BMC Genomics

RESEARCH ARTICLE Open Access

DAIRYdb: a manually curated reference database for improved taxonomy annotation of 16S rRNA gene sequences from dairy products

Marco Meola[1*†], Etienne Rifa[2†], Noam Shani[1], Céline Delbès[2], Hélène Berthoud[1] and Christophe Chassard[2]

Abstract

Background: Reads assignment to taxonomic units is a key step in microbiome analysis pipelines. To date, accurate taxonomy annotation of 16S reads, particularly at species rank, is still challenging due to the short size of read sequences and differently curated classification databases. The close phylogenetic relationship between species encountered in dairy products, however, makes it crucial to annotate species accurately to achieve sufficient phylogenetic resolution for further downstream ecological studies or for food diagnostics. Curated databases dedicated to the environment of interest are expected to improve the accuracy and resolution of taxonomy annotation.

Results: We provide a manually curated database composed of 10'290 full-length 16S rRNA gene sequences from prokaryotes tailored for dairy products analysis (https://github.com/marcomeola/DAIRYdb).
The performance of the DAIRYdb was compared with the universal databases Silva, LTP, RDP and Greengenes. The DAIRYdb significantly outperformed all other databases independently of the classification algorithm by enabling higher accurate taxonomy annotation down to the species rank. The DAIRYdb accurately annotates over 90% of the sequences of either single or paired hypervariable regions automatically.
The manually curated DAIRYdb strongly improves taxonomic annotation accuracy for microbiome studies in dairy environments. The DAIRYdb is a practical solution that enables automatization of this key step, thus facilitating the routine application of NGS microbiome analyses for microbial ecology studies and diagnostics in dairy products.

Keywords: Microbiome, Taxonomy annotation, OTU classification, 16S, Database, Accuracy, Dairy, Cheese, Milk, Whey, Teat, Starter

Background

The exploration of microbial communities has experienced a boost during the last decade with the advent of next-generation sequencing (NGS) technologies [1]. Previously undetectable micro-organisms in soils [2], water [3, 4], airborne [5, 6], snow [7], ice [8], food [9], human gut [10–12] etc. could be unravelled at an unprecedented depth and resolution. Numerous studies have been published describing microbial community structures in various environments, often correlating their dynamic changes over time or space by means of the 16S rRNA gene (16S) [13–15].

First microbiome studies using the 16S were based on fingerprinting techniques, such as Denaturing Gradient Gel Electrophoresis (DGGE), Terminal Restriction Fragment Length Polymorphism (T-RFLP) or Length Heterogeneity Polymerase Chain Reaction (LH-PCR) sometimes

*Correspondence: marco.meola@agroscope.admin.ch
†Marco Meola and Etienne Rifa contributed equally to this work.
[1] Agroscope, Competence Division Methods Development and Analytics, Research Group Fermenting Organisms, Schwarzenburgstrasse 161, 3003 Bern, Switzerland
Full list of author information is available at the end of the article

in combination with Sanger sequencing of the 16S to identify populations of interest. While Sanger sequencing delivered almost the complete 16S at good quality, the throughput was low due to the high workload, preventing researchers to unravel the full array of microbial diversity within a sample [16].

NGS has increased the sequencing depth, uncovering also low abundant micro-organisms, thus overcoming the limitations of Sanger sequencing. The higher sequencing depth of NGS, however, was obtained at the expense of read length, therefore limiting the sequencing to only few hyper variable regions (HVR). The short size of the resulting reads strongly reduces their resolution, while increasing the risk of taxonomic miss-annotation or ambiguous taxonomic classification. The need for trustworthy classification of very short 16S sequences covering only one to three HVR remains a crucial step to obtain robust and accurate taxonomic classification in modern microbiology [17].

Microbiome studies in dairy products are particularly affected by limitation in taxonomic annotation. Dairy environments are highly selective and thus often characterized by only few abundant genera belonging to the lactic acid bacteria (LAB). Therefore, complete microbial biodiversity in dairy products is only visible at species or even strain level, which makes taxonomic annotation at species level crucial for any diagnostics or microbial ecology study. Moreover, short fragment strategies, on single HVRs or HVR pairs, often fail to reliably assign the correct taxonomy at the species level. It is therefore of paramount importance to select the right HVRs to maximize taxonomy resolution in a given environment, especially when differences are visible at species level only.

In recent years, numerous classification algorithms have been developed and optimized to accurately annotate operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) from short reads. Those classification tools have been developed for 16S and other genes based on different mathematical models, such as e.g., k-mer, Bayesian, Hidden Markov-Monte-Carlo model (HMM) etc. The Basic Local Alignment Search Tool (Blast) has long been the gold standard for sequence comparison and annotation [18]. More 16S specific taxonomy predictors have been developed, including RDP Naive Bayesian Classifier (NBC) [19], a naive Bayesian Classifier based on k-mers, GAST [20], MEGAN [21], Metaxa2 [22], riboFrama [23], SPINGO [24], PROTAX [25], SINTAX [26], DynamiC [27], Humidor [28], MAPseq [29], microclass [30], q2-feature-classifier [31], IDTAXA [32] and other tools implemented in the most current 16S pipelines like mothur [33], Qiime v1 [34], Qiime v2 [35] and FROGS [36].

Although classification prediction algorithms have strongly improved, manually curated databases containing only authoritative full-length 16S sequences from type strains and cultivated reference strains can potentially compensate the limitations of short read sequences annotations by means of sophisticated algorithms. To date, three main independent universal repositories dedicated to universal 16S sequences from prokaryotes are widely used: Silva, The Ribosomal Database Project (RDP), and Greengenes (GG) [37].

Silva is the universal 16S repository with the highest number of sequences. The latest release of Silva SSU/LSU 132 (www.arb-silva.de) contained 6’073’181 16S sequences of at least 300 bp, with 2’090’668 good quality sequences with at least 900 bp length [38–40]. Taxonomic rank information of Silva and Living Tree Project (LTP) are based on the Bergey’s Taxonomic Outlines and the List of Prokaryotic Names with Standing Nomenclature (LPSN) [41]. Minimal training sets, such as the SSU Ref NR 99 or the LTP [42], offer a reduced number of sequences for faster classification but still covering the broadest currently known biodiversity.

The second biggest repository, the Ribosomal Database Project (RDP Release 11, Update 5; http://rdp.cme.msu.edu) [43], contained at the time of writing 3’356’809 16S sequences from the International Nucleotide Sequence Database Collaboration (INSDC) [44]. The nomenclature is based on the Bacterial Nomenclature Up-to-Date and the taxonomic rank information on the Bergey’s Manual.

Greengenes v13_5 [45] contains 1’800’000 quality filtered 16S sequences. Classification nomenclature is based on automatic de novo tree construction and rank mapping with the NCBI Taxonomy database [46]. Although frequently used in community studies together with Qiime [34], the last update dates back to 2013 with no indication for an imminent update.

Taxonomic classification of the 16S is not trivial and requires both familiarity with prokaryotic phylogeny and often manual intervention due to poor annotation of the OTUs or by the available 16S databases [47]. Fast and accurate, thus automatized classification of the OTUs is not yet possible at the biologically most significant species rank due to the short sequence fragments and the absence of food-dedicated, thoroughly curated 16S databases. Manually curated databases are of paramount importance to improve reproducibility, speed during the bioinformatics process of microbiome studies, and communication between researchers [48]. Previous studies have highlighted the importance of high-quality data for improving the classification of the obtained OTUs [17, 49, 50]. Although universal 16S databases cover vast prokaryotic biodiversity, they often fail to guarantee accurate classification to the species rank for sequences obtained from a highly studied environment, such as dairy products. In fact, annotation accuracy at lower taxonomic ranks increases with a standard training set encompassing only full-length and good quality representative sequences innate to the investigated environment [17, 48, 50–52].

Here we present a comprehensive reference database, DAIRYdb (Database, Agroscope, Inra, Ribosomal, accuracY), for 16S OTUs classification of next generation sequencing (NGS) reads from dairy products. The main goal was to develop a dedicated database that allows researchers to accurately and automatically annotate short reads of 16S down to the species level. The performance of the DAIRYdb was compared with the universal databases Silva, RDP, Greengenes and LTP using three predictors based on different algorithms and programming languages [53], such as Blast+ [54], Metaxa2, [22, 55], and SINTAX [26]. Manual curation of the database and its restriction to the biodiversity expected in dairy products strongly improves accuracy and reproducibility of phylogenetic classification at all taxonomic ranks.

DAIRYdb is publicly available at https://github.com/marcomeola/DAIRYdb and can be integrated in any classification prediction tool that allows adaptation of customized databases, such as Blast+, Metaxa2, SINTAX, IDTAXA and FROGS.

Results and discussion

Construction

The 16S sequence database of dairy products DAIRYdb was constructed using a set of over 390’000 sequences associated to the selected keywords (cheese, milk, teat, dairy, starter, whey) deposited in NCBI GenBank and ENA/EMBL, as well as sequences with 97% average nucleotide identity (ANI) from Silva, RDP and Greengenes (Fig. 1). About 10’000 best quality reference sequences were retained after filtering based on sequence length (>1300 bp), quality (pintail >75) and potential chimeras. Finally, 16S sequences of important species from cheese and dairy environments [56, 57] lost during the clustering were added subsequently. The final number of 16S sequences consequently reached 10’290.

The observed distribution among the different key words might reflect the unequal distribution of microbiome studies predominantly performed on cheese, dairy and milk samples, as compared to teats and whey. About 1933 sequences of the DAIRYdb were shared among all

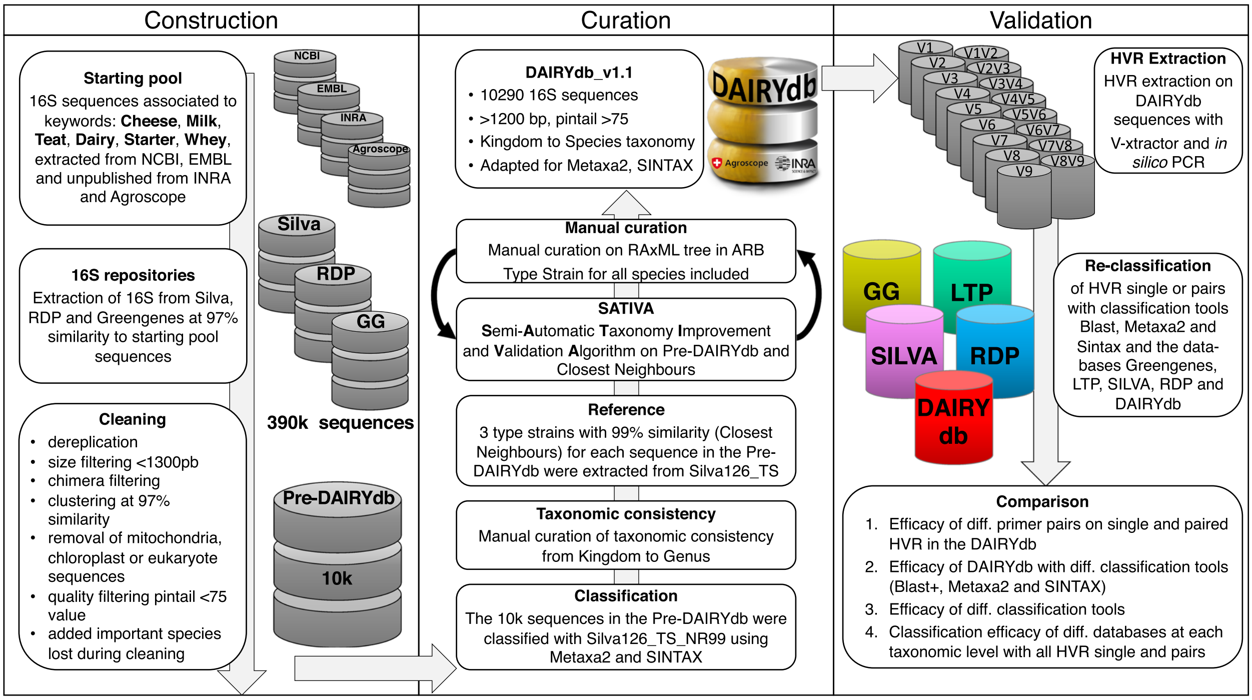
Construction
Curation
Validation
Starting pool
16S sequences associated to keywords: Cheese, Milk, Teat, Dairy, Starter, Whey, extracted from NCBI, EMBL and unpublished from INRA and Agroscope
NCBI
EMBL
INRA
Agroscope
16S repositories
Extraction of 16S from Silva, RDP and Greengenes at 97% similarity to starting pool sequences
Silva
RDP
GG
390k sequences
Cleaning
• dereplication
• size filtering <1300pb
• chimera filtering
• clustering at 97% similarity
• removal of mitochondria, chloroplast or eukaryote sequences
• quality filtering pintail <75 value
• added important species lost during cleaning
Pre-DAIRYdb
10k
DAIRYdb_v1.1
• 10290 16S sequences
• >1200 bp, pintail >75
• Kingdom to Species taxonomy
• Adapted for Metaxa2, SINTAX
DAIRYdb
Agroscope
INRA
Manual curation
Manual curation on RAxML tree in ARB
Type Strain for all species included
SATIVA
Semi-Automatic Taxonomy Improvement and Validation Algorithm on Pre-DAIRYdb and Closest Neighbours
Reference
3 type strains with 99% similarity (Closest Neighbours) for each sequence in the Pre-DAIRYdb were extracted from Silva126_TS
Taxonomic consistency
Manual curation of taxonomic consistency from Kingdom to Genus
Classification
The 10k sequences in the Pre-DAIRYdb were classified with Silva126_TS_NR99 using Metaxa2 and SINTAX
V1
V1V2
V2
V2V3
V3
V3V4
V4
V4V5
V5
V5V6
V6
V6V7
V7
V7V8
V8
V8V9
V9
HVR Extraction
HVR extraction on DAIRYdb sequences with V-xtractor and *in silico* PCR
GG
LTP
SILVA
RDP
DAIRY db
Re-classification
of HVR single or pairs with classification tools Blast, Metaxa2 and Sintax and the databases Greengenes, LTP, SILVA, RDP and DAIRYdb
Comparison
1. Efficacy of diff. primer pairs on single and paired HVR in the DAIRYdb
2. Efficacy of DAIRYdb with diff. classification tools (Blast+, Metaxa2 and SINTAX)
3. Efficacy of diff. classification tools
4. Classification efficacy of diff. databases at each taxonomic level with all HVR single and pairs

Fig. 1 Development of the DAIRYdb consisted in three main steps: construction, curation and validation. For construction, dairy products specific 16S sequences were retrieved from Silva, RDP and Greengenes using Genbank NCBI, EMBL, Agroscope and INRA sequences. Curation was performed based on the cross-validation results from the leave-one-out test of SATIVA and highly iterated RAxML tree, followed by manual curation of taxonomic assignment and consistency throughout all taxonomic ranks, with a particular focus on singleton taxons with no reference sequence. Validation was performed comparing taxonomy annotation accuracy of single and HVR pairs by the five databases (Greengenes 13.8, LTP version, Silva 128 NR99, RDP version and DAIRYdb)

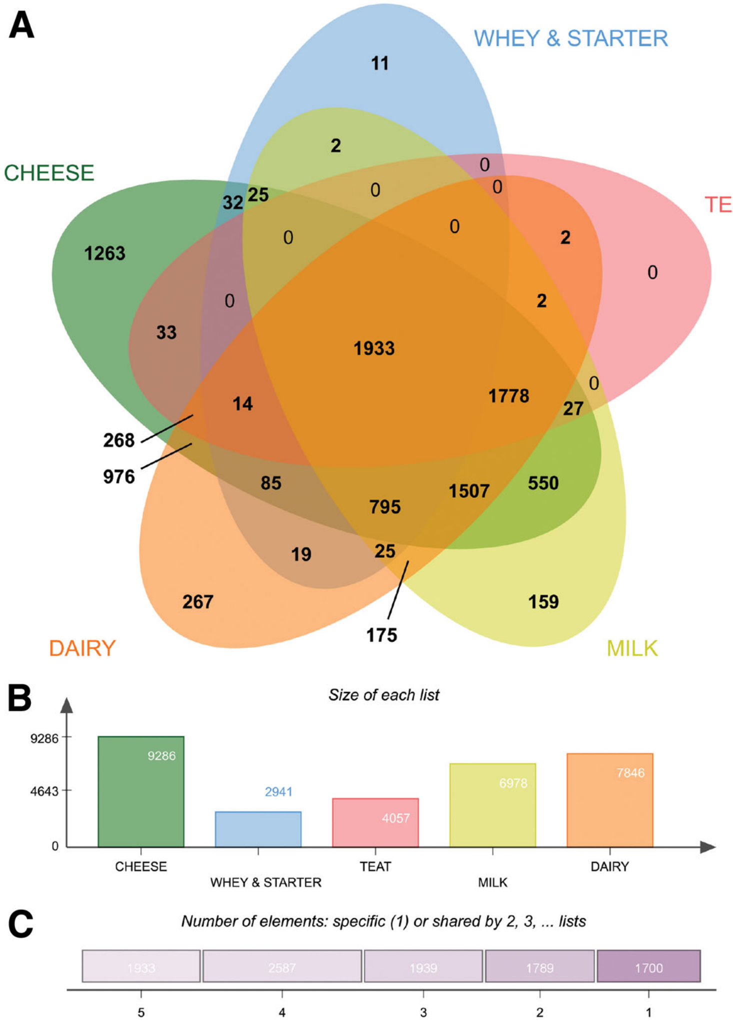

Fig. 2 Origin of sequences in the DAIRYdb. **a** Five-factors Venn diagram comparing the origins of the sequences (9'948) retrieved from the public repositories Genbank NCBI and EMBL associated to the keywords "cheese", "dairy", "milk", "teat" and "whey/starter". About 12.7% (1'263) sequences were only detected in cheese and 15.1% (1'507) were detected in all three cheese, milk and dairy environments. **b** Total number of sequences associated to a particular keyword. **c** Number of sequences shared by 1 to 5 keywords. About 19.4% (1'933) sequences were detected in all 5 keywords, while 17.1% (1'700) sequences were unique to one keyword

keywords (Fig. 2a) and 1778 were shared among the keywords dairy, cheese and milk. In fact, the majority of the sequences composing the DAIRYdb were linked to those three keywords (Fig. 2b). Altogether, 1'700 sequences were associated to just one keyword, with most of the sequences shared by four keywords (Fig. 2c).

Curation

During the first step of data curation, the sequences were taxonomically annotated with Silva by means of SINA [58]. The resulted annotation at all taxonomic ranks underwent a first manual check and cleaning for taxonomic inconsistencies through cross-comparison with the other members of the same taxonomic rank in a phylogenetic tree. No taxonomic overlaps comparable to other databases are present in the DAIRYdb, where different species of the same genus fall under different taxonomic lineages [48]. A maximum of three closest neighbour type strains (CN) with authoritative taxonomy from Silva sharing 99% global sequence similarity to each sequence in the DAIRYdb were added to the 10'290 sequences in the DAIRYdb as reference during the curation process and removed at the end of the curation process. The maximal number of lowest common ancestors (LCA) with an authoritative taxonomy strongly improved the curation process with the Semi-Automatic Taxonomy Improvement and Validation Algorithm (SATIVA) increasing robustness of the proposed changes of miss-annotated environmental sequences within the DAIRYdb [59].

By using only near full-length and curated 16S from type strains as reference sequences, we were able to validate and correct the taxonomy annotation where necessary. The SATIVA results were inspected and taxonomy manually curated using a highly iterated phylogenetic tree. The approach used during the manual curation broadly follows the rationale described in detail in a recently published study [48]. Taxonomy annotations from authoritative type strain sequences were used as reference for the environmental sequences in the tree. For ranks at which no taxonomic annotation was possible with certainty due to the lack of authoritative type strains within the same clade (*i.e.*, commonly labelled "unknown", "uncultured" etc. in universal databases), the *lowest common rank* (LCR) [52] was used down to the species rank with the addition of the unclassified rank. As an example, a sequence assigned to the LCR, the genus *Sporichthya*, was named at species rank *Sporichthya_Species*. This approach avoids the merging of abundance values from different unknown species to biologically uninformative groups, thus improving communication among scientists [60].

DAIRYdb version 1.1 contains 2 kingdoms (Bacteria and Archaea), 47 phyla, 136 classes, 249 orders, 463 families, 1'757 genera and 4'030 unique species-like groups/species complexes (Fig. 3, Additional files 1 and 2). The *Firmicutes* is the predominant phylum with 37% of all sequences, followed by the *Proteobacteria* (22%), *Bacteroidetes* (14%), *Actinobacteria* (9%), *Chloroflexi* (2%), *Acidobacteria* (2%), Archaea (1%) and 34 other minor phyla. The 1% of Archaea is subdivided into *Euryarchaeota* (74%), *Crenarchaetoa* (13%), *Thaumarchaeota* (9%), *Woesearchaeota* (3%) and others (1%). Altogether, the DAIRYdb was able to capture the diversity of known taxa expected to occur in dairy products. Increasing number of whole genome sequences (WGS) will most likely lead to a replacement of

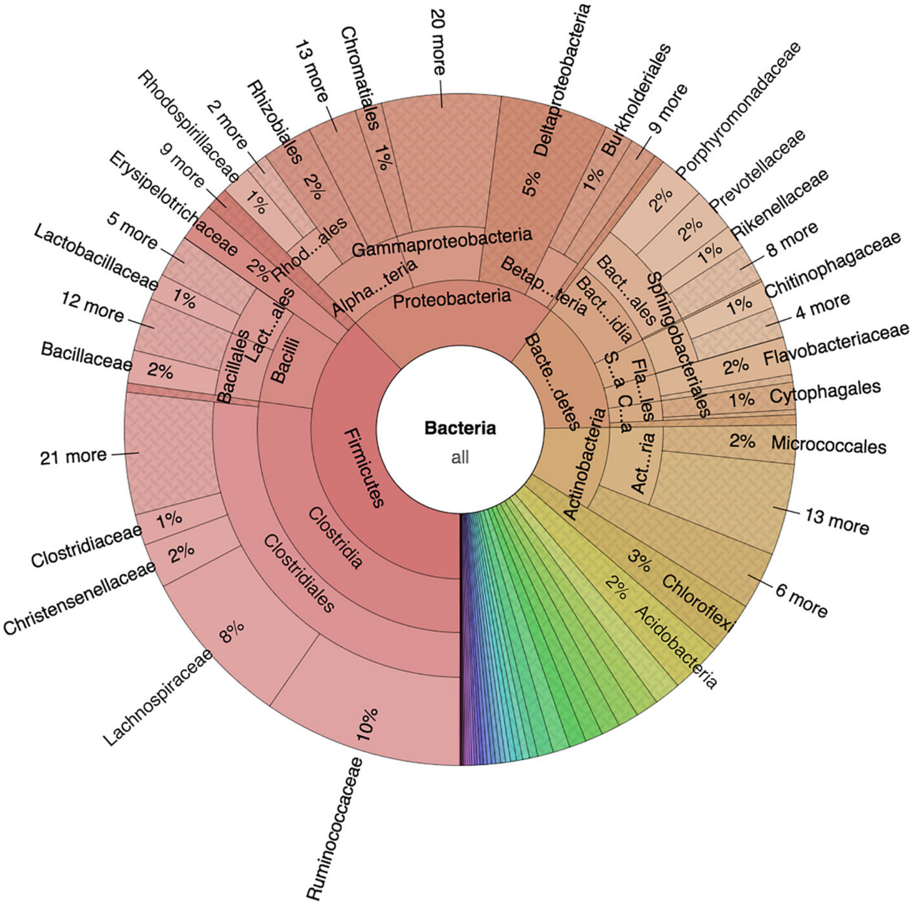

Fig. 3 Complete microbial diversity present in the DAIRYdb. Prokaryotic biodiversity in the DAIRYdb is represented by 2 kingdoms, 47 phyla, 136 classes, 249 orders, 463 families, 1'757 genera and 4'030 unique species-like groups. The most represented phylum is Firmicutes (37% of all sequences), followed by the Proteobacteria (22%), Bacteroidetes (14%), Actinobacteria (9%), Chloroflexi (2%), Acidobacteria (2%), Archaea (1%) and 34 other minor phyla

incomplete 16S sequences in the DAIRYdb by full-length sequences that cover all HVRs.

The cheese microbiome is often dominated by few phylogenetically closely related species of LAB belonging to a few genera (*e.g.*, *Lactobacillus*, *Lactococcus*, *Leuconostoc* and *Streptococcus*) [9]. Therefore, special attention was put into the manual curation of the DAIRYdb sequences at species rank. Despite the genotypic and phenotypic characteristics of the most common LAB in cheese are extensively studied and described, several controversies regarding the nomenclature of some keystone species still remain unsolved, such as for the species *Lactococcus lactis* subsp. *lactis* and *Lactococcus lactis* subsp. *cremoris* [61]. It still is unresolved whether *Streptococcus thermophilus* is a species on its own or a subspecies of *Streptococcus salivarius* [62–64]. The DAIRYdb is composed of sequences retrieved from the Silva database along with their respective taxonomy, which was manually inspected for nomenclature hierarchy conflicts based on the phylogenetic position within the tree. However, some conflicting annotations of the same sequence were detected between the Silva taxonomy and the Bacterial Diversity Metadatabase, such as the species assignment of the type strain sequence Accession AB008205, which is labelled as *L. casei* in Silva and *L. paracasei* in Bacterial Diversity Metadatabase (BacDive) [65]. For the reference sequences of the most crucial species, bacterial names listed in the actual “List of prokaryotic names” according to BacDive were used. However, further disagreements between Silva and BacDive cannot be completely excluded. Moreover, some crucial genera in dairy products may undergo a radical genome-based relabelling in the future to create more homogeneous clusters [64].

Different approaches were applied on impure taxa, *i.e.* taxa that overlap in the tree despite being assigned to different nomenclature [48] in the universal databases. For instance, for the genera *Escherichia* and *Shigella*, Silva, LTP and RDP use the combined genus name *Escherichia–Shigella* but retain well-established species names, such as *Escherichia coli*. Differently, Greengenes leaves their sequences unclassified at ranks below the family *Enterobacteriaceae* [48]. The different taxonomic nomenclature references used by the three databases have an impact on revisions to resolve conflicts with sequence-based phylogenies and the labelling of new candidate groups identified in environmental sequences. However, discussion on the taxonomic inconsistencies and limitations of the universal databases (Silva, LTP, RDP and Greengenes), which the DAIRYdb was compared with, goes beyond the scope of this study and was extensively discussed elsewhere [48, 52].

The DAIRYdb will undergo regular updates in accordance to update on bacterial nomenclature [64], integrating the novelties or correcting the changes. Finally, the inclusion of full-length and high-quality 16S sequences from reference type strains leads to a more robust and confident taxonomic classification [49].

Validation

At present, only short read sequences can be obtained from the most common amplicon NGS sequencers with at least 99% read quality and up to 600 bp in read length (Illumina MiSeq, Ion Torrent S5). Although long reads sequencing technology, such as PacBio and Oxford Nanopore, are steadily improving read quality, they are not yet routinely used for amplicon metabarcoding studies. Therefore, performance of the DAIRYdb was evaluated on short read sequences spanning over either a single HVR or HVR pairs.

The single HVRs and HVR pairs were extracted from randomly subsampled sequences of the DAIRYdb using two methods: V-Xtractor [66] or in silico PCR with mothur [33] (Fig. 4a, c). While V-Xtractor extracted all present HVRs, the in silico PCR also evaluated the theoretical extraction efficiency of the different primer pairs (Table 1). Almost 100% of the sequences in the DAIRYdb span from V2 to V8. The HVR V1 (89%) and V9 (68%) are the regions with the lowest coverage in the DAIRYdb. This is due to the commonly used universal primers 8F and 1492R for the full-length 16S PCR leading to the entire or partial loss of V1 and V9 (Fig. 4a, Table 1). The primer pairs targeting V1 (25%), V2 (58%) and V9 (55%) were less efficient as compared to the primer pairs targeting V3 (84%) to V8 (77%). The primer pair for V4 performed best with 92% coverage, followed by V7 (88%), V5 and V6 (86%). The same hold true for the HVR pairs, where the HVR pairs V1V2 (38%), V2V3 (55%) and V8V9 (51%) performed worse as compared to the central HVR pairs (79–89%) (Fig. 4c).

The ratio between the number of detected HVRs with V-Xtractor and HVRs extracted by in silico PCR determined the biodiversity coverage of the different HVRs achieved with the different primer pairs, which can bias further downstream analyses depending on the HVR targeted (Fig. 4b, d). The net in silico performance of each primer pair is presented as normalized to the total number of sequences detected by V-Xtractor for each HVR (Fig. 4b, d).

The results of microbiome studies are more strongly influenced by the selection of the primer pairs and thereof of the HVR amplified, than by the sequencing technology used for the study [78–80]. The OTU-picking algorithm is mainly dependent on the sequencing technology (clustering vs. denoising) or ASVs instead [81]. Therefore, classification predictors are of secondary importance, although their impact on the outcome is not negligible [82]. The selection of the primer pairs should be made after careful

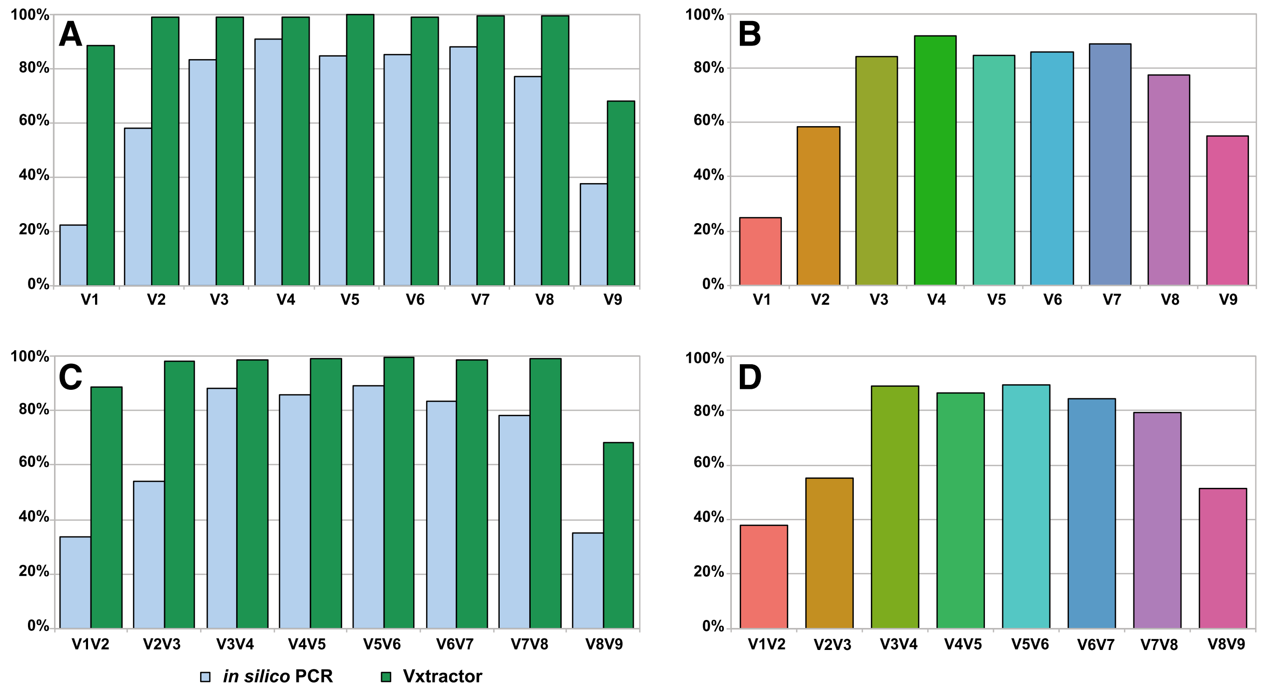

Fig. 4 Presence and extraction efficiency of all HVR in the sequences of the DAIRYdb. Single HVR (**a**) and (**b**) and HVR pairs (**c**) and (**d**) HVRs were extracted using in silio PCR with mothur (**b**) and (**d**) and HVR extraction with V-Xtractor (**a**) and (**b**) from sequences present in DAIRYdb v1.1 to test completeness of the sequences therein. While almost 100% of the 10'290 sequences span over V2 to V8 only 89% contain V1 and 68% contain V9 (**a**) and (**c**). The in silico PCR highlights the theoretical amplification efficiency of the most common universal primers with 0 mismatches normalized to the total number of detected HVR (**b**) and (**d**)

Table 1 Primers used in the in silico PCR extraction of the HVRs

Label	Name - ARB primers	HVR	Location*	bp	Primer sequence	GC%	Reference	Original reference primer
8F_v1f	S-D-Bact-0008-d-S-20	v1F	8-27	20	AGAGTTTGATCMTGGCTCAG	50	[67]	
120R_v1r	*S-D-Bact-0120-e-A-20*	v1R	101-120	20	TTACTCACCCGTNCGCCRCT	55	mod. rev-compl. after [68]	
101F_v2f	S-D-Bact-0101-a-S-20	v2F	101-120	20	AGYGGCGNACGGGTGAGTAA	55	mod. after [68]	
355R_v2r	*S-D-Bact-0355-a-A-18*	v2R	338-355	18	GCWGCCTCCCGTAGGAGT	66	mod. after [69]	
338F_v3f	*S-D-Bact-0338-a-S-20*	v3F	337-354	20	ACWCCTACGGGWGGCAGCAG	65	mod. after [70]	
534R_v3r	S-D-Bact-0518-b-A-17	v3R	518-534	17	ATTACCGCGGCTGCTGG	65	[71]	
515F_v4f	_S-*-Univ-0515-b-S-19_	v4F	515-533	19	GTGNCAGCMGCCGCGGTAA	63	mod. after [72]	
806R_v4r	*S-D-Bact-0756-a-A-20*	v4R	787-806	20	GGACTACHVGGGTWTCTAAT	40	mod. after [72]	
784F_v5f	_S-*-Univ-0779-a-S-15_	v5F	784-798	15	RGGATTAGATACCCY	40	mod. after [73]	
926R_v5r	*S-D-Bact-0907-b-A-20*	v5R	907-926	20	CCGTCAATTYYTTTRAGTTT	25	mod. after [74]	
907F_v6f	S-D-Bact-0907-a-S-20	v6F	907-926	20	AAACTYAAARRAATTGACGG	25	[75]	
1114R_v6r	S-D-Bact-1114-b-A-16	v6R	1099-1114	16	GGGTYKCGCTCGTTRY	50	mod. after [73]	S-D-Bact-1114-a-A-16
1099F_v7f	_S-*-Univ-1099-a-S-16_	v7F	1099-1114	16	RYAACGAGCGMRACCC	50	new primer	S-*-Univ-1100-a-S-15
1200R_v7r	*S-D-Bact-1200-a-A-16*	v7R	1185-1200	16	GAYTTGACRTCVTCCM	38	new primer	
1185F_v8f	*S-D-Bact-1185-a-S-16*	v8F	1185-1200	16	KGGABCACCGCYCGYC	63	new primer	
1407R_v8r	*S-D-Bact-1407-a-A-16*	v8R	1391-1407	16	GRCGRGCGGTGWGTRC	63	mod. after [76]	S-D-Bact-1391-a-A-17
1391F_v9f	*S-D-Bact-1391-a-S-16*	v9F	1391-1407	16	GYACWCACCGCYCGYC	63	new primer	
1510R_v9r	_S-*-Univ-1510-b-A-19_	v9R	1492-1510	19	GGNTACCTTGTTACGACTT	42	mod. after [77]	S-*-Univ-1492-a-A-21

**E. coli* position as a reference

consideration of their coverage in diversity with respect to the studied environment [82]. Although researchers tend to use primers as universal as possible to catch the entire diversity present in the samples, it might be a pragmatic approach to lose some universality while increasing specificity for the studied environment. For dairy products, the DAIRYdb achieves both, covering all the biodiversity expected in these environments, while achieving specificity in taxonomic annotation.

Annotation accuracy of the DAIRYdb was compared with other universal databases, such as Silva128, RDP trainset v16, LTP and Greengenes analysing fragments of single HVRs or HVR pairs extracted from the sequences in the DAIRYdb with V-Xtractor and in silico PCR. About 1000 subsamples, each composed of synthetic HVRs extracted from 100 randomly selected sequences from the DAIRYdb, by either V-Xtractor or in silico PCR, were assigned to all taxonomic ranks by the means of three different classification predictors (Blast+, Metaxa2 and SINTAX) and the aforementioned databases.

Single HVRs extracted with V-Xtractor and annotated with the DAIRYdb using SINTAX achieved annotation accuracy above 75% at all taxonomic ranks (Fig. 5a). Accuracy was highest for the even HVRs (V2, V4, V6 and V8) as compared to the odd HVRs (V1, V3, V5, V7 and V9). The region V2 presented the highest taxonomy annotation accuracy, which is in line with other findings showing that the regions V1 and V2 resulted in a more accurate OTU clustering at 97%, 98% and 99% [27]. Overall, the universal databases were less accurate with decreasing taxonomic rank (Fig. 5b-e). Only the RDP trainset v16 achieved about 25% of correct species annotations, while the other databases only classified to genus rank. Although the RDP trainset v16 performed best among all universal databases, annotation accuracy was below the accuracy values assessed in previous studies [48]. Universal databases achieved highest annotation accuracy with V4, with exception to Silva, which performed best on V2 (Fig. 5). Generally, the difference in annotation accuracy was stable through all HVRs with exception to the Silva database, where bigger oscillations were observed between the HVRs showing a clear drop for V6 and V7 (Fig. 5e). All HVRs taken together, the DAIRYdb achieved a significantly better taxonomy annotation accuracy (adjusted p-value <0.001) of average 88.9% ± 5.5 as compared to the universal databases tested at any taxonomic rank, but particularly at order to species ranks (Fig. 5f; Additional file 4: Tables S1 and S2). Results with Blast+ and Metaxa2 on single HVRs are available in Additional file 3.

The results with the HVR pairs was similar to the single HVRs (Fig. 6a). Annotation accuracy between HVR pairs was less variable between different HVR pairs and within the bootstrapping values of the same HVR pair as compared to the single HVRs, indication for a more robust classification with increasing number of HVRs. The HVR pair V1V2 achieved the highest annotation accuracy at species rank in the DAIRYdb, as well as with RDP and Silva

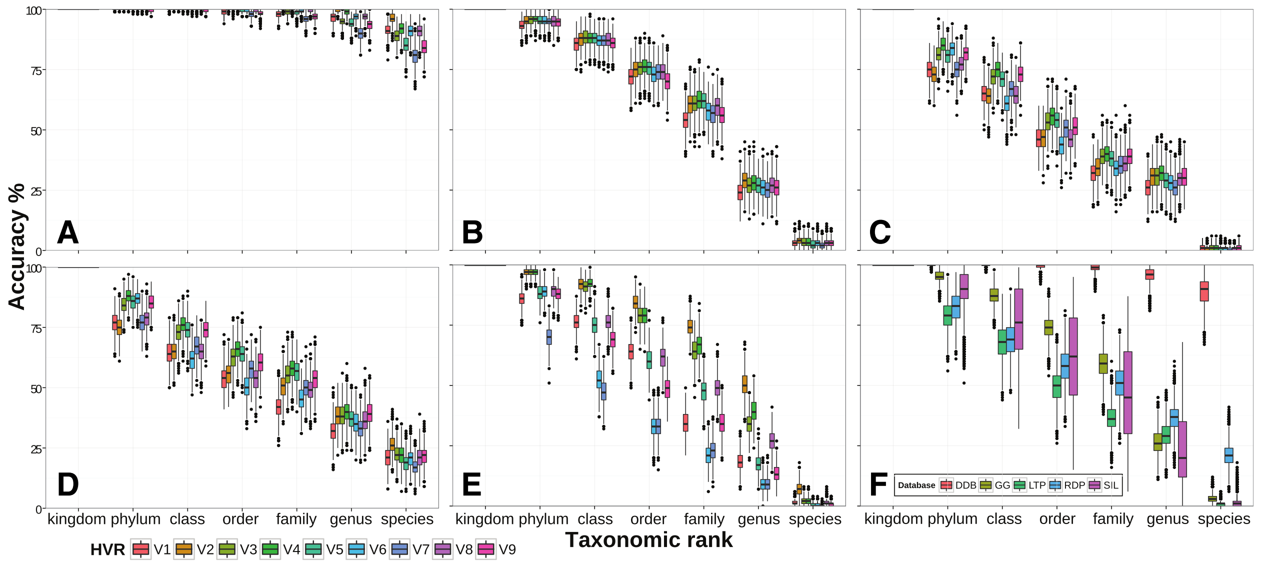

Fig. 5 Taxonomy annotation accuracy of the DAIRYdb on reads extracted with V-Xtractor. Single HVR V1-V9 were re-annotated using three different classification algorithms, Blast+, Metaxa2 or SINTAX, respectively. This figure shows the results with SINTAX (Analyses with Metaxa2 and Blast+ are shown in Additional file 2). Taxonomy annotation was bootstrapped 1000 times with a subset of 100 randomly selected sequences from the DAIRYdb and annotated with DAIRYdb (DDB) (**a**), Greengenes (**b**), LTP (**c**), RDP (**d**) and Silva (**e**). Average performance of all HVR for each database (**f**) (accuracy = correctly annotated/total)

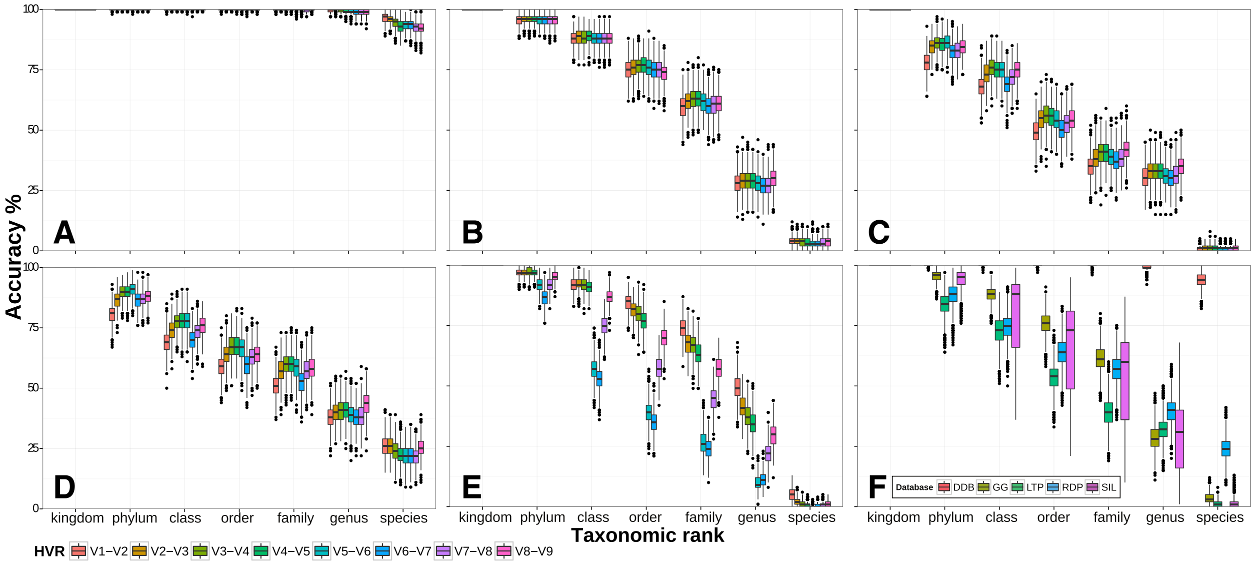

Fig. 6 Taxonomy annotation accuracy of the DAIRYdb on reads extracted with V-Xtractor. The HVR pairs V1-V2, V2-V3, V3-V4, V4-V5, V5-V6, V6-V7, V7-V8, V8-V9 were re-annotated using three different classification algorithms, Blast+, Metaxa2 or SINTAX, respectively. This figure shows the results with SINTAX (Analyses with Metaxa2 and Blast+ are shown in Additional file 2). Taxonomy annotation was bootstrapped 1000 times with a subset of 100 randomly selected sequences from the DAIRYdb and annotated with DAIRYdb (DDB) (**a**), Greengenes (**b**), LTP (**c**), RDP (**d**) and Silva (**e**). Average performance of all HVR for each database (**f**) (accuracy = correctly annotated/total)

(Fig. 6d-e). These results are in agreement with previous studies, where V1 and V2 have been shown to have the highest average classification accuracy and average confidence estimate up to the genus rank [19]. Greengenes species annotation accuracy was similar for all HVRs, while LTP showed very low performance at species rank (Fig. 6b-c). The average accuracy value for correct species annotation of all HVR pairs with the DAIRYdb was over 94% ± 2.8 (Fig. 6f). Only species annotation with the RDP trainset v16 achieved 25% of correct annotations. The BLAST16S database was shown to obtain genus accuracy values ∼50% for V4, which improved with increasing length to ∼60% with V3–V5 and ∼70% with full-length 16S [52]. As expected, the increasing number of HVR increases the confidence in taxonomy annotation. Results with Blast+ and Metaxa2 on HVR pairs are available in Additional file 3.

Different taxonomy predictors using the DAIRYdb showed similar performances and only little difference between single HVRs (Additional file 3: Figures S11) or HVR pairs (Additional file 3: Figure S12). Annotation accuracy performance varied more between different databases than between different classification predictors. The results confirm that OTUs annotation is primarily influenced by the selection of the database, by the HVR, and only at last by the taxonomy predictor (Additional file 3: Figures S1-S10). A comparison of the three classification predictors, Blast+, Metaxa2 and SINTAX with the DAIRYdb confirmed that HVR pairs could be more accurately assigned to the correct species than single HVR (Fig. 7). Among all tools, Blast+ and SINTAX were slightly better than Metaxa2. Since Metaxa2 uses more stringent parameters, as it only assigns the taxa if in agreement with Blast+, the lower performance of Metaxa2 with respect to Blast+ alone is therefore not surprising. Moreover, Metaxa2 performance is strongly dependent on the ANI thresholds used, which were set according to previous studies [83]. On the other hand, the more stringent parameters of Metaxa2 reduced the number of over-classified sequences. Generally, taxonomy annotation results are most robust whilst using different classification predictors with the DAIRYdb. We therefore recommend to use both, Metaxa2 with integrated Blast+ and SINTAX to obtain taxonomy annotations closest to the ground truth. Although a lower SINTAX cutoff of 0.6 increases the risk of over-classification, it is justified by the better quality of the DAIRYdb and the comparison with Metaxa2 for definitive taxonomy annotation (more details on the recommended usage on real samples are described on https://github.com/marcomeola/DAIRYdb).

The main scope of the DAIRYdb is to improve accurate species classification in dairy products. Beyond this, it covers a considerable diversity in agreement with the diversity detected in dairy products so far. However, the DAIRYdb does not necessarily perform better than universal databases on a set of sequences from environments other than dairy, such as the human gut. Annotation accuracy performed on sequences from type strains included

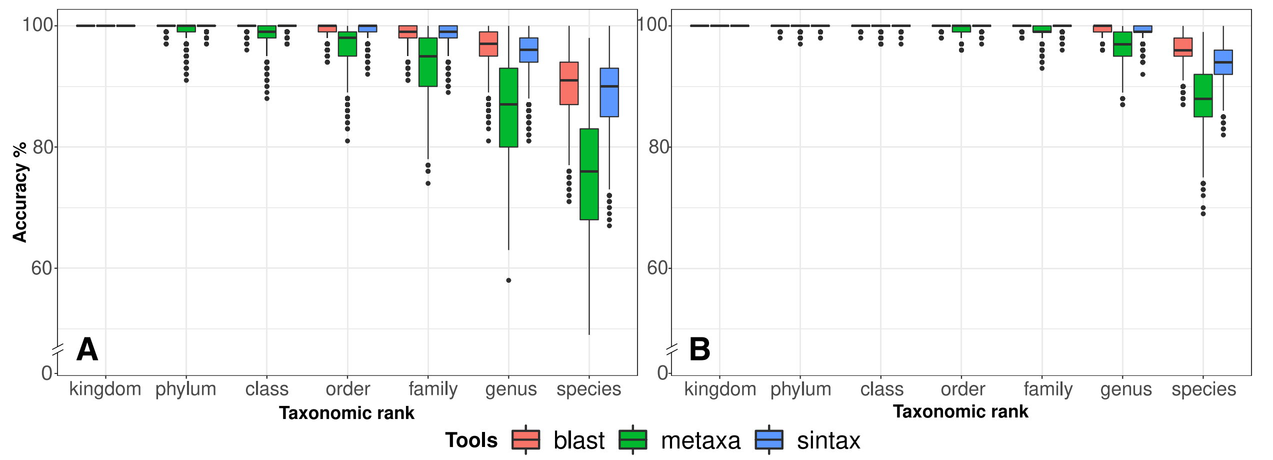

Fig. 7 Comparison of the overall annotation accuracy of the three algorithms, Blast+, Metaxa2 and SINTAX for all single HVR (**a**) and HVR pairs (**b**). Although Blast+ presented a slightly better performance over SINTAX and Metaxa2, it was not statistically significant. All three classification tools assigned more than 75% of the sequences using the DAIRYdb as a reference for all HVR pairs

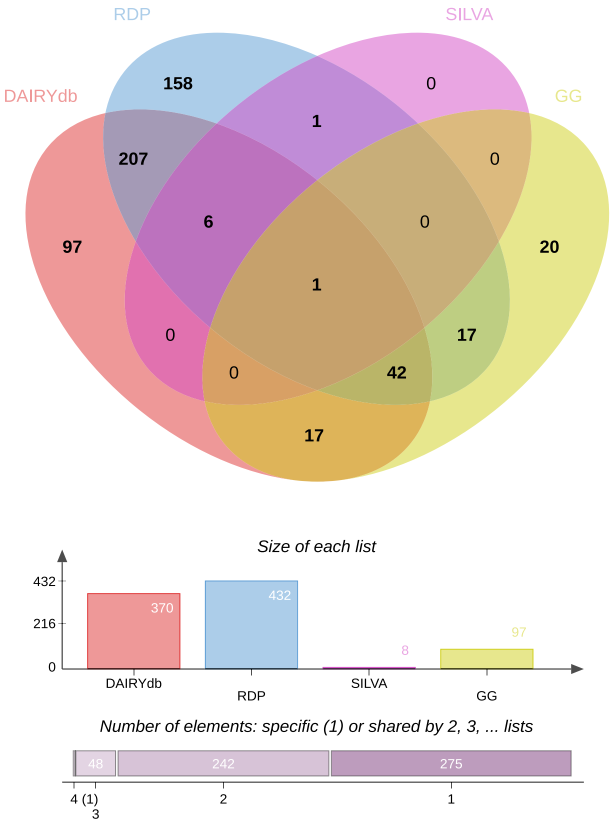

Fig. 8 Taxonomy annotation accuracy test on sequences from the HITdb. Comparison of the taxonomy annotation accuracy at species rank between the DAIRYdb, RDP, Silva and Greengenes on type strain sequences present in the HITdb [50]. On sequences from origin other than dairy products, the DAIRYdb performs better than Silva and Greengenes, but not better than the RPD trainset v16

in the Human Intestinal Tract database (HITdb) showed that the DAIRYdb performed comparably well to the RDP trainset v16 and better than Silva and Greengenes (Fig. 8) [50].

The study of every particular environment calls upon peculiar requirements. Dairy products are no exception, as their bacterial communities are usually dominated by few phylogenetically highly related species, which are often difficult to discern, such as *L. casei*, *L. paracasei* and *L. rhamnosus* or *S. thermophilus* and *S. salivarius*. Particularly for *S. thermophilus*, which is a very important representative bacterium in dairy products, the official name still is *S. salivarius* subsp. *thermophilus* [84]. Although a separate full species status was proposed [85], persistent contention prevented a full ratification by the taxonomic committees [84]. The advances of genomics in microbiology has led to a reassessment of the phylogeny, which still remains a moving target particularly for microbial taxonomy [48, 60].

The correct description of a bacterial community structure remains a challenge in microbiome studies. Any parameter, from wet-lab (*i.e.*, DNA extraction, primer and HVR selection, amplification, sequencing) to the bioinformatic pipeline, can influence the outcome. Although the achievement of over 90% accurate species annotation of short 16S fragments can be considered a dramatic improvement, quality of dairy products is often influenced by different strains of the same species [86]. The resolution at strain or subspecies rank, however, based on full 16S is highly unlikely to be achieved independently from advancing sequencing technology. While on the one hand the definition of strains and subspecies is even more problematic than higher ranks such as species [87], on the other hand, the intraspecies variability of the 16S lacks sufficient resolution to clearly discern between strains and subspecies within the same species [88]. Nevertheless, recent powerful bioinformatics tools, such as Oligotyping [89], Minimal Entropy Decomposition (MED) [90], Dada2 [91] and DiTaxa [92], can be applied to distinguish between ecologically relevant ASVs. The resulting oligotypes or haplotypes within a species might be linked to different metabolic pathways or associated to identified physico-chemical characteristics of cheese or dairy products. Hereof, the DAIRYdb is a powerful improvement as it accurately identifies the sequences belonging to the same species, which can further be decomposed to oligotypes. Finally, links between oligotypes and 16S from WGS could improve the link between phylogeny and ecotypes for a better ecological understanding of the system [81, 87, 93].

Yet, the way and ability to recognize the basic unit for taxonomy of prokaryotes depends on the resolution power of the observational methods available [94]. Increasing sequence read lengths will make it possible to cover three HVRs or even the full 16S, thus further improving taxonomic annotation accuracy at species rank by using a manually curated databases like the DAIRYdb.

Conclusions

Accurate prediction of taxonomy based on the marker gene 16S is a fundamental step in microbial diagnostics and microbial ecology studies. Dairy products, particularly cheeses, are enriched by a few dominant species often belonging to the same genera, such as *Lactobacillus* spp., *Lactococcus* spp., *Streptococcus* spp. An automatic and reliable taxonomic annotation to the correct species is pivotal to further routine microbial diagnostics.

Different to available universal databases, DAIRYdb achieved correct taxonomy annotation for ∼90% of species names on single HVRs and HVR pairs with 16S sequences present in dairy samples [52]. The better performance of the DAIRYdb over universal databases can be explained by the overall reduced number of sequences, only 10’290, with no conflicting taxonomy at all taxonomic ranks. Our results are in disagreement to the recommendation to use the largest and most diverse database possible for 16S classification [95]. On the opposite, manually curated 16S databases with authoritative full-length 16S sequences dedicated to the studied environment enormously improve classification confidence to the species rank [48–50]. Reducing the number of representative sequences to a minimal number in the training set further diminishes the risk of highly similar sequences with conflicting taxonomy, thus lowering the performance of the database used for classification [48, 49].

We therefore propose the manually curated DAIRYdb as a reference database for 16S microbiome studies on cheese and dairy products. The implementation of a curated database may lead to wider consensus and standardization processes reducing conflicts in literature due to the use of different universal databases integrated in different classification tools [82, 96].

Methods

DAIRYdb construction

To retrieve a comprehensive set of near full-length 16S sequences originating from cheese and dairy products, a search was performed against the NCBI Genebank nucleotide database using the command “CHEESE[All Fields] OR MILK[All Fields] OR TEAT[All Fields] OR STARTER[All Fields] OR WHEY[All Fields] OR DAIRY[All Fields] AND “16S ribosomal RNA”[All Fields] AND (“bacteria”[porgn] OR “archaea”[porgn]) AND 1000:2000[SLEN] NOT shotgun”, and the EMBL databases using the command “16S ribosomal RNA cheese, 16S ribosomal RNA milk, 16S ribosomal RNA starter, 16S ribosomal RNA teat, 16S ribosomal RNA dairy, 16S ribosomal RNA whey”. About 12’930 16S sequences were retrieved from NCBI and 51’175 from EMBL.

Mostly unpublished 16S sequences from INRA (171’282), as well as some recently published [97] (NCBI Bioproject PRJNA421256), and Agroscope (1’559) were also included after clustering at 99% (see below for details). The resulting set of 236’946 sequences constituted the starting pool for building up DAIRYdb. The extracted sequences of the starting pool were matched against the Greengenes v13_5, SSURef_NR99_123.1_tax_silva_trunc (Silva123) and RDP current_bacteria, RDP trainset v16_022016 16S databases using the *usearch_global* command of vsearch (v1.11.1) with parameters -id 0.8 and -mid 97. About 34’515 sequences were extracted from Greengenes, 29’181 from Silva123 and 52’686 from RDP, representing all possible full-length 16S sequences associated to the studied environment from the universal repositories prior to OTU clustering.

Altogether, the keywords associated sequences (236’946) and their representative database sequences (116’382) amounted to a total of 391’672 16S sequences of different length and quality. The extracted sequences were first dereplicated with priority to full-length sequences (279’737) and minimal length of 300 bp (268’613) if no full-length was present using the *derep_fulllength* and *derep_prefix* of usearch (v6.0.307_i86osx32), respectively. The remaining 268’613 sequences were divided into “good” sequences, *i.e.,* >1300 bp (50’309), and “bad” sequences, *i.e.,* <1300 bp (218’304), using Prinseq (v0.20.4). The “bad” sequences were matched against all databases (Greengenes, Silva and RDP) to obtain reference sequences of better quality and length using the *usearch_global* command of vsearch (v1.11.1) with parameters -id 0.5 and –maxhits 1. The 146'355 hit sequences were dereplicated using the *derep_fulllength* and *derep_prefix* of usearch6.0.307_i86osx32 resulting in 92'722 unique 16S sequences that were merged with the initial 50'309 "good" sequences. To remove redundancy, the 143'031 16S sequences were dereplicated again, reducing the total amount to 98'590 sequences and subsequently divided based on the length threshold of 1300 bp using Prinseq (v0.20.4) for clustering purposes. The sequences were first subjected to chimeric sequence removal using the 16S reference database available at www.drive5.com/uchime/rdp_gold.fa with usearch (v6.0.307_i86osx32) and subsequently clustered to OTUs separately to a similarity threshold of 97% in steps of 0.5% using the *cluster_otus* command in usearch (v6.0.307_i86osx32). The resulting 18'457 OTUs were matched against a database composed of full-length 16S sequences of type strains from Silva126 using the parameters –id 0.5 –mid 97 -maxhits 1. The sequences with no match were matched again, however, using less stringent parameter –id 0.5 -maxhits 1. Finally, the resulting 14'468 sequences were dereplicated, clustered at 97% and cleaned by removing eukaryotic, chloroplastic and mitochondrial sequences, Pintail value <75 and <1200 bp. The intermediate version of the DAIRYdb contained 9'739 representative 16S sequences.

Species whose sequences were present before the clustering at 97% and absent after clustering have been reintroduced if identical sequence with synonymous taxonomic identity was absent in the DAIRYdb (see List of Prokaryotic Names validly published of the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ)). Moreover, 142 missing species listed in [56] and [57] were added to the DAIRYdb.

DAIRYdb curation

The taxonomy of the 9'963 16S sequences assigned by Silva126 was checked manually upon consistency between the six top taxonomic ranks (Kingdom to Genus). The taxonomy annotations of these sequences were mostly obtained from authoritatively named type strains. While there could be some errors in the taxonomy annotations in Silva, or discordance to DSMZ, taxonomy annotations from authoritative type strains were considered as truth standards for the SATIVA test and manual curation. After a first manual curation, three closest type strain sequences with an authoritative taxonomy [48] were retrieved for all representative sequences in the DAIRYdb from Silva126 at 0.99 ANI and RDP (SequenceMatch: "typestrains", "isolates", "good quality", >1200 bp). The 7'244 closest neighbours (CN) were dereplicated and the resulting 4'545 CN type strain sequences were added to the DAIRYdb. The CN were used as references in the curation process with the Semi-Automatic Taxonomy Improvement and Validation Algorithm (SATIVA [59]) called with the command "sativa.py -s input.phyl -t input_taxonomy.txt -x BAC -T 36 -N 20 -S" using the authoritative CN type strain sequences as reference after alignment with SINA (v1.2.11 [58]).

Manual curation of the taxonomy was performed based on both, SATIVA results and RAxML tree with the sequences in the DAIRYdb and the authoritative CN [98] using the command "raxmplHPC-PTHREADS-AVX2 -T 36 -f a -k -x 237 -m GTRGAMMA -p 1481544944 -N autoMRE -s input.phyl -n output.tre -O -w". Using the algorithms Metaxa2 and SINTAX, an additional training set of Silva126 type strains was used to obtain a taxonomic proposition for sequences with unclear taxonomy. This step was iterated several times until all possible sequences could be assigned to a species group. The sequences with no unequivocal annotation where identified with LPT followed by an underscore and respective taxonomic rank (*e.g.*, *XXX_Family*, *Lactobacillus_Species*).

At least one type strain for each bacterial species present in the DAIRYdb was added if available in the DSMZ database bacdive.dsmz.de/ (https://bacdive.dsmz.de/). After manual curation the 4'545 CN reference sequences with authoritative taxonomy were removed. The final DAIRYdb.v1.1 contains a total of 10'290 16S sequences.

DAIRYdb validation and statistical analyses

Single HVR and pairs of HVR were extracted in silico from the 10'290 16S sequences of the DAIRYdb using primers listed in Table 1 with the *pcr.seqs* function from mothur with 0 missmatches. All present single and pairs of HVRs in the 16S sequences were extracted with V-Xtractor, which is based on a HMM algorithm for the HVR detection and where extraction occurs based on the presence of the HVR with no primer biases.

Taxonomy annotation accuracy was assessed with 1000 subsamples of each 100 randomly selected sequences from within the DAIRYdb. Annotation accuracy is defined as the fraction of sequences that are correctly predicted and classified at each rank [52]. The sequences were taxonomically assigned at all taxonomic ranks by means of three classification tools (Blast+, Metaxa2 and, SINTAX). Annotation accuracy was assessed using 4 universal databases: Greengenes v13.5 (1'262'986 sequences) [45], SILVA v128 SSURef Nr99 (645'151 sequences) [40], LTP v123 SSU (11'939 sequences) [42], RDP training set v16 (13'212 sequences) [43] and the DAIRYdb (10'290 sequences). Correct annotation at all taxonomic ranks was assessed comparing the resulting assignment with the reference DAIRYdb. Pairwise comparisons using Wilcoxon rank sum test were used to highlight significant differences between database assignment accuracy. The *p-values* obtained were adjusted using Bonferroni correction. For more details on the statistical evaluation process see Additional file 4.

Visualizations

Krona

A Krona chart was generated using KronaTools [99]. The total sum of each unique taxonomy of the DAIRYdb was assessed and used as input in the ktImportText command.

Venn diagram

A Venn diagram to determine the potential origin of each sequence composing the DAIRYdb was generated with the web app Jvenn [100]. The global alignment -usearch_global module of vsearch to per a global alignment between the 10'290 sequences of the DAIRYdb and each raw database corresponding to keywords: dairy, cheese, starter, whey, milk. These raw databases were previously dereplicated. We used the parameter -strand both -id 0.9 -maxaccepts 100 -maxrejects 100. For each keywords, the matching sequence identifiers were used as input in Jvenn to generate the Venn diagram.

Additional files

Additional file 1: Species list. The additional file 1 is a text file with all species included in the DAIRYdb in alphabetical order. Can be opened with any text editor or spreadsheet software. (TXT 100 kb)

Additional file 2: Krona diagram. Additional file 2 is an html file with a Krona diagram showing the complete diversity present in the DAIRYdb can be interactively inspected in a webbrowser. (HTML 872 kb)

Additional file 3: Supplementary Information. Additional file 3 contains supplementary figures described in the main manuscript. It is a portable document file (pdf) that can be read with Acrobat Reader. (PDF 3925 kb)

Additional file 4: RMarkdown for reproducibility. The additional file 4 includes the different scripts used to test all HVR primers (for single HVR and HVR pairs), scripts used to customize the different database in order to use them in the different assignation tools, and scripts used for DAIRYdb validation. We mainly used bash and R scripts [101, 102]. The file is in html format and was generated starting from a Rmarkdown file. (HTML 1012 kb)

Abbreviations

ANI: Average nucleotide identity; ASV: Amplicon sequence variant; BacDive: Bacterial diversity metadatabase; CN: Closest neighbour; DAIRYdb: Dairy agroscope INRA ribosomal accuracY database; DGGE: Denaturing gradient gel electrophoresis; DNA: Deoxyribonucleic acid; DSMZ: Deutsche Sammlung von Mikroorganismen und Zellkulturen; GG: Greengenes database; HITdb: Human intestinal tract database; HMM: Hidden Markov-Monte-Carlo model; HVR: Hyper variable region (16S); INSDC: International nucleotide sequence database collaboration; LAB: Lactic acid bacteria; LCA: Lowest common ancestor; LCR: Lowest common rank; LH-PCR: Length heterogeneity polymerase chain reaction; LPSN: List of prokaryotic names with standing in nomenclature; LSU: Large subunit; LTP: Living tree project; MED: Minimal entropy decomposition; NCBI: National center for biotechnology information; NGS: Next generation sequencing; OTU: Operational taxonomic unit; RDP: Ribosomal database project; rRNA: Ribosomal ribonucleic acid; SSU: Small subunit; T-RFLP: Terminal restriction fragment length polymorphism; WGS: Whole genome shotgun

Acknowledgements

Eric Dugat-Bony (INRA Grignon, GMPA) improved the completeness of the DAIRYdb by providing some additional 16S sequences from his own studies related to cheese rind bacteria. We are grateful to the Genotoul bioinformatics platform (Toulouse, Occitanie, https://www.genotoul.fr/) and INRA MIGALE bioinformatics platform (Jouy-en-Josas, Ile-de-France, http://migale.jouy.inra.fr) for providing computing and storage resources.

Authors' contributions

All authors conceived and designed the study. MM and ER constructed, curated and validated the DAIRYdb. MM drafted the manuscript, with help from ER, NS, CD, HB and CC. ER drafted the RMarkdown script. All authors contributed to and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The DAIRYdb is available on a public repository (https://github.com/marcomeola/DAIRYdb). The different databases used to validate the DAIRYdb are available on their own website.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Agroscope, Competence Division Methods Development and Analytics, Research Group Fermenting Organisms, Schwarzenburgstrasse 161, 3003 Bern, Switzerland. [2] Université Clermont Auvergne, INRA, VetAgro Sup, UMRF, 20 côte de Reyne, 15000 Aurillac, France.

Received: 7 August 2018 Accepted: 18 June 2019
Published online: 08 July 2019

References

1. Porter TM, Hajibabaei M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Mol Ecol. 2018;27(2): 313–38. https://doi.org/10.1111/mec.14478.
2. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Agosto Rivera JL, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll JL, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart HP, Guyton K, Haig SJ, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler AV, Krezalek M, Kueneman J, Lamendella R, Landon EM, Lanede Graaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME,

Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB. Po: A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681): 457–63. https://doi.org/10.1038/nature24621.

3. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. Ocean plankton. Structure and function of the global ocean microbiome. Sci (NY). 2015;348(6237): 1261359. https://doi.org/10.1126/science.1261359. NIHMS150003.
4. Moran MA. The global ocean microbiome. Science. 2015;350(6266):. https://doi.org/10.1126/science.aac8455.
5. Meola M, Lazzaro A, Zeyer J. Bacterial composition and survival on Sahara dust particles transported to the European Alps. Front Microbiol. 2015;6(DEC):1–17. https://doi.org/10.3389/fmicb.2015.01454.
6. Pearce DA, Hughes KA, Lachlan-Cope T, Harangozo SA, Jones AE. Biodiversity of air-borne microorganisms at Halley station, Antarctica. Extremophiles. 2010;14(2):145–59. https://doi.org/10.1007/s00792-009-0293-8.
7. Lazzaro A, Wismer A, Schneebeli M, Erny I, Zeyer J. Microbial abundance and community structure in a melting alpine snowpack. Extremophiles. 2015;19(3):631–42. https://doi.org/10.1007/s00792-015-0744-3.
8. Christner B, Skidmore M, Priscu J, Tranter M, Foreman C. Bacteria in subglacial environments. In: Margesin, R, Schinner, F, M, J-C, G, C (Eds.), Psychrophiles: From Biodiversity to Biotechnology. Berlin: SpringerVerlad; 2008, p. 51e71.
9. Yeluri Jonnala BR, McSweeney PLH, Sheehan JJ, Cotter PD. Sequencing of the cheese microbiome and its relevance to industry. Front Microbiol. 2018;9:1020. https://doi.org/10.3389/fmicb.2018.01020.
10. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012;486(7402):215–21. https://doi.org/10.1038/nature11209.
11. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13(4):260–70. https://doi.org/10.1038/nrg3182.
12. Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. Nat Rev Microbiol. 2018;16(3):143–55. https://doi.org/10.1038/nrmicro.2017.157.
13. Fierer N, Nemergut D, Knight R, Craine JM. Changes through time: integrating microorganisms into the study of succession. Res Microbiol. 2010;161(8):635–42. https://doi.org/10.1016/j.resmic.2010.06.002.
14. Meola M, Lazzaro A, Zeyer J. Diversity, resistance and resilience of the bacterial communities at two alpine glacier forefields after a reciprocal soil transplantation. Environ Microbiol. 2014;16(6):1918–34. https://doi.org/10.1111/1462-2920.12435.
15. Shade A, Gregory Caporaso J, Handelsman J, Knight R, Fierer N. A meta-analysis of changes in bacterial and archaeal communities with time. ISME J. 2013;8:1493–506. https://doi.org/10.1038/ismej.2013.54.
16. Ramazzotti M, Bacci G. Chapter 5 - 16s rrna-based taxonomy profiling in the metagenomics era. In: Nagarajan M, editor. Metagenomics. Academic Press; 2018. p. 103–19. https://doi.org/10.1016/B978-0-08-102268-9.00005-7.
17. Vinje H, Liland KH, Almøy T, Snipen L. Comparing K-mer based methods for improved classification of 16S sequences. BMC Bioinformatics. 2015;16(1):1–13. https://doi.org/10.1186/s12859-015-0647-4.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
19. Wang Q, Garrity GM, Tiedje JM, Cole JR. Na??ve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7. https://doi.org/10.1128/AEM.00062-07. Wang, Qiong, 2007, Naive.
20. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. PLOS Genet. 2008;4(11):1–10. https://doi.org/10.1371/journal.pgen.1000255.
21. Mitra S, Stärk M, Huson DH. Analysis of 16s rrna environmental sequences using megan. BMC Genomics. 2011;12(3):17. https://doi.org/10.1186/1471-2164-12-S3-S17.
22. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH. Metaxa2: improved identification and taxonomic classification of small and large subunit rrna in metagenomic data. Mol Ecol Resour. 2015;15(6):1403–14. https://doi.org/10.1111/1755-0998.12399.
23. Ramazzotti M, Berná L, Donati C, Cavalieri D. riboframe: An improved method for microbial taxonomy profiling from non-targeted metagenomics. Front Genet. 2015;6:329. https://doi.org/10.3389/fgene.2015.00329.
24. Allard G, Ryan FJ, Jeffery IB, Claesson MJ. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC Bioinformatics. 2015;1:324. https://doi.org/10.1186/s12859-015-0747-1.
25. Somervuo P, Koskela S, Pennanen J, Henrik Nilsson R, Ovaskainen O. Unbiased probabilistic taxonomic classification for dna barcoding. Bioinformatics. 2016;32(19):2920–7. https://doi.org/10.1093/bioinformatics/btw346.
26. Edgar R. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv. 2016074161. https://doi.org/10.1101/074161.
27. Mysara M, Vandamme P, Props R, Kerckhof F-M, Leys N, Boon N, Raes J, Monsieurs P. Reconciliation between operational taxonomic units and species boundaries. FEMS Microbiol Ecol. 2017;93(4):029. https://doi.org/10.1093/femsec/fix029.
28. Sherman DJ. Humidor: Microbial community classification of the 16 s gene by training cigar strings with convolutional neural networks; 2017.
29. Matias Rodrigues JF, Schmidt TSB, Tackmann J, von Mering C. Mapseq: highly efficient k-mer search with confidence estimates, for rrna sequence analysis. Bioinformatics. 2017;33(23):3808–10. https://doi.org/10.1093/bioinformatics/btx517.
30. Liland KH, Vinje H, Snipen L. microclass: an R-package for 16S taxonomy classification. BMC Bioinformatics. 2017;18(1):172. https://doi.org/10.1186/s12859-017-1583-2.
31. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90. https://doi.org/10.1186/s40168-018-0470-z.
32. Murali A, Bhargava A, Wright ES. Idtaxa: a novel approach for accurate taxonomic classification of microbiome sequences. Microbiome. 2018;6(1):140. https://doi.org/10.1186/s40168-018-0521-5.
33. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41. https://doi.org/10.1128/AEM.01541-09.
34. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. Qiime allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6. https://doi.org/10.1038/nmeth.f.303.
35. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MG, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T,

Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson II MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CH, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Prepr. 2018;6:27295–2. https://doi.org/10.7287/peerj.preprints.27295v2.

36. Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G. Frogs: Find, rapidly, otus with galaxy solution. Bioinformatics. 2018;34(8):1287–94. https://doi.org/10.1093/bioinformatics/btx791.
37. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?. BMC Genomics. 2017;18(S2):114. https://doi.org/10.1186/s12864-017-3501-4.
38. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies J, Westram R, Ludwig W. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotechnol. 2017;February:0–1. https://doi.org/10.1016/j.jbiotec.2017.06.1198.
39. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35(21):7188–96. https://doi.org/10.1093/nar/gkm864.
40. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):590–6. https://doi.org/10.1093/nar/gks1219.
41. Parte AC. Lpsn–list of prokaryotic names with standing in nomenclature. Nucleic Acids Res. 2014;42(Database issue):613–6. https://doi.org/10.1093/nar/gkt1111.
42. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014;42(D1):643–8. https://doi.org/10.1093/nar/gkt1209.
43. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37(SUPPL. 1):141–5. https://doi.org/10.1093/nar/gkn879.
44. Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2016;44(D1):48–50. https://doi.org/10.1093/nar/gkv1323.
45. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72. https://doi.org/10.1128/AEM.03006-05.
46. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2011;39(suppl 1):38–51. https://doi.org/10.1093/nar/gkq1172.
47. Christen R. Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. Microbes Environ. 2008;23(4):253–68.
48. Edgar R. Taxonomy annotation and guide tree errors in 16s rrna databases. PeerJ. 2018;6:5030. https://doi.org/10.7717/peerj.5030.
49. Newton IL, Roeselers G. The effect of training set on the classification of honey bee gut microbiota using the naïve bayesian classifier. BMC Microbiol. 2012;12(1):221. https://doi.org/10.1186/1471-2180-12-221.
50. Ritari J, Salojärvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics. 2015;16(1):1056. https://doi.org/10.1186/s12864-015-2265-y.
51. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM, Albertsen M, Nielsen PH. Midas 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. Database (Oxf). 2017;2017(1):. https://doi.org/10.1093/database/bax016.
52. Edgar RC. Accuracy of taxonomy prediction for 16s rrna and fungal its sequences. PeerJ. 2018;6:4652. https://doi.org/10.7717/peerj.4652.
53. Escobar-Zepeda A, De León AVP, Sanchez-Flores A. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. Front Genet. 2015;6(DEC):1–15. https://doi.org/10.3389/fgene.2015.00348.
54. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
55. Bengtsson-Palme J, Richardson RT, Meola M, Wurzbacher C, Tremblay ÉD, Thorell K, Kanger K, Eriksson KM, Bilodeau GJ, Johnson RM, Hartmann M, Henrik Nilsson R. Metaxa2 database builder: Enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. Bioinformatics482. https://doi.org/10.1093/bioinformatics/bty482.
56. Montel MC, Buchin S, Mallet A, Delbes-Paus C, Vuitton DA, Desmasures N, Berthier F. Traditional cheeses: Rich and diverse microbiota with associated benefits. Int J Food Microbiol. 2014;177(May):136–54.
57. Irlinger F, Layec S, Hélinck S, Dugat-Bony E. Cheese rind microbial communities: diversity, composition and origin. FEMS Microbiol Lett. 2015;362(2):1–11. https://doi.org/10.1093/femsle/fnu015.
58. Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28(14):1823–9. https://doi.org/10.1093/bioinformatics/bts252.
59. Kozlov AM, Zhang J, Yilmaz P, Glöckner FO, Stamatakis A. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res. 2016;44(11):5022–33. https://doi.org/10.1093/nar/gkw396.
60. Konstantinidis KT, Rosselló-Móra R, Amann R. Uncultivated microbes in need of their own taxonomy. ISME Journal. 2017;11(11):2399–406. https://doi.org/10.1038/ismej.2017.113.
61. McAuliffe O. Chapter 9 Genetics of Lactic Acid Bacteria; 2017, pp. 227–47. https://doi.org/10.1016/b978-0-12-417012-4.00009-0. Exported from https://app.dimensions.ai on 2018/12/10.
62. Salvetti E, Torriani S, Felis GE. The genus lactobacillus: A taxonomic update. Probiotics Antimicrob Protein. 2012;4(4):217–26. https://doi.org/10.1007/s12602-012-9117-8.
63. Wuyts S, Wittouck S, De Boeck I, Allonsius CN, Pasolli E, Segata N, Lebeer S. Large-scale phylogenomics of the lactobacillus casei group highlights taxonomic inconsistencies and reveals novel clade-associated features. mSystems. 2017;2(4):. https://doi.org/10.1128/mSystems.00061-17.
64. Salvetti E, Harris HMB, Felis GE, O'Toole PW. Comparative genomics reveals robust phylogroups in the genus lactobacillus as the basis for reclassification. Appl Environ Microbiol. 2018. https://doi.org/10.1128/AEM.00993-18.
65. Søhngen C, Podstawka A, Bunk B, Gleim D, Vetcininova A, Reimer LC, Ebeling C, Pendarovski C, Overmann J. Bacdive - the bacterial diversity metadatabase in 2016. Nucleic Acids Res. 2016;44(D1):581–5. https://doi.org/10.1093/nar/gkv983.
66. Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. J Microbiol Methods. 2010;83(2):250–3. https://doi.org/10.1016/j.mimet.2010.08.008.
67. Hongoh Y, Yuzawa H, Ohkuma M, Kudo T. Evaluation of primers and pcr conditions for the analysis of 16s rrna genes from a natural environment. FEMS Microbiol Lett. 2003;221(2):299–304.
68. Sundquist A, Bigdeli S, Jalili R, Druzin ML, Waller S, Pullen KM, El-Sayed YY, Taslimi MM, Batzoglou S, Ronaghi M. Bacterial flora-typing with targeted, chip-based pyrosequencing. BMC Microbiol. 2007;7:108. https://doi.org/10.1186/1471-2180-7-108.
69. Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA. Combination of 16s rrna-targeted oligonucleotide probes with flow

cytometry for analyzing mixed microbial populations. Appl Environ Microbiol. 1990;56(6):1919–25.

70. el Fantroussi S, Verschuere L, Verstraete W, Top EM. Effect of phenylurea herbicides on soil microbial communities estimated by analysis of 16s rrna gene fingerprints and community-level physiological profiles. Appl Environ Microbiol. 1999;65(3):982–8.
71. Muyzer G, de Waal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16s rrna. Appl Environ Microbiol. 1993;59(3):695–700.
72. Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. Primerprospector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics. 2011;27(8):1159–61. https://doi.org/10.1093/bioinformatics/btr087.
73. Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, Brodie EL, Malamud D, Poles MA, Pei Z. Design of 16s rrna gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol. 2010;16(33):4135–44.
74. Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16s rrna. Appl Environ Microbiol. 1997;63(11):4516–22.
75. Keijser BJF, Zaura E, Huse SM, van der Vossen JMBM, Schuren FHJ, Montijn RC, ten Cate JM, Crielaard W. Pyrosequencing analysis of the oral microflora of healthy adults. J Dent Res. 2008;87(11):1016–20. https://doi.org/10.1177/154405910808701104.
76. Walker JJ, Pace NR. Phylogenetic composition of rocky mountain endolithic microbial ecosystems. Appl Environ Microbiol. 2007;73(11): 3497–504. https://doi.org/10.1128/AEM.02656-06.
77. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16s primers. J Microbiol Methods. 2003;55(3):541–55. https://doi.org/10.1016/j.mimet.2003.08.009.
78. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16s rrna gene sequencing of mock microbial populations- impact of dna extraction method, primer choice and sequencing platform. BMC Microbiol. 2016;16(1):123. https://doi.org/10.1186/s12866-016-0738-z.
79. Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rrna gene-based environmental surveys. Appl Environ Microbiol. 2009;75(16):5227–36. https://doi.org/10.1128/AEM.00592-09.
80. Schloss PD, Westcott SL. Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rrna gene sequence analysis. Appl Environ Microbiol. 2011;77(10):3219–26. https://doi.org/10.1128/AEM.02810-10.
81. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43. https://doi.org/10.1038/ismej.2017.119.
82. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: Attempting to find consensus "best practice" for 16s microbiome studies. Appl Environ Microbiol. 2018;84(7):. https://doi.org/10.1128/AEM.02627-17.
83. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12(9):635–45. https://doi.org/10.1038/nrmicro3330.
84. Burton JP, Chanyi RM, Schultz M. Chapter 19 - common organisms and probiotics: Streptococcus thermophilus (streptococcus salivarius subsp. thermophilus). In: Floch MH, Ringel Y, Walker WA, editors. The Microbiota in Gastrointestinal Pathophysiology. Boston: Academic Press; 2017. p. 165–9. https://doi.org/10.1016/B978-0-12-804024-9.00019-7.
85. Schleifer KH, Ehrmann M, Krusch U, Neve H. Revival of the species streptococcus thermophilus (ex orla-jensen, 1919) nom. rev. Syst Appl Microbiol. 1991;14(4):386–8. https://doi.org/10.1016/S0723-2020(11)80314-0.
86. van Mastrigt O, Di Stefano E, Hartono S, Abee T, Smid EJ. Large plasmidome of dairy lactococcus lactis subsp. lactis biovar diacetylactis fm03p encodes technological functions and appears highly unstable. BMC Genomics. 2018;19(1):620. https://doi.org/10.1186/s12864-018-5005-2.
87. Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E, Cohan FM. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci USA. 2008;105(7):2504–9. https://doi.org/10.1073/pnas.0712205105.
88. Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today. 2006;8:6–9.
89. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Oligotyping: differentiating between closely related microbial taxa using 16s rrna gene data. Methods Ecol Evol. 2013;4(12):1111–9. https://doi.org/10.1111/2041-210X.12114.
90. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 2015;9(4):968–79. https://doi.org/10.1038/ismej.2014.195.
91. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3. https://doi.org/10.1038/nmeth.3869. 15334406.
92. Asgari E, Münch PC, Lesker TR, McHardy AC, Mofrad MRK. Ditaxa: Nucleotide-pair encoding of 16s rrna for host phenotype and biomarker detection. Bioinformatics. 2018954. https://doi.org/10.1093/bioinformatics/bty954.
93. Berry MA, White JD, Davis TW, Jain S, Johengen TH, Dick GJ, Sarnelle O, Denef VJ. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in Freshwater lakes. Front Microbiol. 2017;8(MAR):1–7. https://doi.org/10.3389/fmicb.2017.00365.
94. Rosselló-Móra R. Towards a taxonomy of bacteria and archaea based on interactive and cumulative data repositories. Environ Microbiol. 2012;14(2):318–34. https://doi.org/10.1111/j.1462-2920.2011.02599.x.
95. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16s rrna gene surveys. ISME J. 2012;6(1):94–103. https://doi.org/10.1038/ismej.2011.82.
96. Overcoming hurdles in sharing microbiome data. 2017;2:1573. https://doi.org/10.1038/s41564-017-0077-3.
97. Frétin M, Martin B, Rifa E, Isabelle V-M, Pomiès D, Ferlay A, Montel M-C, Delbès C. Bacterial community assembly from cow teat skin to ripened cheeses is influenced by grazing systems. Sci Rep. 2018;8(1):200. https://doi.org/10.1038/s41598-017-18447-y.
98. Stamatakis A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033.
99. Ondov BD, Bergman NH, Philippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinformatics. 2011;12:385. https://doi.org/10.1186/1471-2105-12-385.
100. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn: an interactive venn diagram viewer. BMC Bioinformatics. 2014;15:293. https://doi.org/10.1186/1471-2105-15-293.
101. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2018. https://www.R-project.org/.
102. Wickham H. Ggplot2: Elegant Graphics for Data Analysis: Springer; 2009. http://ggplot2.org.

Publisher's Note