Phylogenetic Analysis of a Spontaneous Cocoa Bean Fermentation Metagenome Reveals New Insights into Its Bacterial and Fungal Community Diversity

Koen Illeghems, Luc De Vuyst, Zoi Papalexandratou, Stefan Weckx*

Research Group of Industrial Microbiology and Food Biotechnology (IMDO), Faculty of Sciences and Bio-engineering Sciences, Vrije Universiteit Brussel, Brussels, Belgium

Abstract

This is the first report on the phylogenetic analysis of the community diversity of a single spontaneous cocoa bean box fermentation sample through a metagenomic approach involving 454 pyrosequencing. Several sequence-based and composition-based taxonomic profiling tools were used and evaluated to avoid software-dependent results and their outcome was validated by comparison with previously obtained culture-dependent and culture-independent data. Overall, this approach revealed a wider bacterial (mainly *γ-Proteobacteria*) and fungal diversity than previously found. Further, the use of a combination of different classification methods, in a software-independent way, helped to understand the actual composition of the microbial ecosystem under study. In addition, bacteriophage-related sequences were found. The bacterial diversity depended partially on the methods used, as composition-based methods predicted a wider diversity than sequence-based methods, and as classification methods based solely on phylogenetic marker genes predicted a more restricted diversity compared with methods that took all reads into account. The metagenomic sequencing analysis identified *Hanseniaspora uvarum, Hanseniaspora opuntiae, Saccharomyces cerevisiae, Lactobacillus fermentum*, and *Acetobacter pasteurianus* as the prevailing species. Also, the presence of occasional members of the cocoa bean fermentation process was revealed (such as *Erwinia tasmaniensis*, *Lactobacillus brevis*, *Lactobacillus casei*, *Lactobacillus rhamnosus*, *Lactococcus lactis*, *Leuconostoc mesenteroides*, and *Oenococcus oeni*). Furthermore, the sequence reads associated with viral communities were of a restricted diversity, dominated by *Myoviridae* and *Siphoviridae*, and reflecting *Lactobacillus* as the dominant host. To conclude, an accurate overview of all members of a cocoa bean fermentation process sample was revealed, indicating the superiority of metagenomic sequencing over previously used techniques.

Citation: 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(5): e38040. doi:10.1371/journal.pone.0038040

Editor: Mark R. Liles, Auburn University, United States of America

Received February 23, 2012; **Accepted** May 1, 2012; **Published** May 29, 2012

Funding: This research was financed by the Research Council of the Vrije Universiteit Brussel, the Research Foundation Flanders (FWO; project number 1.5.108.09N) and a Titanium FLX-award of the Flanders Institute for Biotechnology (VIB) and Roche Diagnostics Belgium. KI is the receiver of a post-graduate Grant of the Agency for Innovation by Science and Technology (IWT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The Research Group of Industrial Microbiology and Food Biotechnology of the Vrije Universiteit Brussel received a Titanium FLX-award of the Flanders Institute for Biotechnology (VIB) and Roche Diagnostics Belgium. This award was in kind, consisting of a free run on a Genome Sequencer FLX system, for which Roche Diagnostics Belgium provided all materials needed. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: Stefan.Weckx@vub.ac.be

Introduction

Cocoa beans are seeds embedded in a mucilaginous pulp in fruit pods of the cocoa tree, *Theobroma cacao* L., and are used as the basic raw material for chocolate production [1,2]. The desired characteristic cocoa flavor and taste is obtained by fermenting, drying, and roasting of the raw cocoa beans [3,4]. The first step in cocoa processing is a spontaneous three- to six-day fermentation of the cocoa pulp-bean mass, in most cases carried out in heaps or boxes, wherein a succession of microbial activities of yeasts, involved in depectinization and ethanol formation, lactic acid bacteria (LAB), involved in citric acid fermentation and lactic acid production, and acetic acid bacteria (AAB), involved in the oxidation of ethanol produced by the yeasts into acetic acid and overoxidation of acetic acid and of lactic acid produced by LAB into carbon dioxide and water, takes place [5–7]. During fermentation, ethanol and acetic acid diffuse into the beans, and this, in combination with the heat produced during fermentation in general and during ethanol oxidation in particular, causes the death of the seed embryo. This step in turn initiates physical and biochemical changes in the beans, leading to the formation of precursor molecules for the development of a characteristic flavor and color of well-fermented cocoa beans [8–10].

During the last decade, the microbial diversity of spontaneous cocoa bean fermentation processes has been investigated through the application of culture-dependent and culture-independent techniques [11–22]. This has resulted in a better knowledge of this peculiar microbial ecosystem, which is dominated by species such as *Hanseniaspora* sp., *Saccharomyces cerevisiae*, *Lactobacillus fermentum*, *Lactobacillus plantarum*, and *Acetobacter pasteurianus*. However, it is known that both approaches have some drawbacks, undermining an accurate view on the microbial composition of this ecosystem, and implying that more, yet unidentified species, might play a role in the fermentation process. For instance, it has been shown that culture-dependent techniques can enhance the recovery of certain species that are not necessarily the most abundant or important ones in an ecosystem, thereby giving a non-accurate quantitative view [23]. To circumvent this drawback, culture-independent techniques such as denaturing gradient gel electrophoresis of small PCR amplicons of the targeted gene fragments (PCR-DGGE) or rRNA gene clone library sequencing have been used, also in the case of cocoa bean fermentation processes [7,14]. These techniques aim at the identification of both cultivable and yet uncultivable but potentially important players in a microbial ecosystem in a semi-quantitative way, thereby using whole microbial community (metagenomic) DNA. However, these methods might give a biased outcome for several reasons too, as they rely on PCR, thereby suffering from typical artifacts such as preferential DNA amplification. [24]. Moreover, PCR-DGGE is based on the amplification of several, rather small, variable regions of mostly the 16S (bacteria) or 26S rRNA genes (yeasts), of which the resolution within some genera is limited [25–27]. Recently, 454 pyrosequencing has been used to investigate the bacterial communities by sequencing of 16S rRNA gene amplicons solely [28,29]. This has also been done for fermented foods, such as nukadoko and kefir [30,31]. However, as the same short variable regions of the 16S rRNA genes as for PCR-DGGE are targeted, these gene fragments limit this approach.

Whole-community sequence data, obtained by high-throughput parallel sequencing of metagenomic DNA, overcome the limitations of the aforementioned culture-dependent and culture-independent techniques [32]. Concerning industrial fermentations involving bacteria, 454 pyrosequencing has recently been applied for assessing the prokaryotic community composition and functionality of, among others, a biogas fermentation process [33] and a kimchi fermentation process [34]. In the area of eukaryotic metagenomics, only a few studies involving whole-community pyrosequencing have been performed, with a focus on fungal diversity associated with soil and plants [35,36]. To our knowledge, the metagenomic approach has never been used to identify the members of a microbial ecosystem consisting of both prokaryotic and eukaryotic microorganisms, such as in the case of cocoa bean fermentation processes. Yet, to perform such taxonomic profiling, several computational methods are available, tackling either a composition-based [37–39] or a similarity-based [40–42] approach. It is unclear which of these methods result in the best estimate of microbial diversity. Indeed, similarity-based methods will only be accurate if a close evolutionary relative of a generated sequence (read) is present in the database [43] and these methods are known to be computationally expensive [40]. In the case of (supervised) composition-based methods, it is (often incorrectly) assumed that the genomes available in public databases are representative for the microorganisms present in the ecosystem [44]. Also, these methods can suffer from robustness when short sequences (<1 kb) are used [40].

Table 1. Statistics on the environmental reads of two GS FLX Titanium pyrosequencing runs of the metagenomic DNA of a Brazilian spontaneous cocoa bean box fermentation sample.

Statistical parameter	Data set A	Data set B
Number of reads	456,225	1,248,151
Total number of bases	200,550,104	551,635,450
Mean read length	439.08	441.96
Median read length	495	492
% G+C	49.74	49.54

doi:10.1371/journal.pone.0038040.t001

The aim of the present study was to investigate the microbial communities of a single sample of a spontaneous cocoa bean box fermentation process by performing 454 pyrosequencing on metagenomic DNA, and to compare the outcome with previous data of this sample to validate this metagenomic approach. [7,14]. Further, using these data, both similarity-based and composition-based computational methods for taxonomic profiling were evaluated and only operational taxonomic units (OTUs) that were consistently predicted were taken into account to avoid a software-dependent outcome. Hence, a complete and more reliable insight into the microbial diversity of the sample studied could be obtained. The results showed that 454 pyrosequencing can be used to identify the bacterial and fungal community members and to provide an insight into the viral communities of a cocoa bean fermentation sample. Analysis of bacterial diversity with multiple taxonomic profiling tools revealed differences in diversity estimates and abundance, which were consistent on different taxonomic ranks. Overall, a wider community diversity was retrieved compared with previous methods, indicating the superiority of metagenomic sequencing.

Materials and Methods

Total community DNA preparation, pyrosequencing, and sequence data quality control

A spontaneous cocoa bean box fermentation was performed at the 'Leão De Ouro' plantation in Ilhéus (Bahia, Brazil), as described previously [14]. A sample of 500 g was taken 30 h after the start of the fermentation, as at this time point, LAB and AAB species start to control the fermentation, while yeast species, involved during the first hours of the fermentation, are still present [14]. Whole-community metagenomic DNA was isolated in triplicate, each time from 20 g of the sample, as described previously, with minor modifications [12]. Briefly, a NucleoSpin column (Macherey Nagel GmbH, Düren, Germany) was used to remove cocoa pulp compounds, such as polysaccharides, proteins, enzymes, and polyphenols [45]. Furthermore, a second isopropanol precipitation step was applied after RNase treatment, to obtain pure high-quality DNA. The three DNA extracts were pooled and used as template for shotgun pyrosequencing on a Genome Sequencer (GS) FLX system (Roche Applied Science, Mannheim, Germany) using Titanium chemistry, which was performed by the VIB Nucleomics Core Facility (Leuven, Belgium). A DNA library was constructed according to the GS FLX Rapid Library Preparation Kit (Roche Applied Science). The optimal DNA copy per bead ratio was determined by an emulsion PCR titration using a GS FLX Titanium SV emPCR kit (Lib-L; Roche Applied Science). Final emulsion PCR for sequencing production runs was performed using the GS FLX Titanium LV emPCR kit (Lib-L; Roche Applied Science). Two independent pyrosequencing runs were carried out with this DNA library, the first one using two regions of a four-region gasket (half a PicoTiterPlate, the reads were represented by data set A) and the second one using a complete PicoTiterPlate (data set B). To assess the overall comparability of the sequence data sets, average G+C contents of all reads were determined for each data set. Therefore, various Perl scripts were developed to determine the overall and individual G+C contents of the reads. Artificially created duplicate reads were assessed using the Bioconductor software package ShortRead 1.6.2 [46] and cd-hit-454 [47].

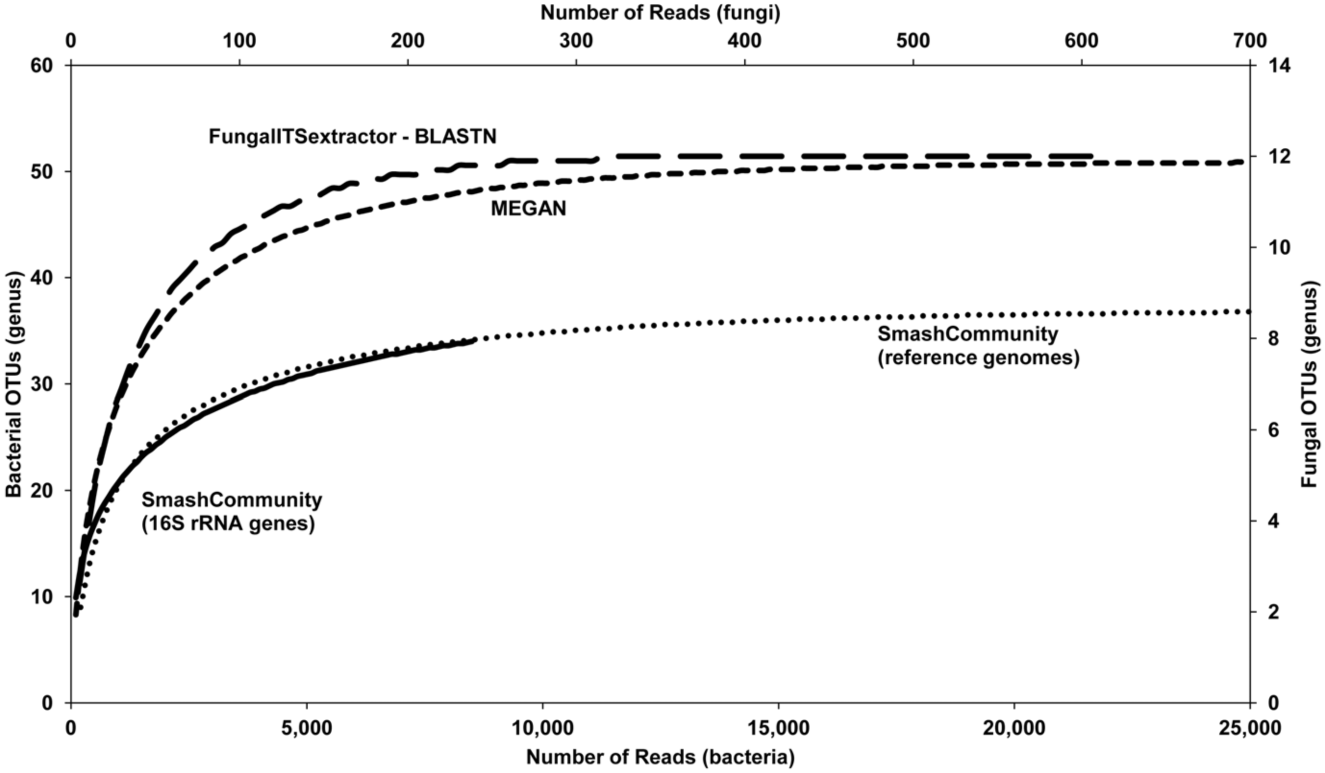

Figure 1. Rarefaction analysis of the genera found with data sets A and B. The rarefaction curves represent an estimation of the number of genera associated with different sampling sizes. As the results of the two 16S rRNA gene-based methods of the SmashCommunity platform were similar, only one method (based on similarity with the 16S rRNA gene sequence database of the SmashCommunity platform) is shown. As the plateau phase of the SmashCommunity reference genomes platform and MEGAN was reached at 25,000 reads, the X-axis is limited to this number of reads.
doi:10.1371/journal.pone.0038040.g001

Bacterial and fungal community richness estimation through rarefaction analysis

Binning of the reads for bacterial and fungal community richness estimation through rarefaction analysis was performed on rank ‘genus’ for data set A as well as for the combined data sets A and B. For bacterial rarefaction analysis, several similarity-based classification tools were used, including tools based on phylogenetic marker genes solely as well as tools that took all environmental reads into account. For the phylogenetic marker gene-based binning approach, SmashCommunity (version 1.5) [48] was used to extract 16S rRNA gene sequence fragments from the data sets. Therefore, two binning approaches (based on either a 16S rRNA gene database or on recognition by the meta_rrna tool) supported by this platform were used for detection of the 16S rRNA gene sequences, both using default parameters. This was followed by their classification with the Ribosomal Database Project (RDP) classifier using default parameters. The binning approach based on all environmental reads was performed with SmashCommunity and MEGAN (version 4.40.5) [42]. SmashCommunity was used to align all metagenomic reads to reference genomes through a BLASTN-based sequence similarity search of a SmashCommunity-compatible reference genome database (microbial reference genomes version 2.0; http://www.bork.embl.de/software/smash/). MEGAN was used with the min support set to 100, the min score set to 100, and the top percent set to 7. Hereto, all reads were aligned to the NCBI-nr database (National Center for Biotechnology Information, Bethesda, Maryland, USA) using the BLASTX algorithm.

Fungal rarefaction analysis was carried out with reads containing (part of) the internal transcribed spacer (ITS) regions ITS1 and/or ITS2. These regions were extracted from the data sets using the FungalITSextractor tool [49]. The reads containing a (partial) predicted ITS1-5.8S-ITS2 region were subsequently used in a BLASTN similarity search using the NCBI-nt database followed by their processing with MEGAN with the min support set to 2, the min score set to 100, and the top percent set to 1. At this stage, only hits within a fungal ITS region were taken into account to reduce false positives.

To assess whether all microbial species, both bacteria and fungi, of the spontaneous cocoa bean box fermentation sample under study were covered by the 454 pyrosequencing reads, rarefaction curves were constructed using the Analytic Rarefaction tool (version 1.3; www.uga.edu/strata/software/Software.html). The estimated numbers of genera associated with different sampling sizes of the environmental reads were expressed as OTUs. To avoid overestimation through misclassification, only genera that had an abundance of more than 0.01% of the data set were taken into account for bacterial rarefaction analysis.

Comparison of different taxonomic profiling tools to estimate the bacterial community composition

To assess the bacterial community composition of the sample under study, several software tools for taxonomic profiling were applied, using the combined data sets A and B. This was performed using both similarity- and composition-based software packages, designed for prokaryotic taxonomic profiling. All analyses were carried out on the taxonomic ranks phylum, class,

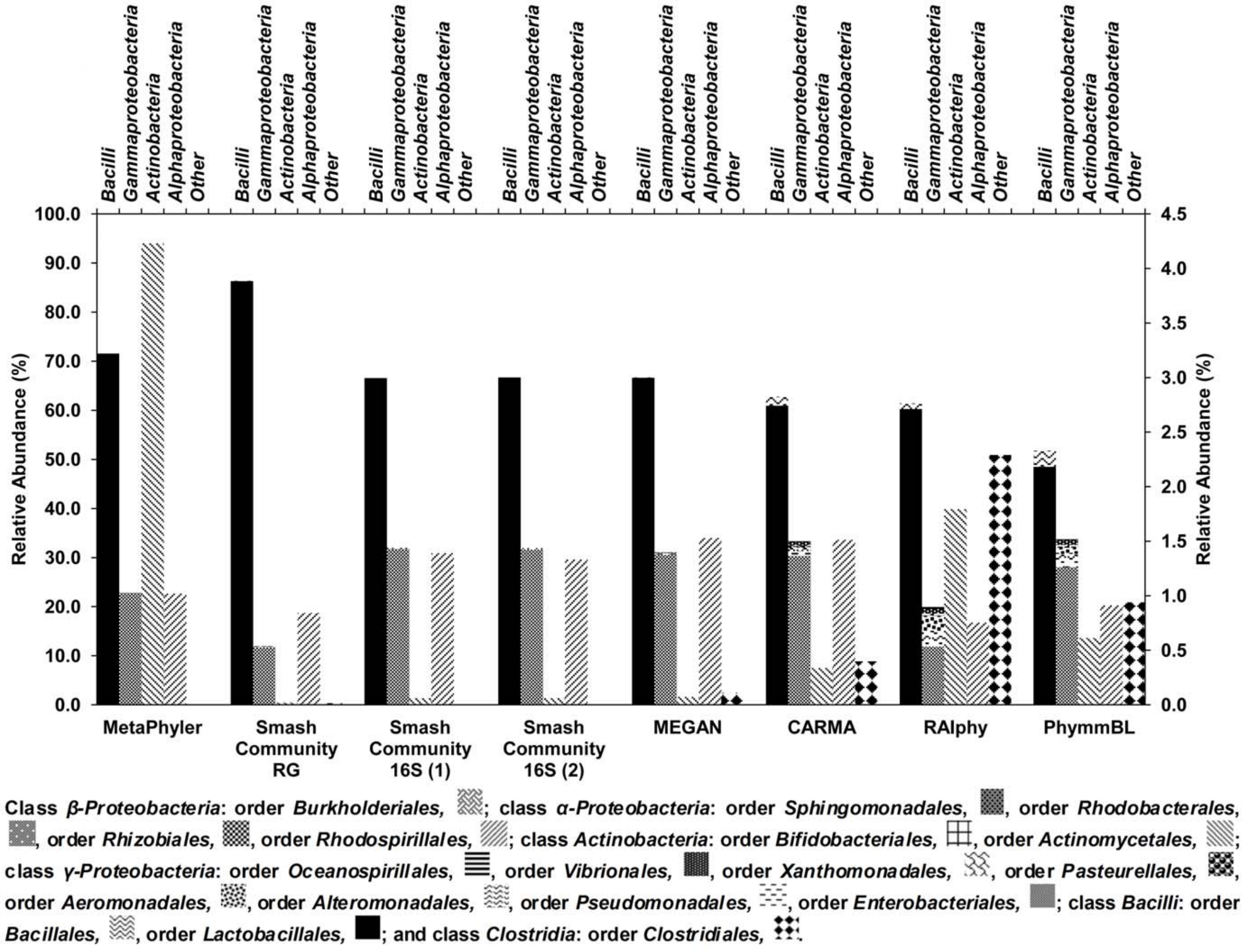

Figure 2. Bacterial composition analysis on ranks class and order by using different taxonomic profiling tools. Classes within the orders *Bacilli* and *γ-Proteobacteria* are shown on the left y-axis; classes within the orders *Actinobacteria*, *α-Proteobacteria*, and others are shown on the right y-axis. 'SmashCommunity RG' depicts SmashCommunity reference genomes, 'SmashCommunity 16S (1)' depicts the SmashCommunity 16S rRNA gene-based method using the meta_rrna approach, 'SmashCommunity 16S (2)' depicts the SmashCommunity 16S rRNA gene-based method using the 16S rRNA gene sequence database approach.
doi:10.1371/journal.pone.0038040.g002

order, family, and genus. To avoid software-dependent results, an OTU was only taken into account if it was predicted by at least five taxonomic profiling tools.

Applying a similarity-based analysis, tools based on extracting and classifying phylogenetic marker gene(s) (SmashCommunity, MetaPhyler) as well as tools based on classifying all environmental reads (SmashCommunity, MEGAN, CARMA) were used. For a phylogenetic marker gene-based analysis, SmashCommunity was used as described above. Analysis with MetaPhyler was performed using default parameters [50]. For an analysis based on all environmental reads, SmashCommunity and MEGAN were used as described above. CARMA (version 3) [51] was applied, based on a HMMER search using the Pfam database (version 24.0) [52].

Applying a composition-based analysis, RAIphy 1.0 [53] and PhymmBL [39] were used. For RAIphy, the binning threshold was set to 60 and a reference database was compiled based on the NCBI reference genomes. For PhymmBL, default parameters were used.

Taxonomic profiling to estimate the fungal community composition

The fungal community composition was assessed using two approaches. In a first approach, MEGAN analysis of the results of a BLASTX search against the NCBI-nr database of data set A was performed with the min support set to 15 and the min score set to 100. The second approach was based on reads extracted from the combined data sets A and B and originating from the ITS region, as described above.

Validation of the metagenomic approach

The results of the phylogenetic analysis of the metagenomic sequence data of the cocoa bean fermentation sample under study were compared with the results of former culture-dependent [(GTG)$_5$-PCR genomic fingerprinting of isolates] and culture-independent community composition analysis methods (PCR-DGGE and/or 16S rRNA gene clone library sequencing of sample DNA) [7,14]. A species was considered to be present in the ecosystem if it could be detected by all five prediction methods that were able to classify reads on rank species, namely SmashCom-

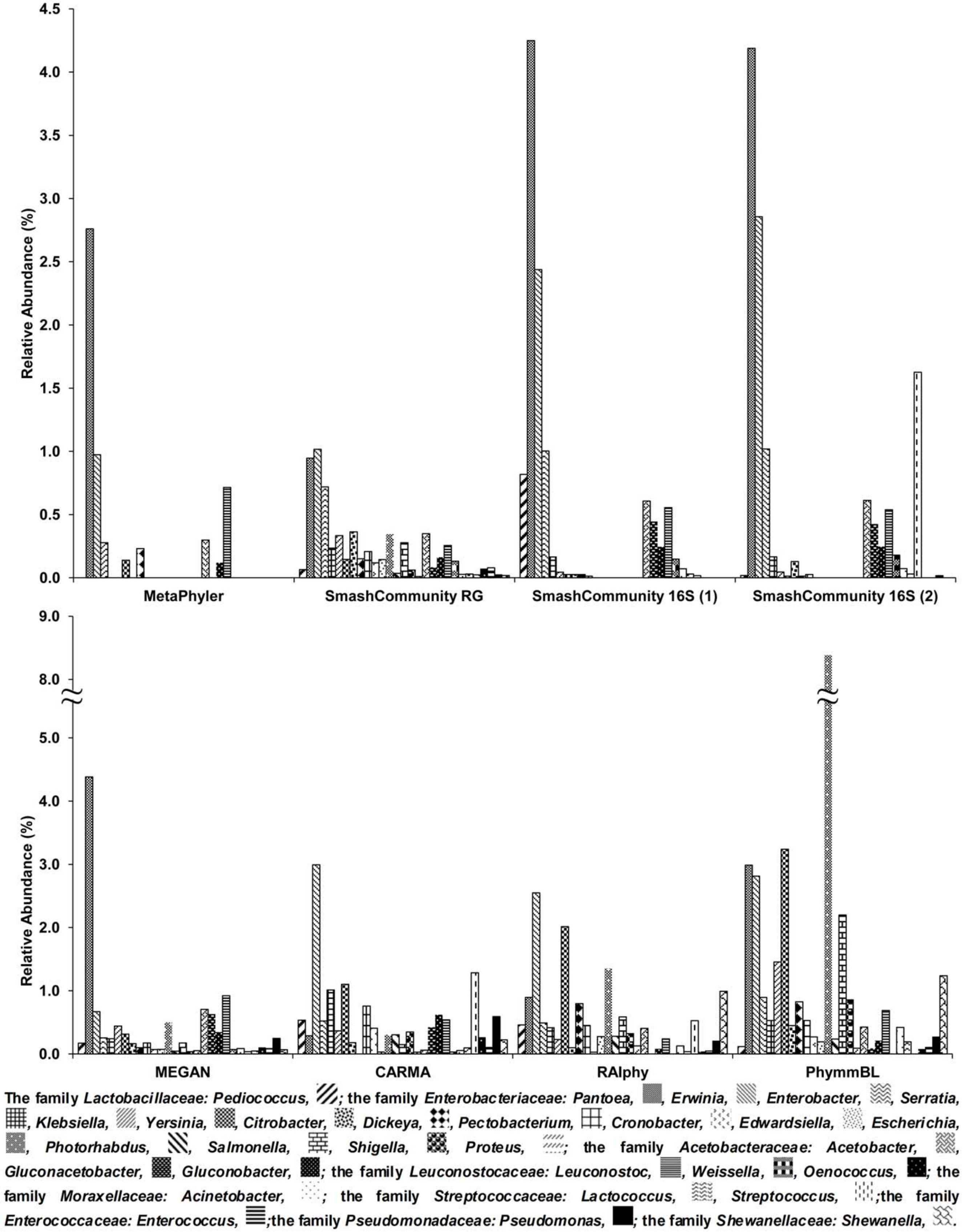

Figure 3. Bacterial composition analysis on rank genus of the low-abundance members by using different taxonomic profiling tools. 'SmashCommunity RG' depicts SmashCommunity reference genomes, 'SmashCommunity 16S (1)' depicts the SmashCommunity 16S rRNA

gene-based method using the meta_rrna approach, 'SmashCommunity 16S (2)' depicts the SmashCommunity 16S rRNA gene-based method using the 16S rRNA gene sequence database approach.
doi:10.1371/journal.pone.0038040.g003

munity reference genomes, MEGAN, CARMA, RAIphy, and PhymmBL.

Data availability

Sequence data from both GS FLX Titanium pyrosequencing runs were deposited in the NCBI Short Read Archive (SRA) under the accession number SRA049973.

Results and Discussion

Pyrosequencing and sequence data quality control

As the total bacterial diversity of a cocoa bean fermentation is limited [5,6] and as only a few microbial species dominate the fermentation process [7,17,22], only half a PicoTiterPlate was initially used for pyrosequencing of the metagenomic DNA library of a Brazilian cocoa bean box fermentation sample. This pyrosequencing run resulted in 456,225 reads with an average length of 439 bases, which accounted for approximately 201-Mb sequence information (data set A; Table 1). To achieve a deeper coverage of the metagenomic DNA and to elucidate its whole complexity, the same DNA library was used for a second pyrosequencing run using a whole PicoTiterPlate, which yielded 1,248,151 reads with an average length of 441 bases and resulted in 552-Mb sequence information (data set B; Table 1). Data set B represented a 2.7-fold increase in coverage of the DNA sample compared to data set A. As the same DNA library was used for both pyrosequencing runs, the G+C contents of the environmental reads were approximately the same (≈49.6%). No decrease in G+C contents for longer read sizes were found for both pyrosequencing runs (Fig. S1), indicating no bias towards microorganisms with a lower G+C content [54]. The Bioconductor software package ShortRead indicated only a few exact duplicates (Table S1), which was confirmed by cd-hit-454 (data not shown). To avoid underestimation of certain microbial groups by removing natural duplicates instead of artificial duplicates, no reads were removed from the data sets.

Bacterial and fungal community richness estimation through rarefaction analysis

The bacterial and fungal community richness of the reads of data set A was estimated using rarefaction curves based on taxonomic classification (Fig. S2). The rarefaction curves for bacteria indicated a gap between 16S rRNA gene-based methods and methods using all reads. Further, no saturation of the curves was reached with the 16S rRNA gene-based methods used. A

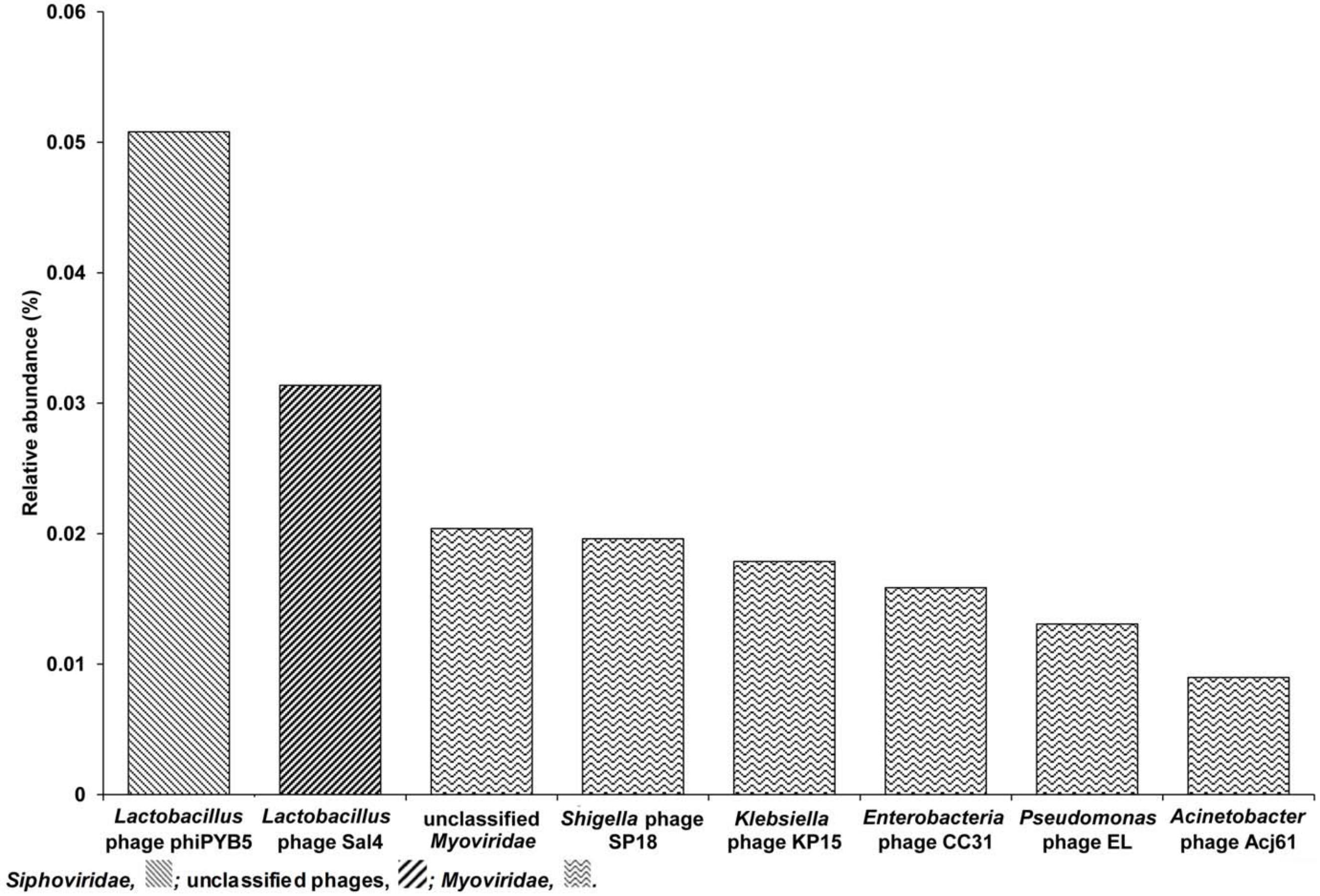

Figure 4. Reads classified as bacteriophages by MEGAN analysis on rank 'species'. The y-axis depicts the relative abundance compared to the total reads assigned by MEGAN.
doi:10.1371/journal.pone.0038040.g004

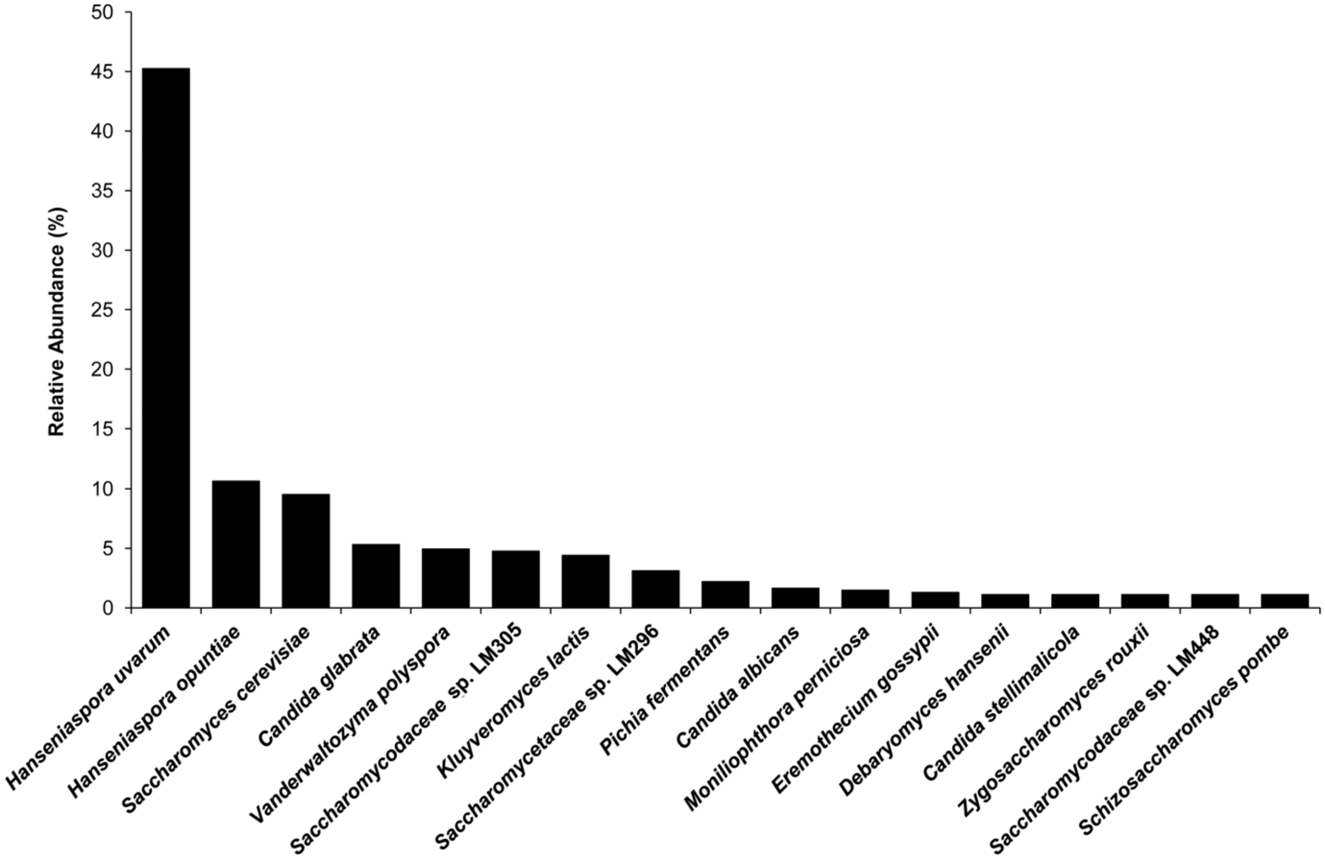

Figure 5. Diversity and richness of fungi on rank 'species'.
doi:10.1371/journal.pone.0038040.g005

rarefaction analysis of the combined data sets A and B revealed that saturation was reached, namely 36 OTUs for SmashCommunity 16S rRNA genes, 37 OTUs for SmashCommunity reference genomes, and 51 OTUs for MEGAN. This indicates that all bacterial members of the cocoa bean fermentation process sample were captured (Fig. 1).

The rarefaction curve for fungi based on data set A indicated that saturation was barely reached (Fig. S2). When using the combined data sets A and B, 755 reads containing the ITS1 and/or ITS2 rRNA gene regions were extracted. A rarefaction analysis of these sequence reads indicated saturation for the fungal communities of the cocoa bean fermentation ecosystem, namely 12 OTUs (Fig. 1).

Comparison of different taxonomic profiling tools to estimate the bacterial community composition

The estimation of the bacterial community diversity varied (for all different taxonomic ranks) between the taxonomic profiling tools when they were evaluated independently (Table S2, rows A). However, when OTUs were only taken into account if they were predicted by five or more different taxonomic profiling tools, a more reliable overview of the members of the ecosystem was obtained (Table S2, rows B). For example, the results of both the similarity-based and composition-based methods were in accordance, when they were applied for high taxonomic ranks. Indeed, on rank phylum, all tools were consistent in predicting the amount of OTUs, *i.e.*, a high abundance of *Firmicutes* and *Proteobacteria* and a low abundance of *Actinobacteria* (data not shown). Analysis on rank class revealed that *Bacilli* were the most abundant, among which *Lactobacillales* was the predominant order, although there was a wide variety of orders between the tools used (from 49% in the case of PhymmBL to 83% in the case of SmashCommunity reference genomes; Fig. 2). Several orders within the class *γ-Proteobacteria* occurred, although this class was dominated by members of the order *Enterobacteriales*. Members of the classes *Actinobacteria* (orders *Actinomycetales* and *Bifidobacteriales*), *α-Proteobacteria* (dominated by order *Rhodospirillales*), and *Clostridia* were also present in the ecosystem under study, but to a lower extent. On rank order, the two composition-based classification tools (RAIphy and PhymmBL) predicted a wider diversity (especially within the class *γ-Proteobacteria*) than tools involving similarity-based methods. Indeed, both composition-based methods predicted the presence of different orders within the *γ-Proteobacteria*, whereas these orders were not, or only to a very low extent, found using similarity-based methods. This discrepancy could be explained by the fact that composition-based methods are able to classify reads without the availability of close relatives in sequence databases, whereas similarity-based methods do not classify these reads on lower ranks if sequence similarity is not above a set threshold [53]. However, the extended bacterial diversity within the *γ-Proteobacteria* predicted by composition-based methods (besides the diverse order *Enterobacteriales*) was consistent for the different methods used. For all taxonomic profiling tools, the family of *Lactobacillaceae* was the most abundant. Moreover, all tools identified *Lactobacillus* as the dominant genus, although large differences were found (from 46% for PhymmBL to 94% for MetaPhyler). *Lactobacillus* is indeed a widespread genus associated with cocoa bean fermentation processes [5,6]. Analysis on rank genus of low-abundant members

Table 2. Comparison of different community composition analysis methods on rank species.

Species	Metagenomic DNA (%)	PCR-DGGE [14]	GTG$_5$-PCR [14]	16S rRNA gene clone library** [7]
Bacilli				
L. brevis	0.5–11.4			
L. casei	0.0–0.9			
L. fermentum	9.3–96.8	X*	X*	X
L. plantarum	0.3–10.2	X	X	
L. reuteri	0.0–2.6	X		
L. rhamnosus	0.0–0.2			
L. vaginalis				X
Lc. lactis	0.0–0.9			
Leuc. mesenteroides	0.1–1.6			
*Leuc. pseudoficulneus****		X*		
*Leuc. pseudomesenteroides****		X		
O. oeni	0.0–1.8			
P. acidilactici			X	
St. salivarius			X*	
α-Proteobacteria				
*A. fabarum****			X	
A. pasteurianus	0.1–0.8	X	X*	X
*A. senegalensis****			X*	
G. oxydans	0.0–3.9		X	
*Ga. saccharivorans****			X	
γ-Proteobacteria				
E. coli	0.0–8.1			
En. cloacae	0.0–0.6			
Er. amylovora	0.0–0.8			
Er. tasmaniensis	0.0–5.6			
K. pneumoniae	0.1–1.9			
Pe. carotovorum	0.0–0.2			
S. enterica	0.0–2.2			
T. citrea		X*		
*T. ptyseos****		X*		

*Detected at 30 h.

**No analysis performed at 30 h.

***Only phylogenetic marker gene(s) sequences available in databases.

Species are only considered as present in the ecosystem if they could be detected by all five taxonomic profiling tools. The relative species abundances, predicted by the different classification tools, are expressed as a range that represent the lowest and highest values obtained. For the other methods, the presence of a species is denoted by ''X''. *A.*: *Acetobacter*; *E.*: *Escherichia*; *En.*: *Enterobacter*; *Er.*: *Erwinia*; *G.*: *Gluconobacter*; *Ga.*: *Gluconacetobacter*; *K.*: *Klebsiella*; *L.*: *Lactobacillus*; *Leuc.*: *Leuconostoc*; *Lc.*: *Lactococcus*; *O.*: *Oenococcus*; *P.*: *Pediococcus*; *Pe.*: *Pectobacterium*; *S.*: *Salmonella*; *St.*: *Streptococcus*; *T.*: *Tatumella*.

doi:10.1371/journal.pone.0038040.t002

only, omitting the genus *Lactobacillus*, revealed differences in predicted OTUs between the classification tools (Fig. 3). Indeed, the three classification tools based on phylogenetic marker genes solely (both SmashCommunity 16S rRNA gene-based methods and MetaPhyler) predicted a restricted diversity compared with tools that took all reads into account. This difference could be ascribed to misclassifications of the latter tools, since binning of reads originating from non-phylogenetic marker genes is more prone to error; alternatively, it could be due to a failure of phylogenetic marker gene(s)-based methods [50]. The latter could originate from misclassifications due to absence of the phylogenetic marker gene sequences in the underlying database, or absence of phylogenetic marker genes in the data set because of insufficient sequencing. Hence, a phylogenetic analysis using only one taxonomic profiling tool should be interpreted carefully. However, predictions on rank genus by tools that took all reads into account were consistent, although some clear differences in abundance of these genera was seen. For instance, the PhymmBL tool classified 8.4% of the reads as *Escherichia*, which was a higher abundance compared with any other classification tool that classified only 0.0 to 1.4% of the reads as *Escherichia*. Database bias towards model bacteria such as *Escherichia coli* might explain this [55,56].

Taxonomic profiling with the MEGAN package revealed the presence of 4,296 reads (0.25% of the total amount of reads) that originated from bacteriophages and that, therefore, were classified as viruses. These viral communities were dominated by *Lactobacillus* phages, although a few other bacterial hosts such as *Enterobacter* and *Klebsiella* were found as well (Fig. 4). As the DNA isolation method targeted bacteria and yeasts, it could be assumed that these reads were indeed from viral origin (such as prophages or remnants of bacteriophages), which were incorporated into the bacterial genomes, as lactobacilli often harbor phage DNA [57–59]. Indeed, a striking similarity between the dominant bacterial genera (*Lactobacillus*) and the dominant predicted bacteriophage hosts (*Lactobacillus*) was found, supporting the assumption that an interaction exists between bacterial hosts and the viral communities [60]. A restricted diversity within these phage DNA-associated sequences was found, as only members of the families *Siphoviridae* and *Myoviridae*, which belong to the order *Caudovirales*, were retrieved. Similarly, the viral communities of fermented shrimp, kimchi, and sauerkraut are dominated by bacteriophages belonging to the viral order *Caudovirales* [61]. This is the first report on the occurrence of bacteriophages of lactobacilli in a cocoa bean fermentation sample. However, it is well known that bacteriophages are associated with LAB involved in food fermentation processes [62,63].

Taxonomic profiling of the fungal community composition

Fungal community composition analysis of data set A classified only 2,032 reads (0.16%) within the kingdom *Fungi*. Almost all reads were classified on high taxonomic ranks. Only 268 out of the 2,032 reads could be classified on genus or species level, the latter being classified as *Hanseniaspora uvarum*, *Kluyveromyces lactis*, *Lachancea thermotolerans*, *Pichia angusta*, *Saccharomyces cerevisiae*, and *Zygosaccharomyces rouxii*. This indicates that a BLASTX-based MEGAN analysis of a whole-community metagenomic data set was not suitable to classify the reads on a low taxonomic rank. However, a combination of extracting reads containing a (partial) ITS region and a subsequent BLASTN similarity search was able to classify a total of 755 reads on rank species (Fig. 5). The present metagenomic analysis indicates that the most prevailing yeast was *H. uvarum*, followed by *Hanseniaspora opuntiae*, and *S. cerevisiae*, which accounted for 45.2%, 10.6%, and 9.5% of all yeast DNA, respectively. These species are commonly associated with cocoa bean fermentation processes [20,64]. Further, other species commonly occurring during cocoa bean fermentations were found in the current sample as well, such as *Candida glabrata*, *K. lactis*, *Pichia fermentans*, *Debaryomyces hansenii*, *Candida stellimalicola*, *Schizosaccharomyces pombe*, and species of the families *Saccharomycodaceae* and *Saccharomycetaceae*. In addition, fungal species that were not yet associated with cocoa bean fermentations were identified. For instance, *Vanderwaltozyma polyspora* is a yeast species previously isolated from a soil ecosystem [65] and *Z. rouxii* has been reported in miso and soy sauce fermentations [66]. *Moniliophthora perniciosa* and *Eremothecium gossypii*, both plant pathogenic filamentous fungi, and the human pathogenic *Candida albicans* were found as well, but their identification could be the result of database bias towards pathogenic fungal species [55,56]. However, as *M. perniciosa* causes witches' broom disease of cocoa trees [67], it is not surprising that this species was present in fermenting cocoa pulp-bean mass.

Validation of the metagenomic approach

A comparison of the results of the phylogenetic analysis using different computational methods with former results of culture-dependent and culture-independent community composition analyses revealed that the metagenomic approach was able to retrieve most of the previously identified members (Table 2). This was even the case on rank species, which is generally regarded as inaccurate [68]. *Lactobacillus fermentum* and *A. pasteurianus* were identified as the prevailing LAB and AAB species, respectively, which was in accordance with 16S rRNA gene-PCR-DGGE and $(GTG)_5$-PCR fingerprinting analyses of this 30-h fermentation sample [7,14]. This underlines the functional role of both species during cocoa bean fermentation [6,69]. Also, *L. plantarum*, *Lactobacillus reuteri*, and *G. oxydans* were identified by metagenomic sequencing of the 30-h fermentation sample, whereas these species were not found by 16S rRNA gene-PCR-DGGE and/or $(GTG)_5$-PCR analysis. Further, the metagenomic analysis revealed the presence of several bacterial species, which were not detected in this fermentation sample by culture-dependent and/or culture-independent analysis. This included opportunistic members of the cocoa bean fermentation process, such as *E. tasmaniensis*, *Lactobacillus brevis*, *Lactobacillus casei*, *Lactococcus lactis*, *Leuconostoc mesenteroides*, and *Oenococcus oeni* [3,7,12,18,70–72]. Additionally, some bacterial species were found that were not yet detected during cocoa bean fermentation processes, such as *Lactobacillus rhamnosus*, an intestinal inhabitant [73]. Further, *Pectobacterium carotovorum* and *Erwinia amylovora* are phytopathogens, causing potato rot diseases and wilt diseases on *Rosaceae*, respectively [74]. The occurrence of *Escherichia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, and *Enterobacter cloacae*, might be unexpected, although contamination with gastro-intestinal (pathogenic) bacteria may occur. However, as only relatively few reads were classified within these species ($<1\%$), their presence could be the result of an overestimation due to a bias towards (pathogenic) model bacteria in the databases used [55,56]. In contrast, other species such as *Fructobacillus pseudoficulneus*, *Acetobacter senegalensis*, and *Tatumella ptyseos* were not retrieved using the metagenomic approach. This was probably due to the lack of sequence information of these species, for which only one or a few phylogenetic marker gene(s) are available, in public databases that were used to perform the taxonomic profiling. Further, it could be due to differences in homogeneity in sample material in the case these species display a low abundance at the particular time point investigated.

Former results indicated that the prevailing yeast in this cocoa bean fermentation sample is a *Hanseniaspora* species, most likely *H. opuntiae* [64]. However, this species could not be distinguished from *H. uvarum* and *Hanseniaspora guilliermondii* through 26S rRNA gene-PCR-DGGE [64]. In contrast, the metagenomic sequencing approach of the present study revealed that *H. uvarum* is the prevailing yeast at this time point. Indeed, using reads originating from the ITS region, it was possible to differentiate between *H. uvarum* and *H. opuntiae*. Hence, it is likely to assume that *H. uvarum*, *H. opuntiae*, and *S. cerevisiae* are the prevailing yeasts during cocoa bean fermentation. Additionally, whereas only a few yeast species were detected before, a wide fungal diversity was found in this sample using a metagenomic approach.

Conclusions

This study is the first report on the taxonomic analysis of a cocoa bean fermentation sample using a metagenomic approach, *i.e.*, 454 pyrosequencing of whole-community DNA. It was shown that this approach, when applying two pyrosequencing runs to obtain a high depth of coverage, was suitable to reveal both dominant and rare bacterial and fungal members of the cocoa bean fermentation ecosystem at a certain time point and to identify associated bacteriophages. However, this approach does not provide information about the community dynamics throughout the whole fermentation process. A combination of different similarity-based and composition-based methods, including both phylogenetic marker gene(s)-based analysis as well as methods using all available sequence information, appears to be necessary to obtain a credible view on the microbial community diversity of a complex microbial ecosystem, such as the cocoa bean fermentation process. Dominant species were *H. uvarum*, *H. opuntiae*, *S. cerevisiae*, *L. fermentum*, and *A. pasteurianus*, which was in accordance with former culture-dependent and culture-independent community analysis methods and underlines their importance in cocoa bean fermentations. In addition, sequence reads associated with viral communities were found, representing only members of the families *Myoviridae* and *Siphoviridae*, with *Lactobacillus* as the dominant microbial host, which is in accordance with the microbial phylogenetic analysis. These results indicate the superiority of metagenomic sequencing over previously used techniques for a phylogenetic characterization of complex matrices such as that involved in the cocoa bean fermentation process. The wider diversity retrieved in the present study is of importance to generate further insights into the functional roles of bacteria, fungi, and bacteriophages during cocoa bean fermentation, which is of great importance to select an appropriate starter culture for homogeneous, fast, and successfully controlled processes [75–77].

Supporting Information

Figure S1 Distribution of the average G+C content as a function of read length. Two pyrosequencing data sets were used, which were the result of a pyrosequencing run using half a PicoTiterPlate and a complete PicoTiterPlate.
(TIF)

Figure S2 Rarefaction analysis of the genera found with data set A. The rarefaction curves represent an estimation of the number of genera associated with different sampling sizes. As the results of the two 16S rRNA gene-based methods of the SmashCommunity platform were similar, only one method (based on the similarity with a 16S rRNA gene sequence database of the SmashCommunity platform) is shown. As the plateau phase of the SmashCommunity reference genomes platform and MEGAN was reached at 25,000 reads, the X-axis is limited to this number of reads.
(TIF)

Table S1 Duplicate reads. Data set A refers to the sequencing run using two regions of a four-region gasket (half a PicoTiterPlate); data set B refers to the sequencing run of a complete PicoTiterPlate.
(DOC)

Table S2 Bacterial community diversity estimations for the eight taxonomic profiling tools used. For each of the taxonomic profiling tools used, the numbers in row A refer to the originally estimated OTUs per rank; the numbers in row B refer to a subsection of the OTUs in row A that were also estimated by at least four other taxonomic profiling tools. The numbers between brackets depict the percentage of reads used to estimate the number of OTUs in row A that are included by the OTUs in row B. 'SmashCommunity RG' depicts SmashCommunity reference genomes, 'SmashCommunity 16S (1)' depicts the SmashCommunity 16S rRNA gene-based method using the meta_rrna approach, 'SmashCommunity 16S (2)' depicts the SmashCommunity 16S rRNA gene-based method using the 16S rRNA gene sequence database approach.
(DOC)

Author Contributions

Conceived and designed the experiments: KI LDV SW. Performed the experiments: KI ZP. Analyzed the data: KI SW. Contributed reagents/materials/analysis tools: KI LDV ZP SW. Wrote the paper: KI LDV SW.

References

1. Becket ST (2009) Industrial Chocolate Manufacture and Use. Chichester, United Kingdom: John Wiley & Sons. 688 p.
2. Afoakwa EO (2010) Chocolate Science and Technology. Oxford, United Kingdom: John Wiley & Sons. 311 p.
3. Thompson SS, Miller KB, Lopez AS (2007) Cocoa and coffee. In: Doyle MP, Beuchat LR, eds. Food Microbiology: Fundamentals and Frontiers. 3rd ed. Washington DC, USA: American Society for Microbiology. pp 837–849.
4. Wood GAR, Lass RA (2001) Cocoa. Oxford, United Kingdom: Blackwell Science. 620 p.
5. Schwan RF, Wheals AE (2004) The microbiology of cocoa fermentation and its role in chocolate quality. Crit Rev Food Sci Nutr 44: 205–221.
6. De Vuyst L, Lefeber T, Papalexandratou Z, Camu N (2010) The functional role of lactic acid bacteria in cocoa bean fermentation. In: Mozzi F, Raya RR, Vignolo GM, eds. Biotechnology of Lactic Acid Bacteria: Novel Applications: Wiley-Blackwell. pp 301–325.
7. Garcia-Armisen T, Papalexandratou Z, Hendryckx H, Camu N, Vrancken G, et al. (2010) Diversity of the total bacterial community associated with Ghanaian and Brazilian cocoa bean fermentation samples as revealed by a 16S rRNA gene clone library. Appl Microbiol Biotechnol 87: 2281–2292.
8. Hansen CE, del Olmo M, Burri C (1998) Enzyme activities in cocoa beans during fermentation. J Sci Food Agric 77: 273–281.
9. Afoakwa EO, Paterson A, Fowler M, Ryan A (2008) Flavor formation and character in cocoa and chocolate: a critical review. Crit Rev Food Sci Nutr 48: 840–857.
10. Rebelo-Lima LJR, Almeida MH, Nout MJR, Zwietering MH (2011) *Theobroma cacao L.*, "The Food of the Gods": quality determinants of commercial cocoa beans, with particular reference to the impact of fermentation. Crit Rev Food Sci Nutr 51: 731–761.
11. Kostinek M, Ban-Koffi L, Ottah-Atikpo M, Teniola D, Schillinger U, et al. (2008) Diversity of predominant lactic acid bacteria associated with cocoa fermentation in Nigeria. Curr Microbiol 56: 306–314.
12. Camu N, De Winter T, Verbrugghe K, Cleenwerck I, Vandamme P, et al. (2007) Dynamics and biodiversity of populations of lactic acid bacteria and acetic acid bacteria involved in spontaneous heap fermentation of cocoa beans in Ghana. Appl Environ Microbiol 73: 1809–1824.
13. 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 Microbiol 28: 964–973.
14. 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.
15. Papalexandratou Z, Falony G, Romanens E, Jimenez JC, Amores F, et al. (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.
16. Jespersen L, Nielsen DS, Hønholt S, Jakobsen M (2005) Occurrence and diversity of yeasts involved in fermentation of West African cocoa beans. FEMS Yeast Res 5: 441–453.
17. Nielsen DS, Hønholt S, Tano-Debrah K, Jespersen L (2005) Yeast populations associated with Ghanaian cocoa fermentations analysed using denaturing gradient gel electrophoresis (DGGE). Yeast 22: 271–284.
18. Nielsen DS, Teniola OD, Ban-Koffi L, Owusu M, Andersson TS, et al. (2007) The microbiology of Ghanaian cocoa fermentations analysed using culture-dependent and culture-independent methods. Int J Food Microbiol 114: 168–186.
19. Lagunes-Gálvez S, Loiseau G, Paredes JL, Barel M, Guiraud J-P (2007) Study on the microflora and biochemistry of cocoa fermentation in the Dominican Republic. Int J Food Microbiol 114: 124–130.
20. Daniel H-M, Vrancken G, Takrama JF, Camu N, De Vos P, et al. (2009) Yeast diversity of Ghanaian cocoa bean heap fermentations. FEMS Yeast Res 9: 774–783.
21. Camu N, González Á, De Winter T, Van Schoor A, De Bruyne K, et al. (2008) 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. Appl Environ Microbiol 74: 86–98.
22. Ardhana MM, Fleet GH (2003) The microbial ecology of cocoa bean fermentations in Indonesia. Int J Food Microbiol 86: 87–99.
23. Giraffa G, Neviani E (2001) DNA-based, culture-independent strategies for evaluating microbial communities in food-associated ecosystems. Int J Food Microbiol 67: 19–34.
24. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol 71: 8966–8969.

25. Mollet C, Drancourt M, Raoult D (1997) *rpoB* sequence analysis as a novel basis for bacterial identification. Mol Microbiol 26: 1005–1011.
26. Devriese LA, Vancanneyt M, Descheemaeker P, Baele M, Van Landuyt HW, et al. (2002) Differentiation and identification of *Enterococcus durans*, *E. hirae* and *E. villorum*. J Appl Microbiol 92: 821–827.
27. Daniel HM, Meyer W (2003) Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. Int J Food Microbiol 86: 61–78.
28. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, et al. (2007) Microbial population structures in the deep marine biosphere. Sci 318: 97–100.
29. Neufeld JD, Mohn WW (2005) Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. Appl Environ Microbiol 71: 5710–5718.
30. Dobson A, O'Sullivan O, Cotter PD, Ross P, Hill C (2011) High-throughput sequence-based analysis of the bacterial composition of kefir and an associated kefir grain. FEMS Microbiol Lett 320: 56–62.
31. Sakamoto N, Tanaka S, Sonomoto K, Nakayama J (2011) 16S rRNA pyrosequencing-based investigation of the bacterial community in nukadoko, a pickling bed of fermented rice bran. Int J Food Microbiol 144: 352–359.
32. Simon C, Daniel R (2009) Achievements and new knowledge unraveled by metagenomic approaches. Appl Microbiol Biotechnol 85: 265–276.
33. Schluter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, et al. (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J Biotechnol 136: 77–90.
34. Jung JY, Lee SH, Kim JM, Park MS, Bae J-W, et al. (2011) Metagenomic analysis of kimchi, the Korean traditional fermented food. Appl Environ Microbiol 77: 2264–2274.
35. Hunt J, Boddy L, Randerson PF, Rogers HJ (2004) An evaluation of 18S rDNA approaches for the study of fungal diversity in grassland soils. Microbial Ecology 47: 385–395.
36. Smit E, Leeflang P, Glandorf B, van Elsas JD, Wernars K (1999) Analysis of fungal diversity in the wheat rhizosphere by sequencing of cloned PCR-amplified genes encoding 18S rRNA and temperature gradient gel electrophoresis. Appl Environ Microbiol 65: 2614–2621.
37. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4: 63–72.
38. Diaz N, Krause L, Goesmann A, Niehaus K, Nattkemper T (2009) TACOA – Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10: 56.
39. Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods 6: 673–676.
40. Monzoorul HM, Ghosh TS, Singh NK, Mande SS (2010) SPHINX – An algorithm for taxonomic binning of metagenomic sequences. Bioinformatics 27: 22–30.
41. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, et al. (2008) Phylogenetic classification of short environmental DNA fragments. NAR 36: 2230–2239.
42. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377–386.
43. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52: 540–542.
44. Kelley D, Salzberg S (2010) Clustering metagenomic sequences with interpolated Markov models. BMC Bioinformatics 11: 544.
45. 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 Microbiol 28: 457–464.
46. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, et al. (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 25: 2607–2608.
47. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.
48. Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P (2011) SmashCommunity: a metagenomic annotation and analysis tool. Bioinformatics 26: 2977–2978.
49. Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, et al. (2010) An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. Fungal Ecol 3: 284–287.
50. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 12: S4.
51. Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. NAR 39: e91.
52. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2008) The Pfam protein families database. NAR 38: D211–D222.
53. Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K (2011) RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics 12: 41.
54. Jaenicke S, Ander C, Bekel T, Bisdorf R, Droge M, et al. (2011) Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. PLoS ONE 6: e14519.
55. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72: 557–578.
56. Huson D, Richter D, Mitra S, Auch A, Schuster S (2009) Methods for comparative metagenomics. BMC Bioinformatics 10: S12.
57. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV (2009) New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18: 11–19.
58. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334–338.
59. Villion M, Moineau S (2009) Bacteriophages of *Lactobacillus*. Front Biosci 14: 1661–1683.
60. Shapiro OH, Kushmaro A, Brenner A (2009) Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. ISME J 4: 327–336.
61. Park E-J, Kim K-H, Abell GCJ, Kim M-S, Roh SW, et al. (2011) Metagenomic analysis of the viral communities in fermented foods. Appl Environ Microbiol 77: 1284–1291.
62. Garneau J, Moineau S (2011) Bacteriophages of lactic acid bacteria and their impact on milk fermentations. Microb Cell Fact 10: S20.
63. Émond E, Moineau S (2007) Bacteriophages in food fermentations. In: McGrath S, Van Sinderen D, eds. Bacteriophage: genetics and molecular biology. Norfolk, UK: Caister Academic Press. pp 93–124.
64. 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 Res 11: 564–574.
65. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, et al. (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A Biol Sci 104: 8397–8402.
66. Kobayashi M, Hayashi S (1998) Supplementation of NaCl to starter culture of the soy yeast *Zygosaccharomyces rouxii*. J Ferment Bioeng 85: 642–644.
67. Aime MC, Phillips-Mora W (2005) The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of *Marasmiaceae*. Mycologia 97: 1012–1022.
68. Cardenas E, Tiedje JM (2008) New tools for discovering and characterizing microbial diversity. Curr Opin Biotechnol 19: 544–549.
69. Lefeber T, Janssens M, Moens F, Gobert W, De Vuyst L (2011) Interesting starter culture strains for controlled cocoa bean fermentation revealed by simulated cocoa pulp fermentations of cocoa-specific lactic acid bacteria. Appl Environ Microbiol 77: 6694–6698.
70. Carr JC, Davies AP, Dougan J (1979) Cocoa fermentation in Ghana and Malaysia. 7th International Cocoa Research Conference. Douala, Cameroon. pp 573–576.
71. Nielsen DS, Schillinger U, Franz CMAP, Bresciani J, Amoa-Awua W, et al. (2007) *Lactobacillus ghanensis* sp. nov., a motile lactic acid bacterium isolated from Ghanaian cocoa fermentations. Int J Syst Evol Microbiol 57: 1468–1472.
72. Passos FML, Silva DO, Lopez A, Ferreira CLLF, Guimarães WV (1984) Characterization and distribution of lactic acid bacteria from traditional cocoa bean fermentations in Bahia. J Food Sci 49: 205–208.
73. Bernardeau M, Guguen M, Vernoux JP (2006) Beneficial lactobacilli in food and feed: long-term use, biodiversity and proposals for specific and realistic safety assessments. FEMS Microbiol Rev 30: 487–513.
74. Hauben L, Moore ERB, Vauterin L, Steenackers M, Mergaert J, et al. (1998) Phylogenetic position of phytopathogens within the *Enterobacteriaceae*. Syst Appl Microbiol 21: 384–397.
75. De Vuyst L, Takrama JF, Ampomah YA, Lefeber T, Camu N (2010) Influence of a lactic acid bacterium/acetic acid bacterium starter culture on cocoa bean heap fermentation dynamics and chocolate flavor. Proceedings of the 16th International Cocoa Research Conference, Nusa Dua, Bali, Indonesia, November 16–20, pp 1325–1332.
76. 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 Microbiol 30: 379–392.
77. Schwann RF (1998) Cocoa fermentations conducted with a defined microbial cocktail inoculum. Appl Environ Microbiol 64: 1477–1483.