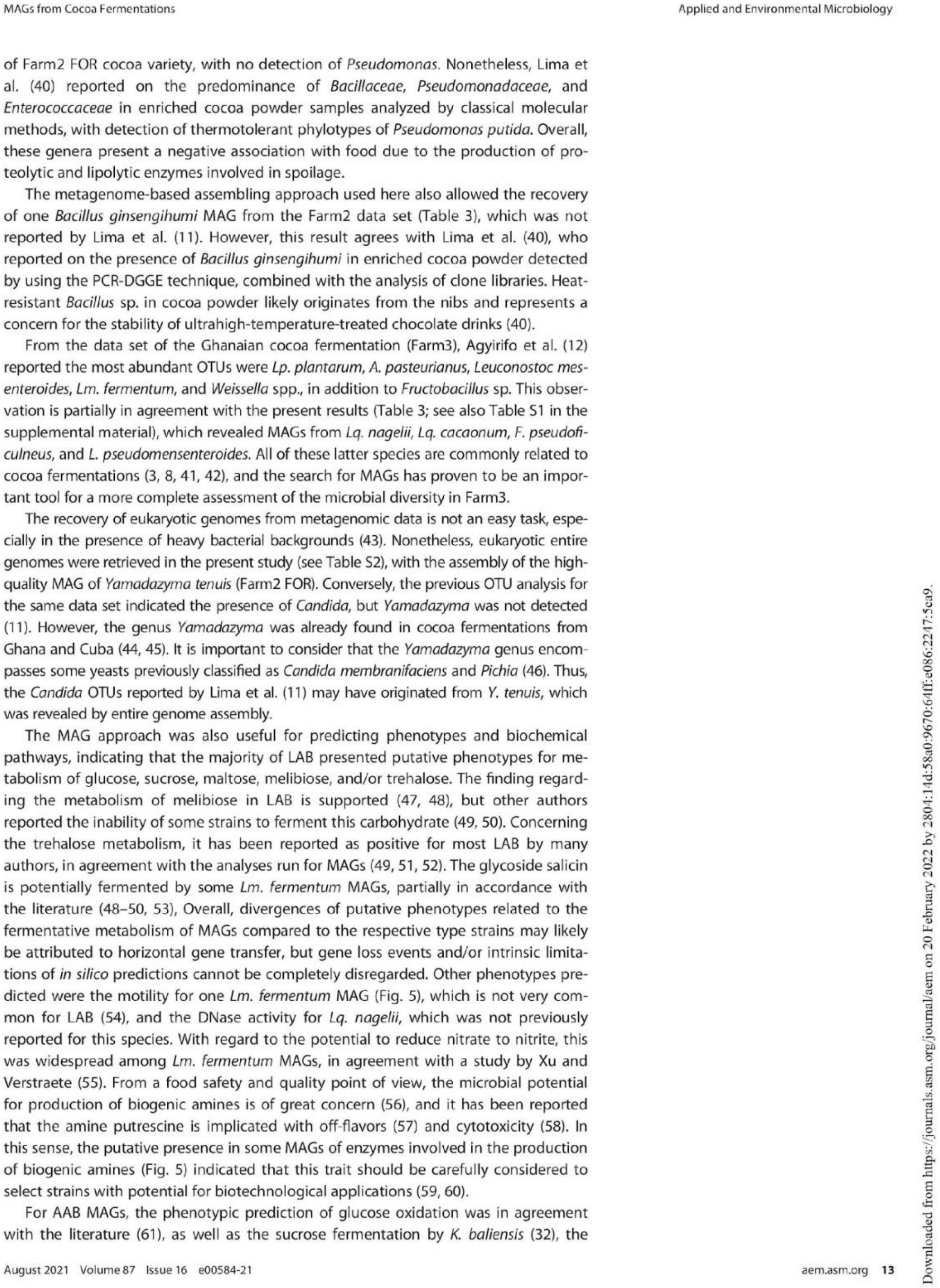
With regard to the Farm1 data set also from Brazil (7), it was previously reported the most abundant OTUs were *Lp. plantarum* (10%) and *Lm. fermentum* (2%) but, intriguingly, the corresponding *Lp. plantarum* MAG was not detected in the present study, only *Lm. fermentum* (Table 3). However, those authors of that study (7) analyzed only a fermentation time of 30 h, which presents a relatively low bacterial load (26). Thus, it is possible that *Lp. plantarum* MAG went undetected because there was not enough bacterial DNA to capture the entire genome from the metagenome data set. Another possible explanation for the divergent results is that, according to Evanovich et al. (27), the genome of *Lp. plantarum* (~3.2 Mb) would be more difficult to reconstruct because it is larger than that of *Lm. fermentum* (~1.9 Mb). Moreover, it is known that *Lp. plantarum* and *Lm. fermentum* share many gene sets at the family level, and the similarity among strains originated from the same environment can be very high due to the occurrence of horizontal gene transfers (27). In the case of high similarity between two species, the analysis of short reads (OTUs) may lead to an inaccurate taxonomical attribution, which can be disambiguated when entire genomes are considered.

The cocoa microbiome previously reported for the Brazilian Farm2 data set (11) was composed mainly by OTUs from *Acetobacter*, namely, *A. senegalensis* and *A. pasteurianus*. In the present research (Table 3), one entire genome of *A. pasteurianus* was retrieved with medium quality (65.39% completeness; 14.1% contamination). The present results also revealed a *Lq. vini* MAG in cocoa fermentation, but this bacterium had been previously associated only with grape fermentation (28) and bioethanol production (29).

The genus *Komagataeibacter* was in low abundance among the MAGs retrieved in the Brazilians data sets, but it presented a high diversity, as indicated by the detection of the species *K. melaceti*, *K. maltaceti*, and *K. saccharivorans*, which are all novel in cocoa fermentation environments. *Komagataeibacter melaceti* AV382^T was recently isolated from apple cider vinegar in Slovenia (23), and it was defined as the type strain for the species. The phylogenetic relationship between the Slovenian *Komagataeibacter melaceti* AV382^T and the cocoa-related *Komagataeibacter melaceti* MAG from Brazil can be considered unequivocal, since the ANI analysis revealed a similarity of 99.20%. Thus, it is likely the contamination degree of 20.84% observed for *Komagataeibacter melaceti* MAG can be attributed to coaggregation of genes from multiple strains. This is in accordance with Pacheco-Montealegre et al. (24), who pointed out a wide diversity of oligotypes can be found in cocoa fermentations. Another result that further corroborates this hypothesis is the higher GC content and the larger genome size of *Komagataeibacter melaceti* MAG, in comparison with the type strain (Fig. 4). With regard to *K. saccharivorans* and *K. maltaceti*, the type strains for these species were reported from beet juice (NCBI type strain genome accession no. NKTY00000000.1) and malt vinegar fermentation (30), respectively, with a more recent detection of *K. saccharivorans* in fruit fly gut (31). The detection of *Komagataeibacter* MAGs in this research indicated indeed this genus is widely distributed, although a clear role for it in cocoa fermentation is still not defined. *Kozakia baliensis* was another MAG recovered in this study, corroborating results from the OTU-based analysis of Lima et al. (11). This AAB was originally isolated from palm brown sugar in Indonesia (32) and presents a marked ability to produce exopolysaccharides (33), which are particularly interesting for texture development in food matrices (34).

Since the MAGs of *Lq. vini*, *K. saccharivorans*, *K. maltaceti*, and *K. melaceti* are novel in cocoa fermentation, it was hypothesized that WGA against their respective type strains could reveal potentially interesting results. The presence of more hypothetical genes in the MAGs in comparison to the respective type strains (Fig. 1 and 3) was suggestive of the occurrence of horizontal gene transfers (35), although the coassembly of various strains in the same bin and/or bioinformatics misprediction could not be ruled out (36). In the other hand, for the *K. melaceti* MAG the presence multiple copies of the genes *acsC*, *acsAB*, and *addA* compared to the type strain may indicate a niche-specific adaptation to protect the cells from environmental stresses (37, 38). With regard to *K. melaceti*, both the MAG and the type strain presented many copies of the *czcA* gene, which has been implicated in bacterial resistance to heavy metals (39).

Also with regard to Brazilian data sets, two MAGs from nonfermenter Gram-negative bacteria, identified as *Brevundimonas* and *Pseudomonas* (Table 3) and both from Farm2, were retrieved in the present study. This is partially in accordance with Lima et al. (11), who detected low proportions of *Brevundimonas* OTUs in the fermentation



of Farm2 FOR cocoa variety, with no detection of *Pseudomonas*. Nonetheless, Lima et al. (40) reported on the predominance of *Bacillaceae*, *Pseudomonadaceae*, and *Enterococcaceae* in enriched cocoa powder samples analyzed by classical molecular methods, with detection of thermotolerant phylotypes of *Pseudomonas putida*. Overall, these genera present a negative association with food due to the production of proteolytic and lipolytic enzymes involved in spoilage.

The metagenome-based assembling approach used here also allowed the recovery of one *Bacillus ginsengihumi* MAG from the Farm2 data set (Table 3), which was not reported by Lima et al. (11). However, this result agrees with Lima et al. (40), who reported on the presence of *Bacillus ginsengihumi* in enriched cocoa powder detected by using the PCR-DGGE technique, combined with the analysis of clone libraries. Heat-resistant *Bacillus* sp. in cocoa powder likely originates from the nibs and represents a concern for the stability of ultrahigh-temperature-treated chocolate drinks (40).

From the data set of the Ghanaian cocoa fermentation (Farm3), Agyirifo et al. (12) reported the most abundant OTUs were *Lp. plantarum*, *A. pasteurianus*, *Leuconostoc mesenteroides*, *Lm. fermentum*, and *Weissella* spp., in addition to *Fructobacillus* sp. This observation is partially in agreement with the present results (Table 3; see also Table S1 in the supplemental material), which revealed MAGs from *Lq. nagelii*, *Lq. cacaonum*, *F. pseudoficulneus*, and *L. pseudomensenteroides*. All of these latter species are commonly related to cocoa fermentations (3, 8, 41, 42), and the search for MAGs has proven to be an important tool for a more complete assessment of the microbial diversity in Farm3.

The recovery of eukaryotic genomes from metagenomic data is not an easy task, especially in the presence of heavy bacterial backgrounds (43). Nonetheless, eukaryotic entire genomes were retrieved in the present study (see Table S2), with the assembly of the high-quality MAG of *Yamadazyma tenuis* (Farm2 FOR). Conversely, the previous OTU analysis for the same data set indicated the presence of *Candida*, but *Yamadazyma* was not detected (11). However, the genus *Yamadazyma* was already found in cocoa fermentations from Ghana and Cuba (44, 45). It is important to consider that the *Yamadazyma* genus encompasses some yeasts previously classified as *Candida membranifaciens* and *Pichia* (46). Thus, the *Candida* OTUs reported by Lima et al. (11) may have originated from *Y. tenuis*, which was revealed by entire genome assembly.

The MAG approach was also useful for predicting phenotypes and biochemical pathways, indicating that the majority of LAB presented putative phenotypes for metabolism of glucose, sucrose, maltose, melibiose, and/or trehalose. The finding regarding the metabolism of melibiose in LAB is supported (47, 48), but other authors reported the inability of some strains to ferment this carbohydrate (49, 50). Concerning the trehalose metabolism, it has been reported as positive for most LAB by many authors, in agreement with the analyses run for MAGs (49, 51, 52). The glycoside salicin is potentially fermented by some *Lm. fermentum* MAGs, partially in accordance with the literature (48–50, 53), Overall, divergences of putative phenotypes related to the fermentative metabolism of MAGs compared to the respective type strains may likely be attributed to horizontal gene transfer, but gene loss events and/or intrinsic limitations of *in silico* predictions cannot be completely disregarded. Other phenotypes predicted were the motility for one *Lm. fermentum* MAG (Fig. 5), which is not very common for LAB (54), and the DNase activity for *Lq. nagelii*, which was not previously reported for this species. With regard to the potential to reduce nitrate to nitrite, this was widespread among *Lm. fermentum* MAGs, in agreement with a study by Xu and Verstraete (55). From a food safety and quality point of view, the microbial potential for production of biogenic amines is of great concern (56), and it has been reported that the amine putrescine is implicated with off-flavors (57) and cytotoxicity (58). In this sense, the putative presence in some MAGs of enzymes involved in the production of biogenic amines (Fig. 5) indicated that this trait should be carefully considered to select strains with potential for biotechnological applications (59, 60).

For AAB MAGs, the phenotypic prediction of glucose oxidation was in agreement with the literature (61), as well as the sucrose fermentation by *K. baliensis* (32), the fermentation of melibiose, and the nitrate metabolism by *A. senegalensis* MAG (62, 63). Another predicted phenotype, for *Acetobacter* species, was the formation of yellow pigment, which has been reported for other AAB and related either to ethanol oxidation or production of riboflavin (61, 64). On the other hand, the prediction of starch metabolism in *A. senegalensis* is not in accordance with the literature (63), and it may be due to an inaccurate phenotypic prediction by the pipeline or to the mixed binning of reads from starch-degrading bacilli (65) or because this species acquired genes for hydrolysis of starch via horizontal gene transfer. The data analysis also revealed the phenotype of capnophilia for *Acetobacter senegalensis* MAGs, indicating a novel possibility for culture to overcome problems due to a viable but nonculturable state (66).

Taking into account the complexity of cocoa microbiome, it is key to understand the metabolic pathways driving the microbial succession that transform high-molecular-weight compounds into simple carbohydrates, free amino acids, and oligopeptides that serve as substrates for the Maillard reaction, which releases volatile organic compounds typical of chocolate flavor (2, 67–69).

In this sense, the functional annotation performed here was important to reveal the main metabolic routes linked to the MAGs assembled. The yeast *Y. tenuis* MAG was enriched in genes related to the cycle of tricarboxylic acid and presented potential to utilize starch, sucrose, mannose, fructose, and galactose, in accordance with the results of Haase et al. (70). Next in the microbial succession, LAB can produce lactate, acetic acid, and ethanol, which are all substrates for AAB growth and generation of acetic acid through exothermic reactions (26). In addition, we demonstrated here the potential of LAB to metabolize D-galacturonate and D-glucuronate (*Lq. nagelii* and *L. pseudomesenteroides*), which can indicate the ability for the direct uptake of these products from pectin hydrolysis by yeasts and *Bacillus* spp. (71). Members of the AAB group are the last colonizers of the microbial succession in cocoa fermentation, with a role in heat killing the embryos of cocoa seeds and the formation of flavor precursors (72). In particular, there were putative genes for butanol biosynthesis in *A. senegalensis* MAGs, which can provoke flower- and candy-related notes in the fermented seeds (26, 73).

In conclusion, the MAGs presented novel findings on cocoa microbiome, phenotypes, and functional microbial pathways, with potential to aid in the selection of key members for the improvement and possible standardization of cocoa fermentations.

MATERIALS AND METHODS

Data acquisition, quality control, and coassembly. A search for shotgun metagenomic DNA sequencing data was performed in public repositories (National Center for Biotechnology Information [NCBI], MG-RAST, GOLD, and ENA). This search retrieved three bioprojects from the NCBI and another one from MG-RAST. One of the NCBI bioprojects (11) was composed of two data sets from different cultivars: FOR (Forastero) and MIX (hybrids PS1319 and CCN51).

All of the bioprojects used sequencing by synthesis technology from Illumina (San Diego, CA), except for the "Farm1" data set that was generated with 454 Pyrosequencing from Roche (Basel, Switzerland). Another publicly available data set obtained with Illumina sequencing was included from our previous report (9). The metadata of the five data sets used in this study are presented in detail in Table 1.

The raw reads were downloaded and quality processed using bbduk version 38.82, from the BBtools package (74), to remove adaptor sequences and to select reads with Phred values of +33 (bbduk parameters: hdist = 1 tpe tbo qtrim=rl trimq = 30 maq = 30). Moreover, the reads selected had a length of at least 100 bp, except for NCBI BioProject PRJNA527768, with a minimum of 50 bp. Before coassembly, the quality-processed reads were concatenated in single data sets according to the geographical proximity, as follows: CEPLAC-CEPEC, Farm1, Farm2 FOR, Farm2 MIX, and Farm3.

The metagenome coassembly was performed using MEGAHIT version 1.2.9 (75) with default settings and k-mer extensions as follows: (i) CEPLAC-CEPEC: 21, 29, 39, 59, and 79; (ii) Farm1: 21, 29, 39, 59, 79, 99, 119, and 141; (iii) Farm2 FOR: 21, 29, 39, 59, 79, 99, 119, and 141; (iv) Farm2 MIX: 21, 29, 39, 59, 79, 99, 119, and 141; and (v) Farm3: 21, 29, 39, 59, 79, 99, 119, and 141. In a first processing step, contigs smaller than 1,800 bp were discarded using a custom Perl script to keep only primary contigs longer than 1,800 bp. These were dereplicated with CD-HIT-EST tool version 4.8.1 (76) to combine contigs with minimum 99% global sequence similarity (-c 0.99) in order to eliminate redundancy and to decrease computational loads. To generate longer contigs, dereplicated primary contigs from each fermentation site were coassembled once more, using CAP3 assembler (77), with an overlap percent identity cutoff of ≥97% and an overlap length cutoff of ≥97% (parameters: -p 97 -o 97). The resulting coassemblies were

referred as secondary contigs, and for downstream analyses they were concatenated with the primary contigs previously assembled, generating a single file.

Binning procedures and quality control for bacterial MAGs. To obtain bacterial MAGs, three independent binners were used: Maxbin2 version 2.2.5 (78), Metabat2 version 2.12.1 (79), and CONCOCT version 1.0.0 (80). This approach was done to minimize biases and to maximize the certainty of results, taking into account that there is no consensus in the literature about the best binning algorithm currently available.

The sequencing depth (coverage) was calculated using default parameters of Bowtie2 version 2.3.4.3 (81), with coassembled data sets as indexes. The raw reads were mapped against the indexed assemblies to generate SAM files, which were subsequently converted to sorted and indexed BAM files. The script "jgi_summarize_bam_contig_depths" from the Metabat2 pipeline was used to calculate the coverages of the assemblies, which were summarized in mapping files.

The binning with the software Maxbin2 was performed by inputting the coassembled data sets and the calculated mapping file, using default parameters of a scaffold length of ≥2,000 bp and a 107-marker-gene set.

For binning with Metabat2, the coassembled data sets and the calculated mapping files were input, following the default guidelines of the software, with a contig length of ≥2,000 bp.

With CONCOCT, the binning was performed by slicing the coassembled contigs into smaller sequences to generate a coverage table using the previous BAM files from Bowtie2 alignments. The composition of sequences was calculated, the coassembled sequences were clustered, and the bins were parsed into different files using the built-in script "extract_fasta_bins.Py."

DAStool version 1.1.2 (82) was used to compare data generated with the three independent binning pipelines and to aggregate redundant bins, based on the composition of single-copy genes (SCGs).

Also, the metrics of completeness and degree of contamination were calculated for the bins obtained, by using the CheckM tool version 1.1.2 (83) with the parameters "lineage_wf." Finally, the genomes obtained were classified according to Bowers et al. (84), as follows: "high-quality draft" (completeness > 90% and < 5% contamination); "medium-quality draft" (completeness ≥ 50% and < 10% contamination); "low-quality draft" (completeness < 50% and < 10% contamination); and "not categorized." The genome assemblies' statistics were calculated using Meta-QUAST pipeline version 5.1. (85).

Binning procedures and quality control for fungal MAGs. To capture particularly contigs with eukaryotic signatures, the bins previously generated by the CONCOCT pipeline were analyzed with the script "filter_euk_bins.py" from the EukCC pipeline (86) that includes the EukRep pipeline (87). This pipeline was used to scan the bins and to select those containing over 20% of eukaryotic bases, with the objective of excluding putative bacterial sequences. Moreover, GeneMark-ES (88) was used to predict proteins from the bins (parameters: –ES, –fungus, –min_contig 5000), in order to estimate the completeness and contamination degrees with BUSCO (89), available in the EukCC pipeline.

Taxonomical identification of bacterial MAGs. For this, the Bin Annotator Tool (BAT) version 5.1.2 (90) was used, and all of the proteins from individual bacterial MAGs were predicted (–meta parameter) with the prodigal version 2.6.3 (91). The predicted proteins were blasted against the NCBI nonredundant database using DIAMOND version 0.9.34 (92), and then each CDS was assigned to its LCA, and the scores were calculated (90). In addition, to confirm the assignments at the species level obtained with BAT, the ANI was determined for each MAG. For this purpose, genomes were downloaded from the RefSeq database using the script "ncbi-genome-download" (see Table S5), and the ANIs were calculated with the fastANI tool (93).

If the taxonomic assignment from BAT was not confirmed with ANI metrics (see Table S3), further processing was done by downloading all type and non-type genomes for the genus from the RefSeq database, as were the cases for *Komagataeibacter* sp. (see Table S4) and *Weissella* sp. (see Table S6). For these analyses, the fastANI tool was used, with a 95% threshold for species assignment.

Additional analyses were done for medium-quality MAGs of *Lq. vini*, *K. saccharivorans*, *K. maltaceti*, and *K. melaceti*, which had never been reported for cocoa-related environments. Their genomic organizations were compared to each respective type strain by using PROKKA for annotation (94) and the software CGView (95) for genome visualization.

Taxonomical identification of fungal MAGs. A customized database was built by downloading all nonredundant RefSeq fungal proteins, and they were concatenated in a single file, which was indexed in DIAMOND version 0.9.34 for blastp alignment according to GeneMark-ES. To assign the MAGs to the most probable species, the results were filtered for hits with low E values ($<1 \times 10^{-5}$) and high identities.

Functional annotation of MAGs. Genes were predicted using prodigal version 2.6.3 (parameter: –meta) and annotated on Traitar pipeline (96) (version 1.1.2), according to the criteria of presence/absence of the protein in a given phenotype (phypat) and similar data sum with evolutionary information about gene gain/loss for protein families (model phypat+PGL). The MAGs' functional repertoire related to the metabolism of carbohydrates was annotated on the RAS*tk* pipeline (97).

Protein sequences for fungi were predicted using GeneMark-ES from EukCC pipeline. To summarize the main metabolic pathways of carbohydrates, the putative proteins were blasted against KEGG database using BlastKOALA, with "genus_eukaryotes" parameter (98).

Data availability. MAGs from this research are available at GitHub (https://github.com/Otavio20/Cocoa_MAGs).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.4 MB.

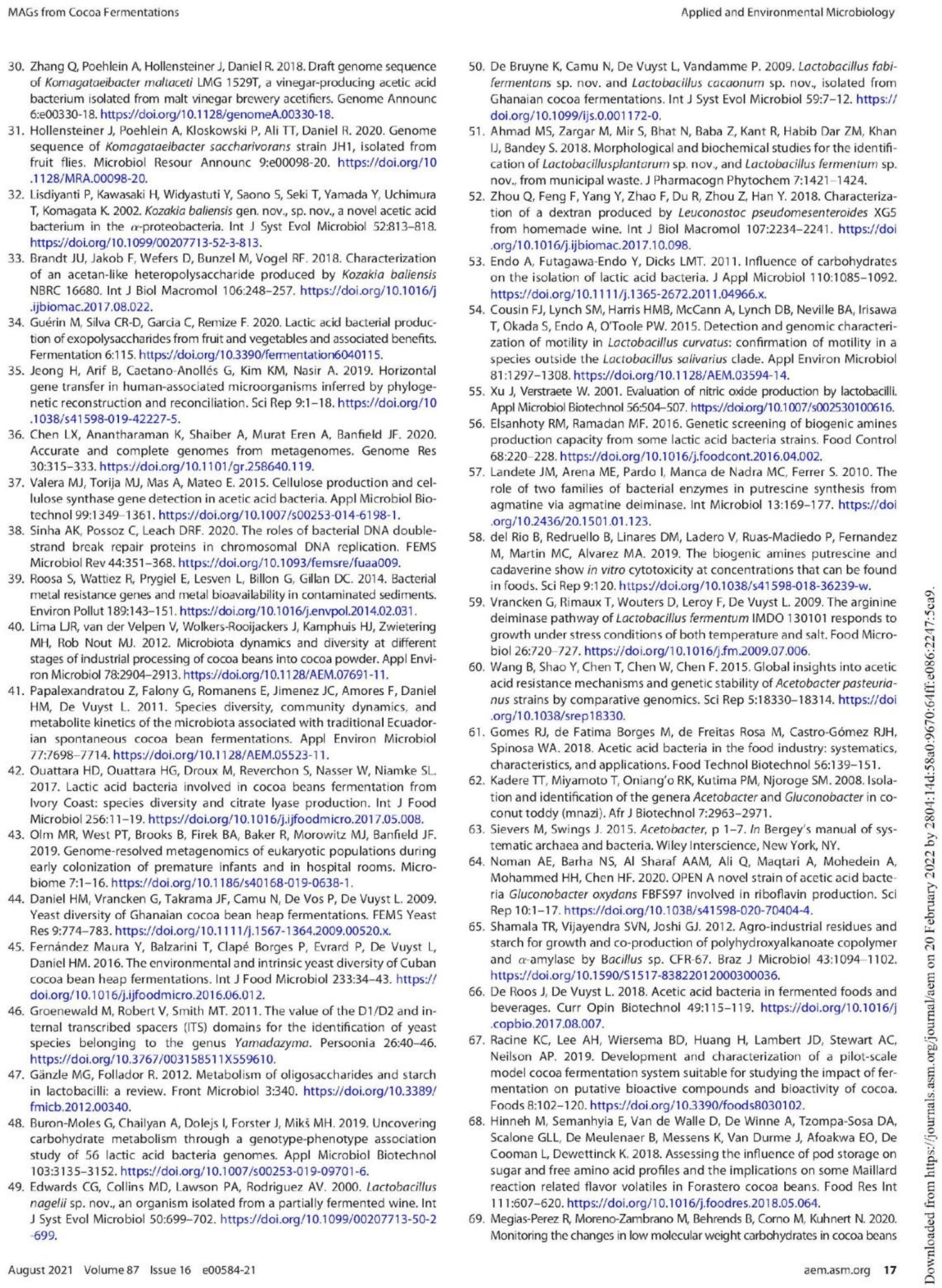
ACKNOWLEDGMENTS

We thank the São Paulo Research Foundation (FAPESP)–Brazil for a research grant (process 2018/13564-3). FAPESP is especially acknowledged by O.G.G.A. for a Ph.D. fellowship (process 2017/137596). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brazil (CAPES; finance code 001). E.C.P.D.M. is a research fellow of the National Council for Scientific and Technological Development (CNPq)–Brazil.

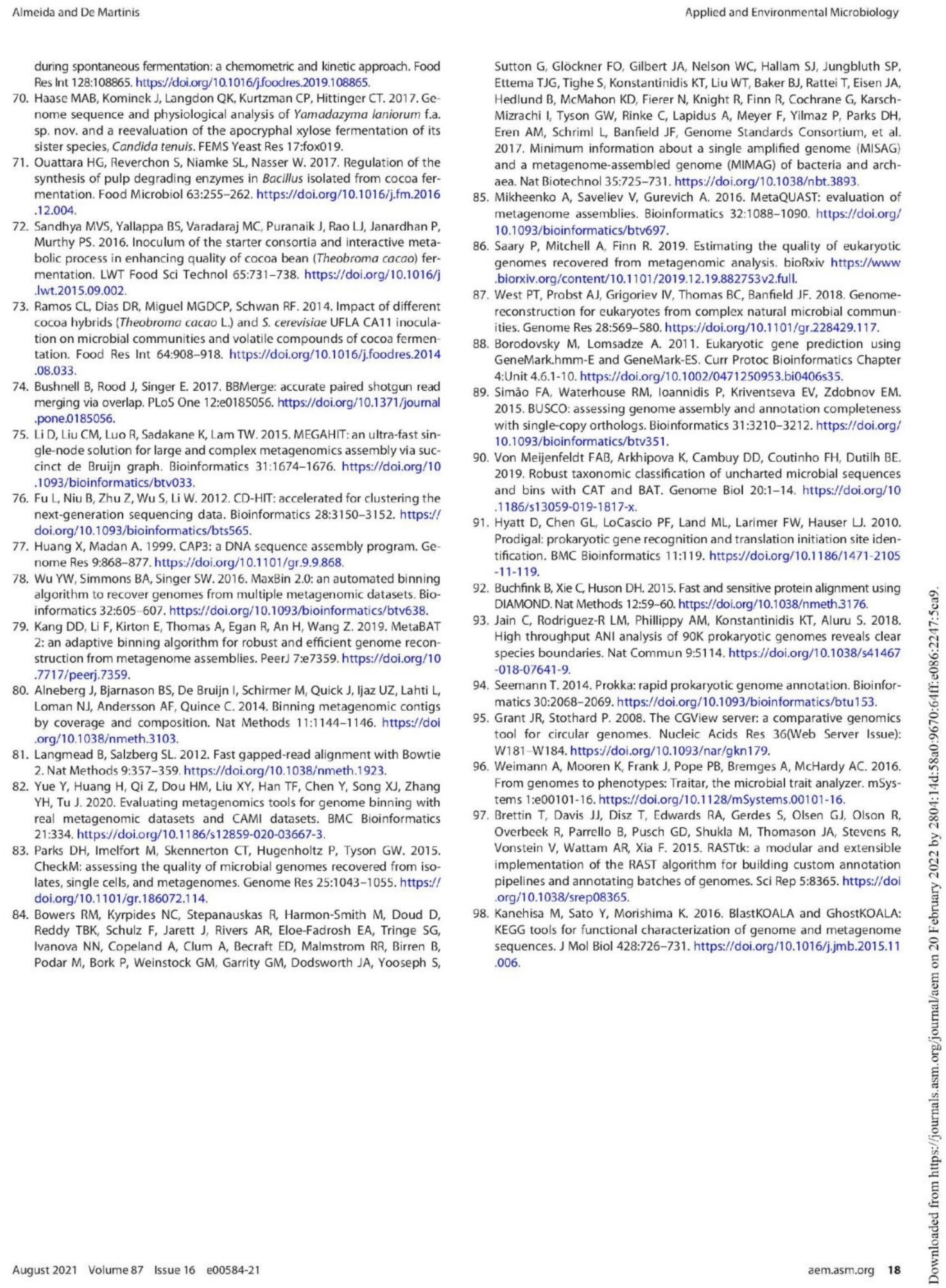
We have no conflicts of interest to declare.

REFERENCES

1. Pereira GVDM, Miguel MGDCP, Ramos CL, Schwan RF. 2012. Cocoa fermentations and screening of yeast and bacterial strains to develop a defined starter culture. Appl Environ Microbiol 78:5395–5405. https://doi.org/10.1128/AEM.01144-12.
2. De Vuyst L, Weckx S. 2016. The cocoa bean fermentation process: from ecosystem analysis to starter culture development. J Appl Microbiol 121:5–17. https://doi.org/10.1111/jam.13045.
3. Papalexandratou Z, Vrancken G, de Bruyne K, Vandamme P, de Vuyst L. 2011. Spontaneous organic cocoa bean box fermentations in Brazil are characterized by a restricted species diversity of lactic acid bacteria and acetic acid bacteria. Food Microbiol 28:1326–1338. https://doi.org/10.1016/j.fm.2011.06.003.
4. Meersman E, Steensels J, Mathawan M, Wittocx PJ, Saels V, Struyf N, Bernaert H, Vrancken G, Verstrepen KJ. 2013. Detailed analysis of the microbial population in Malaysian spontaneous cocoa pulp fermentations reveals a core and variable microbiota. PLoS One 8:e81559. https://doi.org/10.1371/journal.pone.0081559.
5. Viesser JA, de Melo Pereira GV, de Carvalho Neto DP, Vandenberghe LPDS, Azevedo V, Brenig B, Rogez H, Góes-Neto A, Soccol CR. 2020. Exploring the contribution of fructophilic lactic acid bacteria to cocoa beans fermentation: isolation, selection, and evaluation. Food Res Int 136:109478. https://doi.org/10.1016/j.foodres.2020.109478.
6. Illeghems K, de Vuyst L, Papalexandratou Z, Weckx S. 2012. Phylogenetic analysis of a spontaneous cocoa bean fermentation metagenome reveals new insights into its bacterial and fungal community diversity. PLoS One 7:e38040. https://doi.org/10.1371/journal.pone.0038040.
7. Illeghems K, Weckx S, De Vuyst L. 2015. Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. Food Microbiol 50:54–63. https://doi.org/10.1016/j.fm.2015.03.005.
8. Bortolini C, Patrone V, Puglisi E, Morelli L. 2016. Detailed analyses of the bacterial populations in processed cocoa beans of different geographic origin, subject to varied fermentation conditions. Int J Food Microbiol 236:98–106. https://doi.org/10.1016/j.ijfoodmicro.2016.07.004.
9. Almeida OGG, Pinto UM, Matos CB, Frazilio DA, Braga VF, von Zeska-Kress MR, De Martinis ECP. 2020. Does quorum sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An *in silico* perspective. Food Res Int 131:109034. https://doi.org/10.1016/j.foodres.2020.109034.
10. Serra JL, Moura FG, Pereira GDM, Soccol CR, Rogez H, Darnet S. 2019. Determination of the microbial community in Amazonian cocoa bean fermentation by Illumina-based metagenomic sequencing. LWT Food Sci Technol 106:229–239. https://doi.org/10.1016/j.lwt.2019.02.038.
11. Lima COC, Vaz ABM, De Castro GM, Lobo F, Solar R, Rodrigues C, Martins Pinto LR, Vandenberghe L, Pereira G, Miúra da Costa A, Benevides RG, Azevedo V, Trovatti Uetanabaro AP, Soccol CR, Góes-Neto A. 2021. Integrating microbial metagenomics and physicochemical parameters and a new perspective on starter culture for fine cocoa fermentation. Food Microbiol 93:103608. https://doi.org/10.1016/j.fm.2020.103608.
12. Agyirifo DS, Wamalwa M, Otwe EP, Galyuon I, Runo S, Takrama J, Ngeranwa J. 2019. Metagenomics analysis of cocoa bean fermentation microbiome identifying species diversity and putative functional capabilities. Heliyon 5:e02170. https://doi.org/10.1016/j.heliyon.2019.e02170.
13. Stefani FOP, Bell TH, Marchand C, De La Providencia IE, El Yassimi A, St-Arnaud M, Hijri M. 2015. Culture-dependent and -independent methods capture different microbial community fractions in hydrocarbon-contaminated soils. PLoS One 10:e0128272–16. https://doi.org/10.1371/journal.pone.0128272.
14. Almeida OGG, De Martinis ECP. 2019. Bioinformatics tools to assess metagenomic data for applied microbiology. Appl Microbiol Biotechnol 103:69–82. https://doi.org/10.1007/s00253-018-9464-9.
15. Vitorino LC, Bessa LA. 2018. Microbial diversity: the gap between the estimated and the known. Diversity 10:46. https://doi.org/10.3390/d10020046.
16. Kumar Awasthi M, Ravindran B, Sarsaiya S, Chen H, Wainaina S, Singh E, Liu T, Kumar S, Pandey A, Singh L, Zhang Z. 2020. Metagenomics for taxonomy profiling: tools and approaches. Bioengineered 11:356–374. https://doi.org/10.1080/21655979.2020.1736238.
17. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF. 2015. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol 16:1–18. https://doi.org/10.1186/s13059-015-0834-7.
18. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW, Roehe R, Watson M. 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun 9:1–11. https://doi.org/10.1038/s41467-018-03317-6.
19. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176:649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.
20. Hervé V, Liu P, Dietrich C, Sillam-Dussès D, Stiblik P, Šobotník J, Brune A. 2020. Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. PeerJ 8:e8614–27. https://doi.org/10.7717/peerj.8614.
21. Uritskiy G, Di Ruggiero J. 2019. Applying genome-resolved metagenomics to deconvolute the halophilic microbiome. Genes (Basel) 10:220. https://doi.org/10.3390/genes10030220.
22. Zheng J, Wittouck S, Salvetti E, Franz CMAP, Harris HMB, Mattarelli P, O'Toole PW, Pot B, Vandamme P, Walter J, Watanabe K, Wuyts S, Felis GE, Gänzle MG, Lebeer S. 2020. A taxonomic note on the genus *Lactobacillus*: description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. Int J Syst Evol Microbiol 70:2782–2858. https://doi.org/10.1099/ijsem.0.004107.
23. Marič L, Cleenwerck I, Accetto T, Vandamme P, Trček J. 2020. Description of *Komagataeibacter melaceti* sp. nov. and *Komagataeibacter melomenusus* sp. nov. isolated from apple cider vinegar. Microorganisms 8:1178. https://doi.org/10.3390/microorganisms8081178.
24. Pacheco-Montealegre ME, Dávila-Mora LL, Botero-Rute LM, Reyes A, Caro-Quintero A. 2020. Fine-resolution analysis of microbial communities provides insights into the variability of cocoa bean fermentation. Front Microbiol 11:650. https://doi.org/10.3389/fmicb.2020.00650.
25. Bobay LM, Ochman H. 2017. The evolution of bacterial genome architecture. Front Genet 8:72. https://doi.org/10.3389/fgene.2017.00072.
26. Ordoñez-Araque RH, Landines-Vera EF, Urresto-Villegas JC, Caicedo-Jaramillo CF. 2020. Microorganisms during cocoa fermentation: systematic review. Foods Raw Mater 8:155–162.
27. Evanovich E, De Souza Mendonça Mattos PJ, Guerreiro JF. 2019. Comparative genomic analysis of *Lactobacillus plantarum*: an overview. Int J Genomics 2019:4973214. https://doi.org/10.1155/2019/4973214.
28. Rodas AM, Chenoll E, Macián MC, Ferrer S, Pardo I, Aznar R. 2006. *Lactobacillus vini* sp. nov., a wine lactic acid bacterium homofermentative for pentoses. Int J Syst Evol Microbiol 56:513–517. https://doi.org/10.1099/ijs.0.63877-0.
29. Mendonça AA, da Silva PKN, Calazans TLS, de Souza RB, Elsztein C, de Morais Junior MA. 2020. Gene regulation of the *Lactobacillus vini* in response to industrial stress in the fuel ethanol production. Microbiol Res 236:126450. https://doi.org/10.1016/j.micres.2020.126450.



30. Zhang Q, Poehlein A, Hollensteiner J, Daniel R. 2018. Draft genome sequence of *Komagataeibacter maltaceti* LMG 1529T, a vinegar-producing acetic acid bacterium isolated from malt vinegar brewery acetifiers. Genome Announc 6:e00330-18. https://doi.org/10.1128/genomeA.00330-18.
31. Hollensteiner J, Poehlein A, Kloskowski P, Ali TT, Daniel R. 2020. Genome sequence of *Komagataeibacter saccharivorans* strain JH1, isolated from fruit flies. Microbiol Resour Announc 9:e00098-20. https://doi.org/10.1128/MRA.00098-20.
32. Lisdiyanti P, Kawasaki H, Widyastuti Y, Saono S, Seki T, Yamada Y, Uchimura T, Komagata K. 2002. *Kozakia baliensis* gen. nov., sp. nov., a novel acetic acid bacterium in the α-proteobacteria. Int J Syst Evol Microbiol 52:813–818. https://doi.org/10.1099/00207713-52-3-813.
33. Brandt JU, Jakob F, Wefers D, Bunzel M, Vogel RF. 2018. Characterization of an acetan-like heteropolysaccharide produced by *Kozakia baliensis* NBRC 16680. Int J Biol Macromol 106:248–257. https://doi.org/10.1016/j.ijbiomac.2017.08.022.
34. Guérin M, Silva CR-D, Garcia C, Remize F. 2020. Lactic acid bacterial production of exopolysaccharides from fruit and vegetables and associated benefits. Fermentation 6:115. https://doi.org/10.3390/fermentation6040115.
35. Jeong H, Arif B, Caetano-Anollés G, Kim KM, Nasir A. 2019. Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation. Sci Rep 9:1–18. https://doi.org/10.1038/s41598-019-42227-5.
36. Chen LX, Anantharaman K, Shaiber A, Murat Eren A, Banfield JF. 2020. Accurate and complete genomes from metagenomes. Genome Res 30:315–333. https://doi.org/10.1101/gr.258640.119.
37. Valera MJ, Torija MJ, Mas A, Mateo E. 2015. Cellulose production and cellulose synthase gene detection in acetic acid bacteria. Appl Microbiol Biotechnol 99:1349–1361. https://doi.org/10.1007/s00253-014-6198-1.
38. Sinha AK, Possoz C, Leach DRF. 2020. The roles of bacterial DNA double-strand break repair proteins in chromosomal DNA replication. FEMS Microbiol Rev 44:351–368. https://doi.org/10.1093/femsre/fuaa009.
39. Roosa S, Wattiez R, Prygiel E, Lesven L, Billon G, Gillan DC. 2014. Bacterial metal resistance genes and metal bioavailability in contaminated sediments. Environ Pollut 189:143–151. https://doi.org/10.1016/j.envpol.2014.02.031.
40. Lima LJR, van der Velpen V, Wolkers-Rooijackers J, Kamphuis HJ, Zwietering MH, Rob Nout MJ. 2012. Microbiota dynamics and diversity at different stages of industrial processing of cocoa beans into cocoa powder. Appl Environ Microbiol 78:2904–2913. https://doi.org/10.1128/AEM.07691-11.
41. Papalexandratou Z, Falony G, Romanens E, Jimenez JC, Amores F, Daniel HM, De Vuyst L. 2011. Species diversity, community dynamics, and metabolite kinetics of the microbiota associated with traditional Ecuadorian spontaneous cocoa bean fermentations. Appl Environ Microbiol 77:7698–7714. https://doi.org/10.1128/AEM.05523-11.
42. Ouattara HD, Ouattara HG, Droux M, Reverchon S, Nasser W, Niamke SL. 2017. Lactic acid bacteria involved in cocoa beans fermentation from Ivory Coast: species diversity and citrate lyase production. Int J Food Microbiol 256:11–19. https://doi.org/10.1016/j.ijfoodmicro.2017.05.008.
43. Olm MR, West PT, Brooks B, Firek BA, Baker R, Morowitz MJ, Banfield JF. 2019. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. Microbiome 7:1–16. https://doi.org/10.1186/s40168-019-0638-1.
44. Daniel HM, Vrancken G, Takrama JF, Camu N, De Vos P, De Vuyst L. 2009. Yeast diversity of Ghanaian cocoa bean heap fermentations. FEMS Yeast Res 9:774–783. https://doi.org/10.1111/j.1567-1364.2009.00520.x.
45. Fernández Maura Y, Balzarini T, Clapé Borges P, Evrard P, De Vuyst L, Daniel HM. 2016. The environmental and intrinsic yeast diversity of Cuban cocoa bean heap fermentations. Int J Food Microbiol 233:34–43. https://doi.org/10.1016/j.ijfoodmicro.2016.06.012.
46. Groenewald M, Robert V, Smith MT. 2011. The value of the D1/D2 and internal transcribed spacers (ITS) domains for the identification of yeast species belonging to the genus *Yamadazyma*. Persoonia 26:40–46. https://doi.org/10.3767/003158511X559610.
47. Gänzle MG, Follador R. 2012. Metabolism of oligosaccharides and starch in lactobacilli: a review. Front Microbiol 3:340. https://doi.org/10.3389/fmicb.2012.00340.
48. Buron-Moles G, Chailyan A, Dolejs I, Forster J, Mikš MH. 2019. Uncovering carbohydrate metabolism through a genotype-phenotype association study of 56 lactic acid bacteria genomes. Appl Microbiol Biotechnol 103:3135–3152. https://doi.org/10.1007/s00253-019-09701-6.
49. Edwards CG, Collins MD, Lawson PA, Rodriguez AV. 2000. *Lactobacillus nagelii* sp. nov., an organism isolated from a partially fermented wine. Int J Syst Evol Microbiol 50:699–702. https://doi.org/10.1099/00207713-50-2-699.
50. De Bruyne K, Camu N, De Vuyst L, Vandamme P. 2009. *Lactobacillus fabifermentans* sp. nov. and *Lactobacillus cacaonum* sp. nov., isolated from Ghanaian cocoa fermentations. Int J Syst Evol Microbiol 59:7–12. https://doi.org/10.1099/ijs.0.001172-0.
51. Ahmad MS, Zargar M, Mir S, Bhat N, Baba Z, Kant R, Habib Dar ZM, Khan IJ, Bandey S. 2018. Morphological and biochemical studies for the identification of *Lactobacillusplantarum* sp. nov., and *Lactobacillus fermentum* sp. nov., from municipal waste. J Pharmacogn Phytochem 7:1421–1424.
52. Zhou Q, Feng F, Yang Y, Zhao F, Du R, Zhou Z, Han Y. 2018. Characterization of a dextran produced by *Leuconostoc pseudomesenteroides* XG5 from homemade wine. Int J Biol Macromol 107:2234–2241. https://doi.org/10.1016/j.ijbiomac.2017.10.098.
53. Endo A, Futagawa-Endo Y, Dicks LMT. 2011. Influence of carbohydrates on the isolation of lactic acid bacteria. J Appl Microbiol 110:1085–1092. https://doi.org/10.1111/j.1365-2672.2011.04966.x.
54. Cousin FJ, Lynch SM, Harris HMB, McCann A, Lynch DB, Neville BA, Irisawa T, Okada S, Endo A, O'Toole PW. 2015. Detection and genomic characterization of motility in *Lactobacillus curvatus*: confirmation of motility in a species outside the *Lactobacillus salivarius* clade. Appl Environ Microbiol 81:1297–1308. https://doi.org/10.1128/AEM.03594-14.
55. Xu J, Verstraete W. 2001. Evaluation of nitric oxide production by lactobacilli. Appl Microbiol Biotechnol 56:504–507. https://doi.org/10.1007/s002530100616.
56. Elsanhoty RM, Ramadan MF. 2016. Genetic screening of biogenic amines production capacity from some lactic acid bacteria strains. Food Control 68:220–228. https://doi.org/10.1016/j.foodcont.2016.04.002.
57. Landete JM, Arena ME, Pardo I, Manca de Nadra MC, Ferrer S. 2010. The role of two families of bacterial enzymes in putrescine synthesis from agmatine via agmatine deiminase. Int Microbiol 13:169–177. https://doi.org/10.2436/20.1501.01.123.
58. del Rio B, Redruello B, Linares DM, Ladero V, Ruas-Madiedo P, Fernandez M, Martin MC, Alvarez MA. 2019. The biogenic amines putrescine and cadaverine show *in vitro* cytotoxicity at concentrations that can be found in foods. Sci Rep 9:120. https://doi.org/10.1038/s41598-018-36239-w.
59. Vrancken G, Rimaux T, Wouters D, Leroy F, De Vuyst L. 2009. The arginine deiminase pathway of *Lactobacillus fermentum* IMDO 130101 responds to growth under stress conditions of both temperature and salt. Food Microbiol 26:720–727. https://doi.org/10.1016/j.fm.2009.07.006.
60. Wang B, Shao Y, Chen T, Chen W, Chen F. 2015. Global insights into acetic acid resistance mechanisms and genetic stability of *Acetobacter pasteurianus* strains by comparative genomics. Sci Rep 5:18330–18314. https://doi.org/10.1038/srep18330.
61. Gomes RJ, de Fatima Borges M, de Freitas Rosa M, Castro-Gómez RJH, Spinosa WA. 2018. Acetic acid bacteria in the food industry: systematics, characteristics, and applications. Food Technol Biotechnol 56:139–151.
62. Kadere TT, Miyamoto T, Oniang'o RK, Kutima PM, Njoroge SM. 2008. Isolation and identification of the genera *Acetobacter* and *Gluconobacter* in coconut toddy (mnazi). Afr J Biotechnol 7:2963–2971.
63. Sievers M, Swings J. 2015. *Acetobacter*, p 1–7. *In* Bergey's manual of systematic archaea and bacteria. Wiley Interscience, New York, NY.
64. Noman AE, Barha NS, Al Sharaf AAM, Ali Q, Maqtari A, Mohedein A, Mohammed HH, Chen HF. 2020. OPEN A novel strain of acetic acid bacteria *Gluconobacter oxydans* FBFS97 involved in riboflavin production. Sci Rep 10:1–17. https://doi.org/10.1038/s41598-020-70404-4.
65. Shamala TR, Vijayendra SVN, Joshi GJ. 2012. Agro-industrial residues and starch for growth and co-production of polyhydroxyalkanoate copolymer and α-amylase by *Bacillus* sp. CFR-67. Braz J Microbiol 43:1094–1102. https://doi.org/10.1590/S1517-83822012000300036.
66. De Roos J, De Vuyst L. 2018. Acetic acid bacteria in fermented foods and beverages. Curr Opin Biotechnol 49:115–119. https://doi.org/10.1016/j.copbio.2017.08.007.
67. Racine KC, Lee AH, Wiersema BD, Huang H, Lambert JD, Stewart AC, Neilson AP. 2019. Development and characterization of a pilot-scale model cocoa fermentation system suitable for studying the impact of fermentation on putative bioactive compounds and bioactivity of cocoa. Foods 8:102–120. https://doi.org/10.3390/foods8030102.
68. Hinneh M, Semanhyia E, Van de Walle D, De Winne A, Tzompa-Sosa DA, Scalone GLL, De Meulenaer B, Messens K, Van Durme J, Afoakwa EO, De Cooman L, Dewettinck K. 2018. Assessing the influence of pod storage on sugar and free amino acid profiles and the implications on some Maillard reaction related flavor volatiles in Forastero cocoa beans. Food Res Int 111:607–620. https://doi.org/10.1016/j.foodres.2018.05.064.
69. Megias-Perez R, Moreno-Zambrano M, Behrends B, Corno M, Kuhnert N. 2020. Monitoring the changes in low molecular weight carbohydrates in cocoa beans



during spontaneous fermentation: a chemometric and kinetic approach. Food Res Int 128:108865. https://doi.org/10.1016/j.foodres.2019.108865.

70. Haase MAB, Kominek J, Langdon QK, Kurtzman CP, Hittinger CT. 2017. Genome sequence and physiological analysis of *Yamadazyma laniorum* f.a. sp. nov. and a reevaluation of the apocryphal xylose fermentation of its sister species, *Candida tenuis*. FEMS Yeast Res 17:fox019.

71. Ouattara HG, Reverchon S, Niamke SL, Nasser W. 2017. Regulation of the synthesis of pulp degrading enzymes in *Bacillus* isolated from cocoa fermentation. Food Microbiol 63:255–262. https://doi.org/10.1016/j.fm.2016.12.004.

72. Sandhya MVS, Yallappa BS, Varadaraj MC, Puranaik J, Rao LJ, Janardhan P, Murthy PS. 2016. Inoculum of the starter consortia and interactive metabolic process in enhancing quality of cocoa bean (*Theobroma cacao*) fermentation. LWT Food Sci Technol 65:731–738. https://doi.org/10.1016/j.lwt.2015.09.002.

73. Ramos CL, Dias DR, Miguel MGDCP, Schwan RF. 2014. Impact of different cocoa hybrids (*Theobroma cacao* L.) and *S. cerevisiae* UFLA CA11 inoculation on microbial communities and volatile compounds of cocoa fermentation. Food Res Int 64:908–918. https://doi.org/10.1016/j.foodres.2014.08.033.

74. Bushnell B, Rood J, Singer E. 2017. BBMerge: accurate paired shotgun read merging via overlap. PLoS One 12:e0185056. https://doi.org/10.1371/journal.pone.0185056.

75. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

76. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

77. Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res 9:868–877. https://doi.org/10.1101/gr.9.9.868.

78. Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607. https://doi.org/10.1093/bioinformatics/btv638.

79. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359. https://doi.org/10.7717/peerj.7359.

80. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. https://doi.org/10.1038/nmeth.3103.

81. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.

82. Yue Y, Huang H, Qi Z, Dou HM, Liu XY, Han TF, Chen Y, Song XJ, Zhang YH, Tu J. 2020. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinformatics 21:334. https://doi.org/10.1186/s12859-020-03667-3.

83. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.

84. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Genome Standards Consortium, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893.

85. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32:1088–1090. https://doi.org/10.1093/bioinformatics/btv697.

86. Saary P, Mitchell A, Finn R. 2019. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis. bioRxiv https://www.biorxiv.org/content/10.1101/2019.12.19.882753v2.full.

87. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Res 28:569–580. https://doi.org/10.1101/gr.228429.117.

88. Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics Chapter 4:Unit 4.6.1-10. https://doi.org/10.1002/0471250953.bi0406s35.

89. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

90. Von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol 20:1–14. https://doi.org/10.1186/s13059-019-1817-x.

91. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.

92. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.

93. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.

94. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

95. Grant JR, Stothard P. 2008. The CGView server: a comparative genomics tool for circular genomes. Nucleic Acids Res 36(Web Server Issue): W181–W184. https://doi.org/10.1093/nar/gkn179.

96. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC. 2016. From genomes to phenotypes: Traitar, the microbial trait analyzer. mSystems 1:e00101-16. https://doi.org/10.1128/mSystems.00101-16.

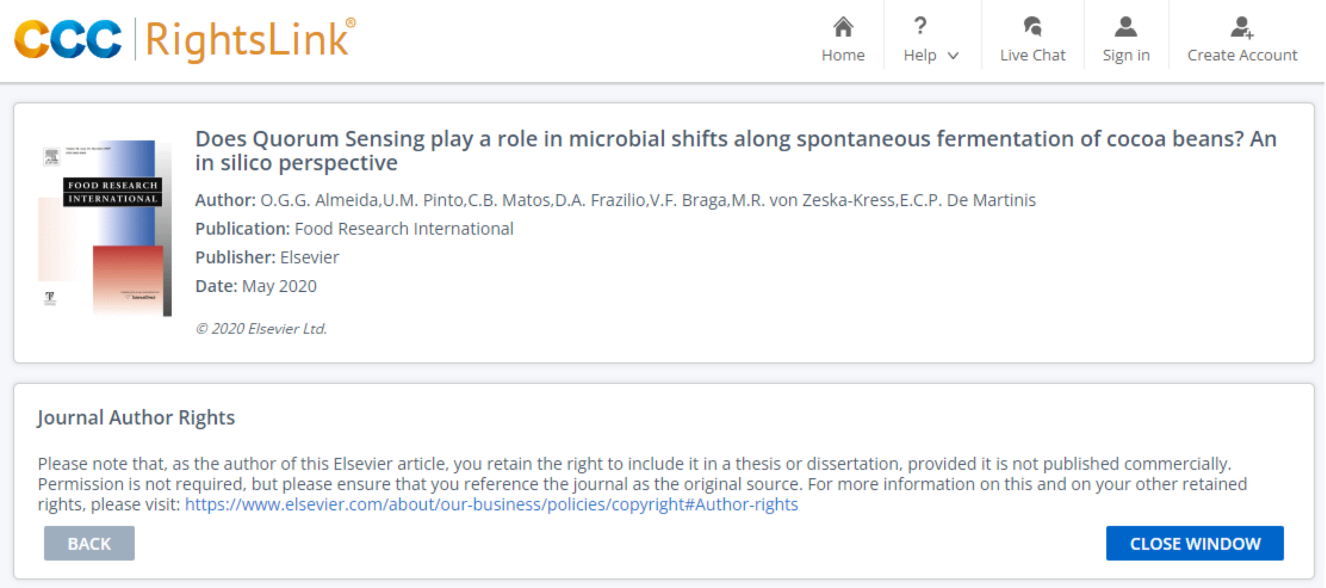
97. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365. https://doi.org/10.1038/srep08365.

98. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726–731. https://doi.org/10.1016/j.jmb.2015.11.006.

Attachments

Attachment A

Journal Author Rights of *Food Research International* for the article presented in Chapter 1.



Does Quorum Sensing play a role in microbial shifts along spontaneous fermentation of cocoa beans? An in silico perspective

Author: O.G.G. Almeida,U.M. Pinto,C.B. Matos,D.A. Frazilio,V.F. Braga,M.R. von Zeska-Kress,E.C.P. De Martinis

Publication: Food Research International

Publisher: Elsevier

Date: May 2020

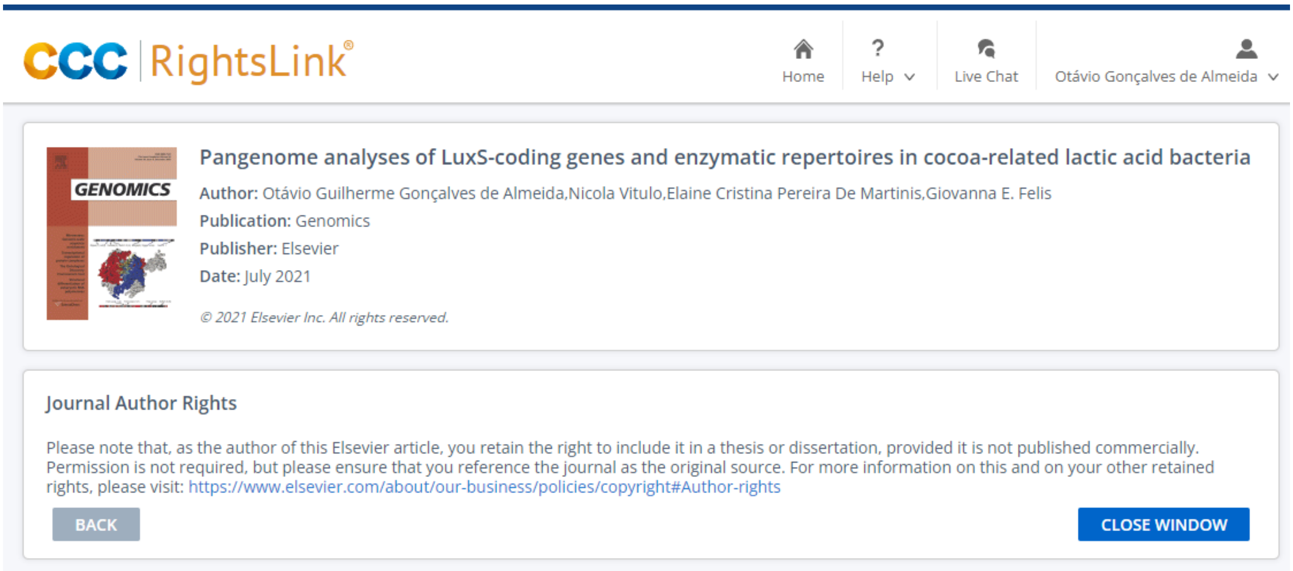
© 2020 Elsevier Ltd.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

Attachment B

Journal Author Rights of *Genomics* Journal for the article presented in the Chapter 2.



Pangenome analyses of LuxS-coding genes and enzymatic repertoires in cocoa-related lactic acid bacteria

Author: Otávio Guilherme Gonçalves de Almeida,Nicola Vitulo,Elaine Cristina Pereira De Martinis,Giovanna E. Felis

Publication: Genomics

Publisher: Elsevier

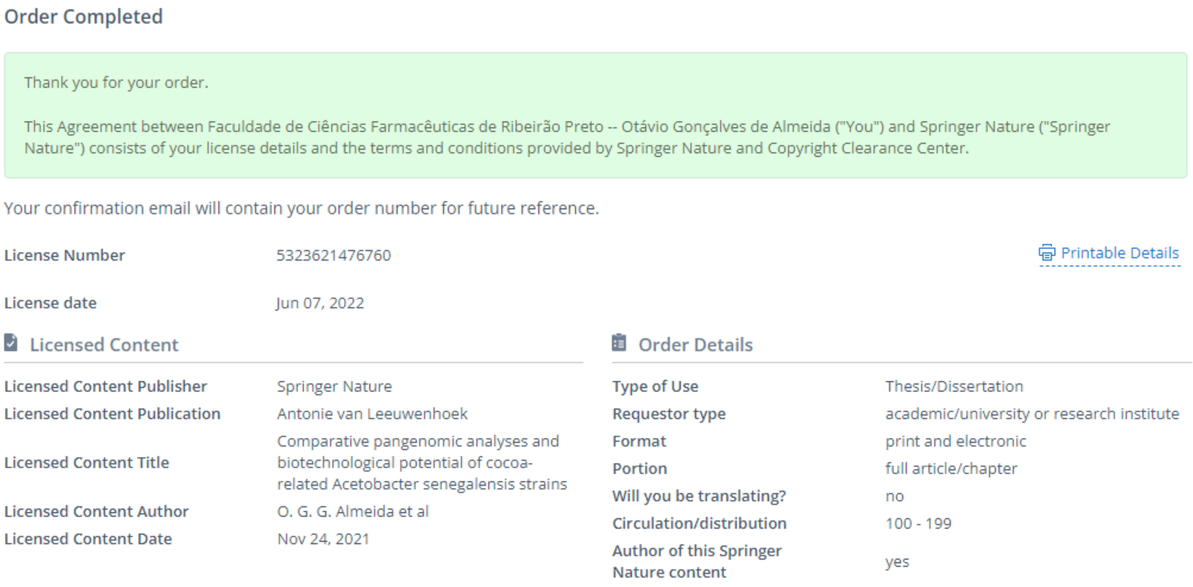
Date: July 2021

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

Attachment C

Journal Author Rights of *Antonie van Leeuwenhoek* Journal for the article presented in the Chapter 3.



Order Completed

Thank you for your order.

This Agreement between Faculdade de Ciências Farmacêuticas de Ribeirão Preto -- Otávio Gonçalves de Almeida ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number	5323621476760
License date	Jun 07, 2022

Licensed Content

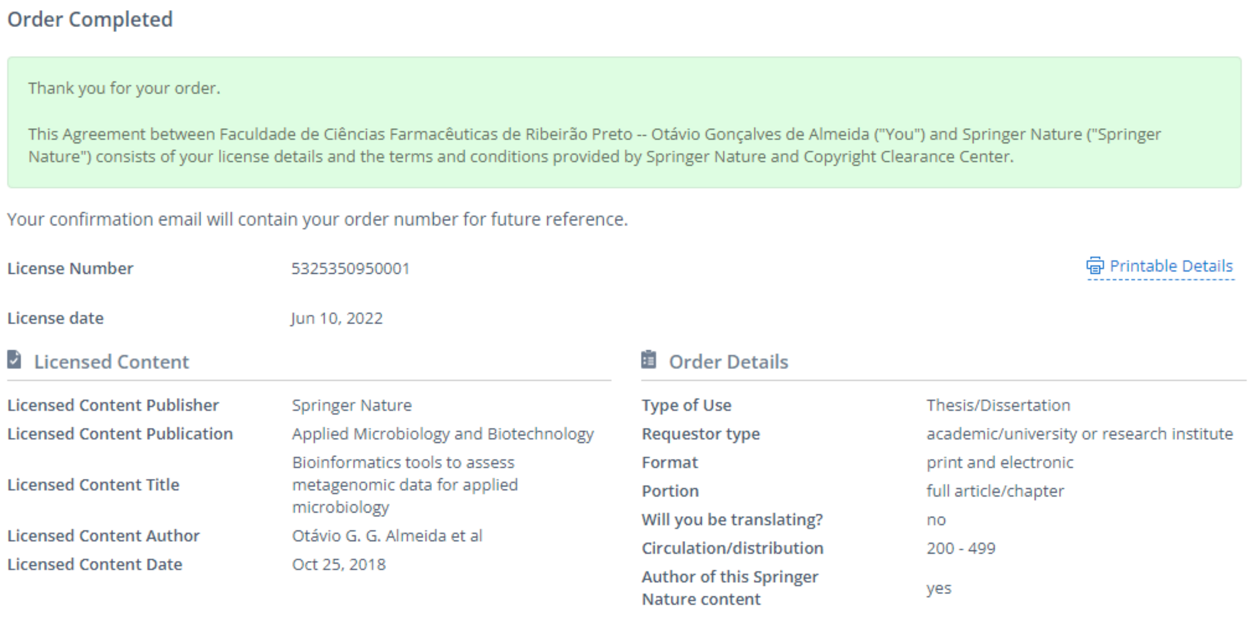
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Antonie van Leeuwenhoek
Licensed Content Title	Comparative pangenomic analyses and biotechnological potential of cocoa-related Acetobacter senegalensis strains
Licensed Content Author	O. G. G. Almeida et al
Licensed Content Date	Nov 24, 2021

Order Details

Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	100 - 199
Author of this Springer Nature content	yes

Attachment D

Journal Author Rights of *Applied Microbiology and Biotechnology* Journal for the article presented in the Appendix A.



Order Completed

Thank you for your order.

This Agreement between Faculdade de Ciências Farmacêuticas de Ribeirão Preto -- Otávio Gonçalves de Almeida ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number	5325350950001
License date	Jun 10, 2022

Licensed Content

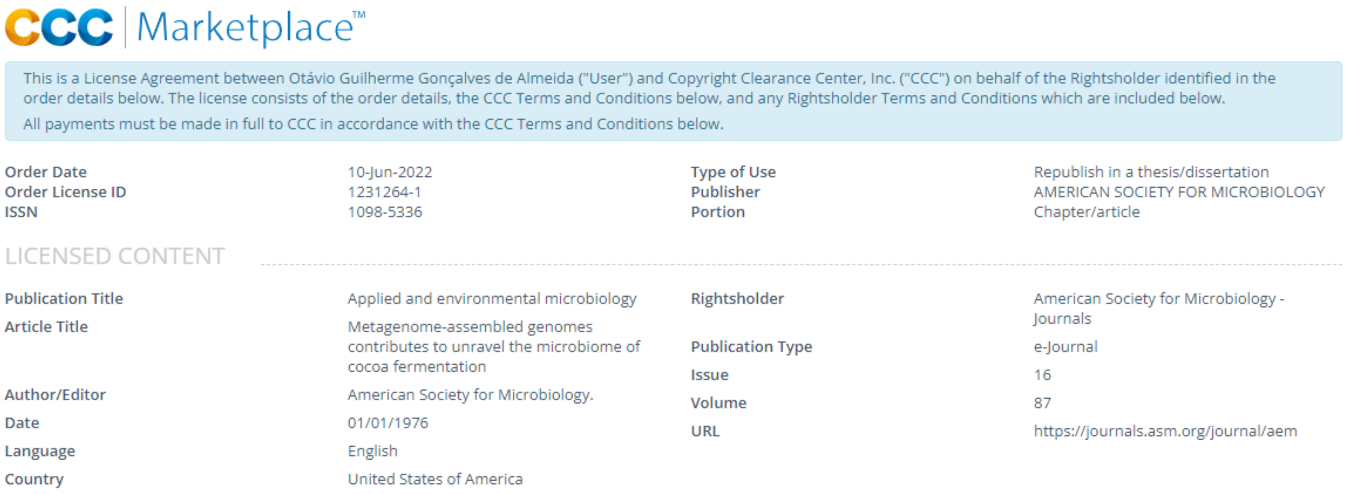
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Applied Microbiology and Biotechnology
Licensed Content Title	Bioinformatics tools to assess metagenomic data for applied microbiology
Licensed Content Author	Otávio G. G. Almeida et al
Licensed Content Date	Oct 25, 2018

Order Details

Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	200 - 499
Author of this Springer Nature content	yes

Attachment E

Journal Author Rights of *Applied and Environmental Microbiology* Journal for the article presented in the Appendix B.



CCC | Marketplace™

This is a License Agreement between Otávio Guilherme Gonçalves de Almeida ("User") and Copyright Clearance Center, Inc. ("CCC") on behalf of the Rightsholder identified in the order details below. The license consists of the order details, the CCC Terms and Conditions below, and any Rightsholder Terms and Conditions which are included below.

All payments must be made in full to CCC in accordance with the CCC Terms and Conditions below.

Order Date	10-Jun-2022	**Type of Use**	Republish in a thesis/dissertation
Order License ID	1231264-1	**Publisher**	AMERICAN SOCIETY FOR MICROBIOLOGY
ISSN	1098-5336	**Portion**	Chapter/article

LICENSED CONTENT

Publication Title	Applied and environmental microbiology	**Rightsholder**	American Society for Microbiology - Journals
Article Title	Metagenome-assembled genomes contributes to unravel the microbiome of cocoa fermentation	**Publication Type**	e-Journal
		Issue	16
Author/Editor	American Society for Microbiology.	**Volume**	87
Date	01/01/1976	**URL**	https://journals.asm.org/journal/aem
Language	English		
Country	United States of America		

UNIVERSIDADE DE SÃO PAULO

SCIENTIA VINCES