

# DAIRYdb: A manually curated gold standard reference database for improved taxonomy annotation of 16S rRNA gene sequences from dairy products

Marco Meola<sup>1\*</sup>, Etienne Rifa<sup>2\*</sup>, Noam Shani<sup>1</sup>, Céline Delbès<sup>2</sup>, Hélène Berthoud<sup>1</sup>, and Christophe Chassard<sup>2</sup>

<sup>1</sup>Research Group Fermenting Organisms, Competence Division Methods Development and Analytics, Agroscope, Schwarzenburgstrasse 161, 3003 Bern, Switzerland

<sup>2</sup>Université Clermont Auvergne, INRA, UMR 0545 UMRF Unité Mixte de Recherche sur le Fromage, 20 côte de Reyne, 15000 Aurillac, France

\*Equal contributor

Reads assignment to taxonomic units is a key step in microbiome analysis pipelines. To date, accurate taxonomy annotation, particularly at species rank, is still challenging due to the short size of read sequences and differently curated classification databases. However, the close phylogenetic relationship between species encountered in dairy products requires accurate species annotation to achieve sufficient phylogenetic resolution for further downstream ecological studies or for food diagnostics. Taxonomy annotation in universal 16S databases with environmental sequences like Silva, RDP or Greengenes is based on predictions rather than on studies of type strains or isolates. 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 classification 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.

microbiome | taxonomy annotation | OTU classification | 16S | database | accuracy | dairy | cheese | milk | whey | teat | starter |

Correspondence: [marco.meola@agroscope.admin.ch](mailto:marco.meola@agroscope.admin.ch)

## Introduction

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, since unculturable, 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. An infinite number of descriptive studies have been published describing microbial community structures in various environments, often correlating

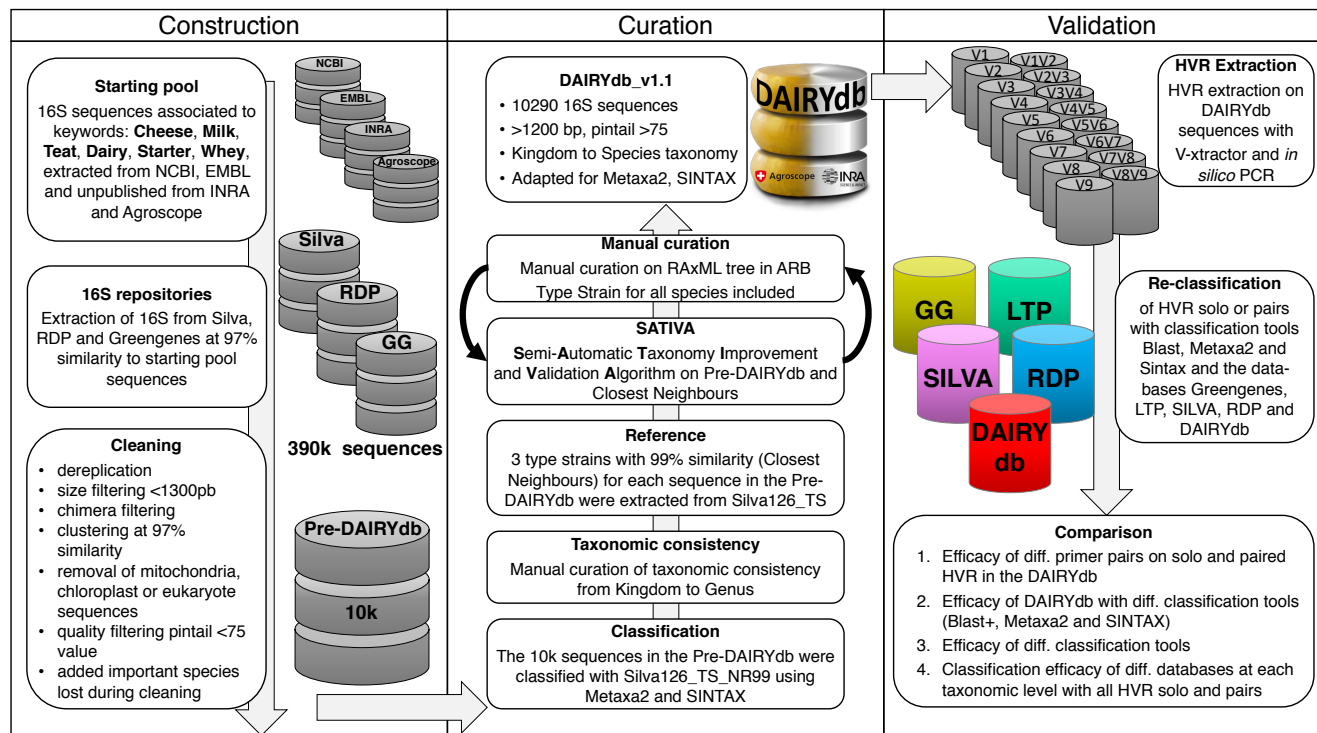
their dynamic changes over time or space by means of the 16S rRNA gene (13–15).

**The marker gene 16S rRNA.** The 16S rRNA gene (16S) was proposed by Carl Woese and George Fox in the 1980s as the gold standard marker gene for molecular taxonomic research (16–18). Several characteristics of the 16S make it a marker gene for surveys of microbial diversity: i) it is an ubiquitous and highly conserved gene in prokaryotes, ii) it has a functional degree and size presenting clock-like mutation rates like an evolutionary chronometer, iii) and the presence of alternating conserved and hypervariable regions (HVRs) permit to design universal primers on the conserved regions and to use the HVRs (V1–V9) for taxonomic classification (19, 20).

Before the advent of NGS, microbial community studies were based on fingerprinting techniques, such as DGGE, T-RFLP or LH-PCR sometimes in combination with Sanger sequencing of the complete 16S spanning over about 1550 bp. 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 (20).

**Classification tools.** 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 operational taxonomic units (OTUs) by the available 16S databases (21). On the one hand, NGS has triggered the acquisition of enormous amounts of sequencing data, offering the possibility to overcome the limitations of Sanger sequencing. On the other hand it has brought huge computational challenges, such as the risk of taxonomic mis-annotations or ambiguous results during taxonomic classification steps. Despite the steadily increasing read length obtained by NGS, 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 (22).

Numerous classification predictors algorithms have been developed and optimized in recent years with the aim to ac-



**Fig. 1.** Development of the DAIRYdb consisted in three main steps: construction, curation and validation. For construction, dairy products specific 16S rRNA gene 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 identification accuracy of single and HVR pairs by the five databases (Greengenes 13.8, LTP version, Silva 128 NR99, RDP version and DAIRYdb).

curately annotate the taxonomy of OTUs 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 (23). In recent years, more 16S specific taxonomy predictors have been developed, including RDP Naive Bayesian Classifier (NBC) (24), a naive Bayesian Classifier based on k-mers, GAST (25), MEGAN (26), Metaxa2 (27), riboFrama (28), SPINGO (29), PROTAX (30), SINTAX (31), DynamiC (32), Humidor (33), MAPseq (34), microclass (35) and other tools implemented in the most current 16S pipelines like mothur (36), Qiime v1 (37), Qiime v2 (<https://qiime2.org>) and FROGS (38).

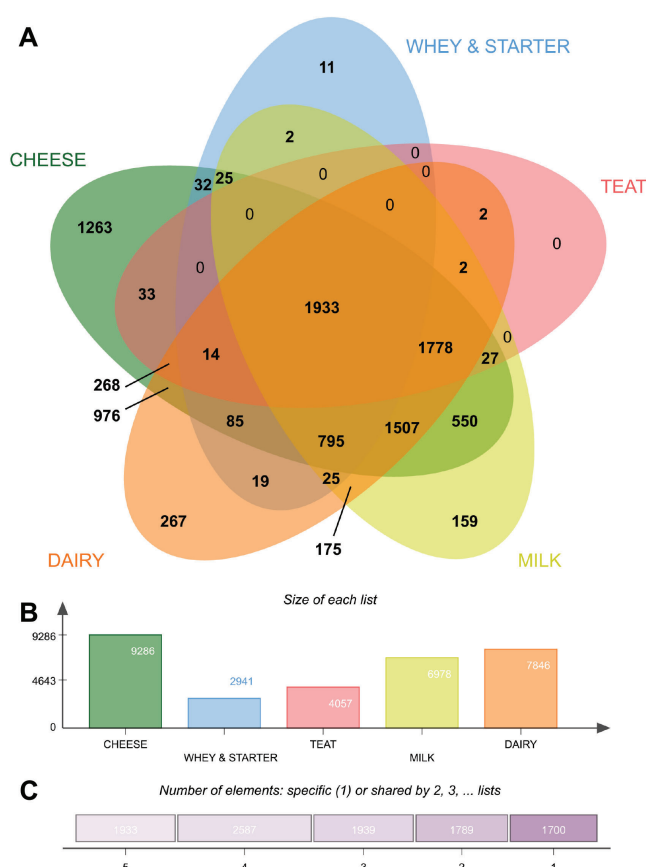
Here we used three taxonomy predictors based on different algorithms and programming languages (39), in which dedicated databases can be integrated and used for classification. One of the three classification tools used in this study to test the performance of the DAIRYdb was the updated version of Blast, Blast+ (40). It is based on an heuristic method for identification of database sequences that resemble the query sequence above a certain threshold. It searches for short sequence matches, which are locally aligned after the first match (40).

Another classification tool used in this study was Metaxa2, an HMM based software tool for automated detection and classification of short NGS fragments, such as ribosomal small

and large subunits, SSU and LSU, respectively, or any gene of interest useful for classification of any organism (27, 41). The third taxonomy predictor used in this study was SINTAX, a non-Bayesian taxonomy classifier specific for 16S sequences, which uses k-mer similarity to identify the top hit in a reference database providing bootstrap confidence values at each taxonomic rank (31).

**16S repositories.** 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 (42).

Silva is the universal 16S repository with the highest number of sequences. The latest release of Silva SSU/LSU 132 ([www.arb-silva.de](http://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 (43–45). 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) (46). Minimal training sets, such as the SSU Ref NR 99 or the LTP (47), offer a reduced number of sequences for faster classification but still cover-

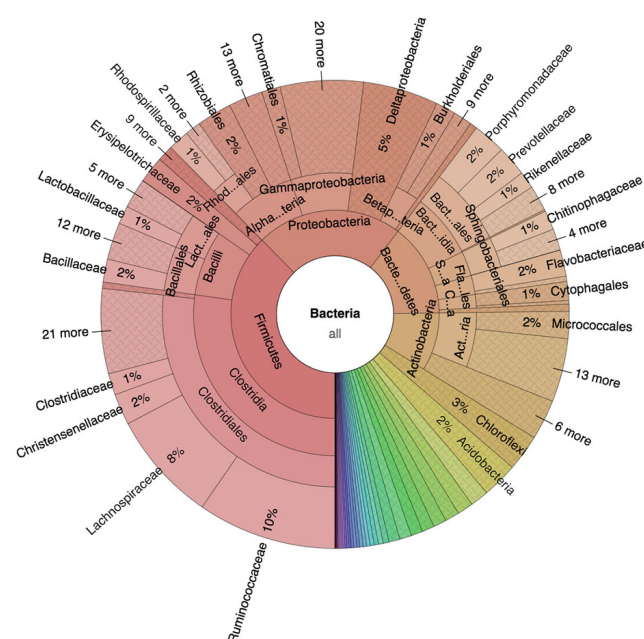


**Fig. 2.** Origin of sequence in the DAIRYdb. A) Five-factors Venn diagram comparing the origins of the sequences (9'948) retrieved from the public repositories Genbank NCBI and EMBL associated to the keywords "cheese", "dairy", "milk", "teat" and "whey/starter". About 12.7% (1'263) sequences were only 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.

ing the broadest currently known biodiversity.

The second biggest repository, the Ribosomal Database Project (RDP Release 11, Update 5; <http://rdp.cme.msu.edu>) (48), contained at the time of writing 3'356'809 16S sequences from the International Nucleotide Sequence Database Collaboration (INSDC) (49). The nomenclature is based on the Bacterial Nomenclature Up-to-Date and the taxonomic rank information on the Bergey's Manual. Greengenes v13\_5 (50) 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 (51). Although frequently used in community studies together with Qiime (37), the last update dates back to 2013 with no indication for an imminent update.

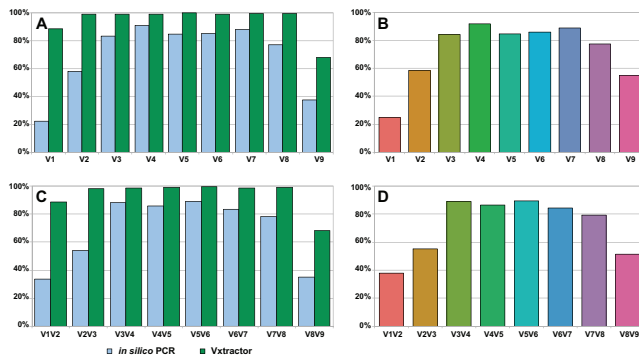
**Taxonomy annotation in microbiology.** Phylogenetic classification has tailored taxa by means of phylogenetic, phenotype and genomic coherence that make taxonomic units unique within the classification schema (52). Phylogenetic coherence is determined by the 16S for which the previously mentioned databases provide a valuable tool for tax-



**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 phyla 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.

onomic classification (52). However, the exponential increase of 16S sequences from previously unknown and uncultured bacteria led to an explosion of exotic labels at any taxonomic rank with often contrasting taxonomic classifications between databases based on different taxonomic catalogues (e.g., Bergey's Taxonomic Outlines, List of Prokaryotic Names with Standing Nomenclature (LPSN), International Sequence Database Collaboration (INSDC), Bacterial Nomenclature Up-to-Date (53, 54)).

The lack of consensus on a widely accepted taxonomy, as well as the lack of taxonomic characterisation of yet uncultured bacteria, are severely limiting communication among scientists and may lead to incorrect annotations and thus "poison every experiment that makes use of them" (52, 55). In worst cases, incorrectly annotated bacteria are included in databases further used to classify new sequences. In fact, it has been shown that there are numerous unambiguous disagreements between the nomenclature hierarchies, where many taxon names are placed into different parent taxa in the databases Silva and Greengenes or the taxonomy is not consistent with the tree (54). In addition to hierarchy disagreements, about 34% of identical sequences in Silva and Greengenes databases presented annotation conflicts, 24% of which were blanks (unclassified) in either of both database (54). While the blanks imply a false negative by one database or a false positive by the other, differently annotated sequences with the same Accession number or identical sequence string must be due to annotation error in one or both databases (54).



**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 silico PCR with mothur (B) and (D) and HVR extraction with V-Extractor (A) and (C) 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).

While NGS has allowed researchers to obtain deep insights into the microbial community structures inhabiting various environments, the complexity of the analytical process and taxonomy annotation on short read sequences is still challenging and prevents researchers to deploy microbial community analysis for diagnostic purposes in an accurate and reproducible way (1, 67). Previous studies have highlighted the importance of high-quality data for improving the classifications of the obtained OTUs (22, 68, 69). 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, classification accuracy at lower taxonomic ranks increases with a gold standard training set encompassing only full-length and good quality representative sequences innate to the investigated environment (22, 54, 69, 70).

In microbiology, distinction is made between the concept and the definition of species (52). The species concept explains the idea of what is considered to be a species as a unit of biodiversity, the meaning of the patterns of recurrence observed in nature, and the reason for their existence (71). The species definition, however, is concretely the set of parameters that are applied to circumscribe the category (72).

Thanks to the dropping costs, NGS is increasingly applied routinely as diagnostic technology for quality assessments and microbial community analyses in dairy products. Several initiatives aimed at tracking from "Farm to Fork" the range of expected microbial communities along the food supply chain, such as Food Safety Consortium with IBM, Mars and Bio-Rad Laboratories ([www-03.ibm.com/press/us/en/pressrelease/45938.wss](http://www-03.ibm.com/press/us/en/pressrelease/45938.wss) and [www-03.ibm.com/press/us/en/pressrelease/52690.wss](http://www-03.ibm.com/press/us/en/pressrelease/52690.wss)). However, 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, particularly. Therefore,

**Table 1.** Primers used in the *in silico* PCR extraction of the HVRs. \*E. coli position as a reference.

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	(56)	
120R_v1r	S-D-Bact-0120-e-A-20	v1r	101-120	20	TTACTCACCCGTGCGCCCT	55	mod. rev-compl. after (57)	
101F_v2f	S-D-Bact-0101-a-S-20	v2f	101-120	20	AGYGGCGNACGGGTGAGTAA	55	mod. after (57)	
355R_v2r	S-D-Bact-0355-a-A-18	v2r	338-355	18	GCWGGCTCCCGTAGAGT	66	mod. after (58)	
338F_v3f	S-D-Bact-0338-a-S-20	v3f	337-354	20	ACWCCTACGGGCGGAGCAG	65	mod. after (59)	
534R_v3r	S-D-Bact-0518-b-A-17	v3r	518-534	17	ATTACCGCGGCTGCTGG	65	(60)	
515F_v4f	S*-Univ-0515-b-S-19	v4f	515-533	19	GTGNCAGCMGCCCGGTAA	63	mod. after (61)	
806R_v4r	S-D-Bact-0756-a-A-20	v4r	787-806	20	GGACTACHVGGGTWCTAAT	40	mod. after (61)	
784F_v5f	S*-Univ-0779-a-S-15	v5f	784-798	15	RGGATTAGATACCCY	40	mod. after (62)	
926R_v5r	S-D-Bact-0907-b-A-20	v5r	907-926	20	CCGTCATTTTTRAGTTT	25	mod. after (63)	
907F_v6f	S-D-Bact-0907-a-S-20	v6f	907-926	20	AAACTYAAARRAATTGACCG	25	(64)	
1114R_v6r	S-D-Bact-1114-b-A-16	v6r	1099-1114	16	GGGTTCGCTCGTTRY	50	mod. after (62)	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	GAYTTGACRTCVTCM	38	new primer	
1185F_v8f	S-D-Bact-1185-a-S-16	v8f	1185-1200	16	KGGABCAACCGCYCGYC	63	new primer	
1407R_v8r	S-D-Bact-1407-a-A-16	v8r	1391-1407	16	GRCGRGCGGTGWTTC	63	mod. after (65)	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	GGNTACCTTGTACGACTT	42	mod. after (66)	S*-Univ-1492-a-A-21

manually curated databases are of paramount importance to improve reproducibility, speed during the bioinformatics process of microbial community studies and communication between researchers (54).

Here we present a comprehensive gold standard database, DAIRYdb (Database, Agroscope, Inra, Ribosomal, accuracy), for 16S OTUs classification from NGS data of dairy products. The main goal was to develop a dedicated database that allow researchers to accurately and automatically annotate short reads of 16S down to the species level. Manual curation of the database and its restriction to the biodiversity expected in dairy products strongly improves accuracy and reproducibility of phylogenetic classification to all taxonomic ranks. DAIRYdb is publicly available at <https://github.com/marcomeola/DAIRYdb> and can be integrated in any classification prediction tool that allows the integration of customized databases, such as Blast+, Metaxa2, SINTAX and FROGS.

## Results

**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% ANI from Silva, RDP and Greengenes (Figure 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 (73, 74), whose sequences were lost during the clustering, were added, resulting to the final number of 10'290 16S sequences.

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 keywords (Figure 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 (Figure 2B). Altogether, 1'700 sequences were associated to just one keyword, with most of the sequences shared by four keywords (Figure 2C).

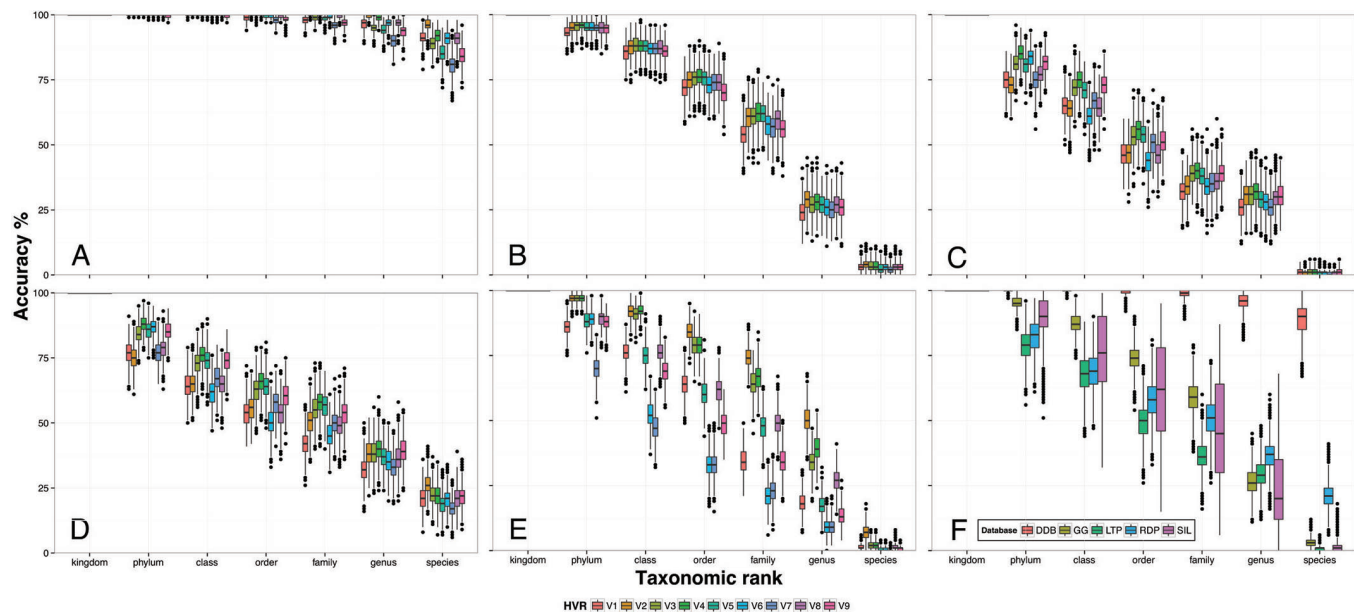
**Curation.** During the first step of data curation, the sequences were taxonomically annotated with Silva by means of SINA (75). The resulted annotation at all taxonomic ranks underwent a first manual check and cleaning for taxonomic inconsistencies through cross-comparison with the other members of 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 (54). A maximum of three closest neighbour type strains (CN) with authoritative taxonomy (CN) 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 (76).

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 (54). 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) (70) was used down to the species rank with the addition of the unclassified rank. Although not all OTUs in a microbiome study will be classified to the species rank, at least they will not all be merged to the same unclassified species rank, but taking over the LCR to differentiate between all unclassified OTUs. 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 biological uninformative groups, thus improving communication among scientists (53).

DAIRYdb version 1.1 contains 2 kingdoms (Bacteria and Archaea), 47 phyla, 136 classes, 249 orders, 463 families, 1757 genera and 4030 unique species-like groups/species complexes (Figure 3, Additional File 1 and Additional File 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%), *Crenarchaeota* (13%), *Thaumarchaeota* (9%), *Woeisearchaeota* (3%) and others (1%). Altogether, the DAIRYdb is able to capture the diversity of known taxa expected to occur in dairy products. Increasing number of whole genome sequences (WGS) will more likely lead to a replacement of 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 lactic acid bacteria (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



**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 (A), Greengenes (B), LTP (C), RDP (D) and Silva (E). Average performance of all HVR for each database (F) (accuracy = correctly annotated/total).

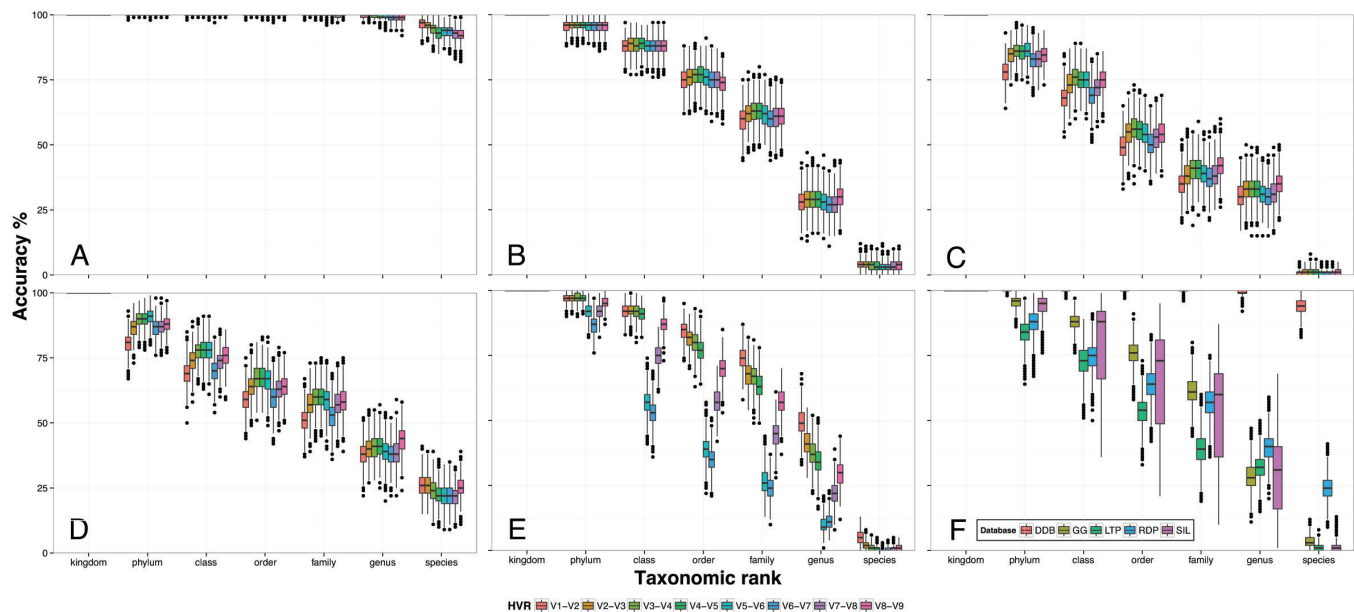
extensively studied and described, several unresolved controversies regarding the nomenclature of some keystone species still remain unsolved, such as for the species *L. helveticus* and *L. gallinarum*, *S. thermophilus* and *S. salivarius*, *L. casei* and *L. paracasei* and *L. zeae*, *L. plantarum* and *L. paraplantarum* (77–79). 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 BacDive (80). For the reference sequences of the most crucial species, it was tried to use bacterial names listed in the actual "List of prokaryotic names" according to BacDive, however, further disagreements between Silva and BacDive cannot be completely excluded. Moreover, it is also possible that some crucial genera in dairy products may undergo a radical genome-based relabelling in order to have more homogeneous clusters (79).

Different approaches were applied on inpure taxa, *i.e.* taxa that overlap in the tree despite being assigned to different nomenclature (54), by the universal databases. For instances, 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* (54). 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(54, 70).

The DAIRYdb will undergo regular updates in accordance to update on bacterial nomenclature (79), 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(68).

**Validation.** At present, only short read sequences can be obtained from the most common amplicon NGS sequencer with at least 99% quality and up to 600 bp in 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 from the DAIRYdb using two methods: V-Xtractor (81) (Figure 4A,C) or *in silico* PCR with mothur (36). V-Xtractor was used to evaluate the general annotation accuracy of all HVRs present in the DAIRYdb. The *in silico* PCR with universal primers (Table 1) highlighted the theoretical extraction efficiency of the primer pairs adapted to the pool of sequences in the DAIRYdb. While V-Xtractor extracted the HVRs, the *in silico* PCR also evaluated the theoretical extraction efficiency of different primer pairs. The ratio between the number of detected HVRs with V-Xtractor and HVRs extracted by *in silico* PCR determined the biodiversity coverage of the



**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 (A), Greengenes (B), LTP (C), RDP (D) and Silva (E). Average performance of all HVR for each database (F) (accuracy = correctly annotated/total).

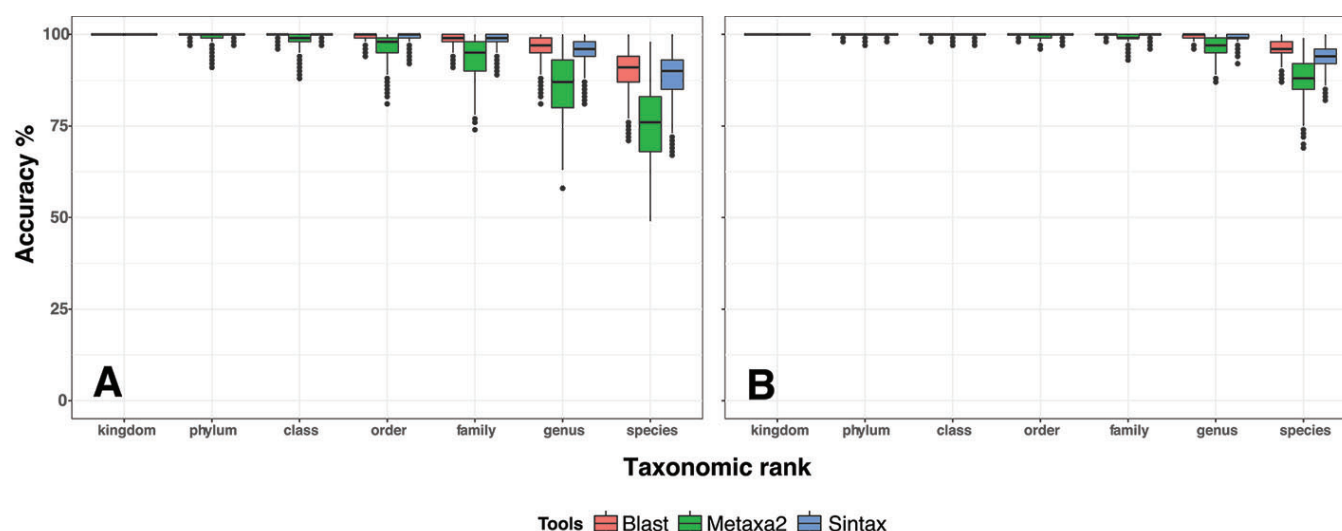
different HVRs achieved with the different primer pairs and potential biases in community structure in downstream analyses depending on the different HVR analysed (Figure 4B,D).

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 least 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 entirely or partial loss of V1 and V9 (Figure 4A, Table 1). The primer pairs targeting the V1, V2 and V9 were less efficient as compared to the primer pairs targeting V3 to V8. The primer pairs for V4 performed best with 90% coverage, followed by V7 (88%), V5 and V6 (85%), and V3 (83%). The same hold true for the HVR pairs, where the HVR pairs V1V2, V2V3 and V8V9 performed less well as compared to the central HVR pairs (Figure 4C, 1).

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 (Figure 4B,D). Percentage values of the single HVRs slightly increased while confirming the overall picture. The largest biodiversity coverage by the DAIRYdb was achieved by the single HVR V4 (92%), followed by the HVR pairs V5V6 and V3V4 (89%).

The taxonomy 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. The synthetic HVR fragments were extracted from 1000 subsamples of each 100 randomly selected sequences from the DAIRYdb by either V-Xtractor or *in silico* PCR and assigned to all taxonomic ranks by the means of three different classification predictors (Blast+, Metaxa2 and

SINTAX) using the aforementioned databases. Taxonomic annotation accuracy of the single HVR extracted with V-Xtractor with the DAIRYdb using SINTAX was above 75% at all taxonomic ranks (Figure 5). 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 greatest classification 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% (32). Overall, the universal databases were less accurate with decreasing taxonomic rank (Figure 5B-D). 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 (54). Different to the DAIRYdb, the HVR V4 performed best with the universal databases with exception to Silva, where V2 achieved a higher accuracy (Figure 5). Generally, the difference in classification 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 (Figure 5E). All HVRs taken together, the DAIRYdb achieved a significantly better taxonomy annotation accuracy of average  $88.9\% \pm 5.5$  as compared to the universal databases tested, particularly at order to species ranks (Figure 5F). The annotation accuracy results with Blast+ and Metaxa2 of single HVR extracted with *in silico* PCR (Additional File 3, Figures 1, 3 and 7), V-Xtractor (Additional File 3, Figures 5 and 9), and HVR pairs with *in silico* PCR (Additional File 3, Figures 2, 4 and 8), V-Xtractor (Additional File 3, Figures 6 and 10) are available in Additional File 3.



**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.

The results with the HVR pairs was similar to the single HVRs (Figure 6). Classification confidence 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 classification accuracy at species rank in the DAIRYdb, as well as with RDP and Silva. 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 (24). Greengenes species annotation accuracy was similar for all HVRs, while LTP showed very low performance at species rank. The average accuracy value for correct species annotation of all HVR pairs with the DAIRYdb was over  $94\% \pm 2.8$  (Figure 6F). Only species annotation with the RDP trainset v16 achieved 25% of correct annotations. The BLAST16S database was shown to obtain genus accuracies  $\sim 50\%$  for V4, which improves with increasing length to  $\sim 60\%$  with V3–V5 and  $\sim 70\%$  with full-length 16S (70). As expected, the increasing number of HVR increases the confidence in taxonomy annotation.

Taxonomy annotation accuracy varied only little between different taxonomy predictors with the DAIRYdb and not significantly with either both, single HVR (Additional File 3, Figure 11) or HVR pairs and (Additional File 3, Figure 12). In fact, classification annotation accuracy performance varied more dependent on the database rather than the classification predictor. These results indicate that annotation of the members of the bacterial community is primarily influenced by the selection of the database, by the HVR, and only then by the taxonomy predictor (Additional File 3, Figures 1–10). 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 (Figure 7). Among all tools, Blast+ and SINTAX were slightly yet not significantly 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 not surprising. Moreover, Metaxa2 performance is strongly dependent on the average nucleotide identity (ANI) thresholds used, which were set according to (82). On the other hand, the more stringent parameters of Metaxa2 reduce 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>).

**Outlook.** The advances of genomics in microbiology has led to a reassessment of the phylogeny, which still remains a moving target particularly for microbial taxonomy (53, 54). The complexity of microbial taxonomy was already reflected in the statement by S. T. Cowan saying that "Taxonomy is the most subjective branch of any biological discipline, and in many ways is more an art than a science" (52, 83). As of 2014, the number of species of prokaryotes with validly published names was about 11'000 (84). However, microbial systematics is failing in its fundamental mission to precisely provide the ecological properties of an organism that is classified to a species (85).

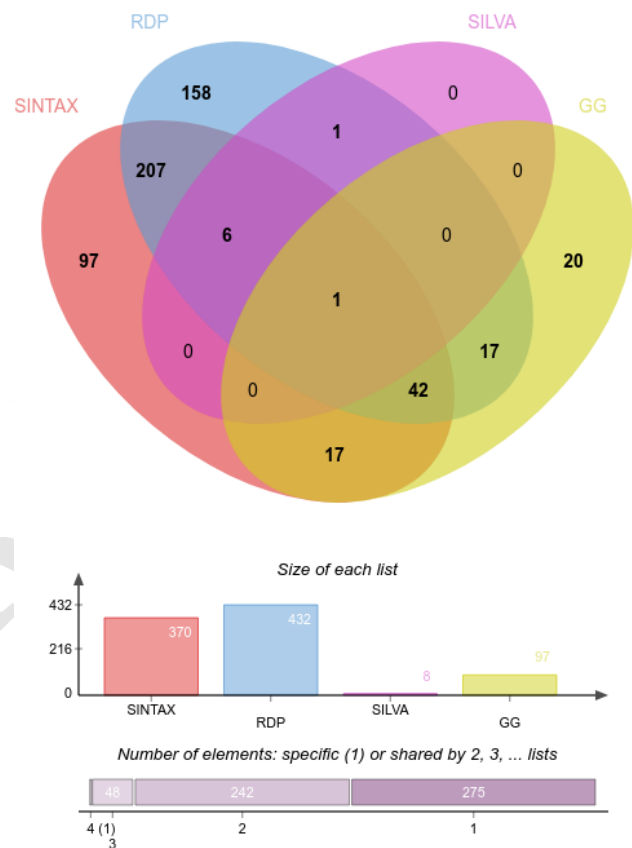
The correct definition of a bacterial community structure remains a bioinformatic challenge, where any parameter, from wet-lab (*i.e.*, DNA extraction, primer and HVR selection, amplification, sequencing) to the bioinformatic pipeline, can influence the outcome. The results of microbiome

studies are most strongly influenced by the selection of the primer pairs and thereof of the HVR amplified, rather than on the sequencing technology used for the study (86–88). The OTU-picking algorithm dependent on the sequencing technology (clustering vs. denoising) or ASVs instead (89), the classification predictor are of secondary importance, although their impact on the outcome is not negligible (67). The selection of the primer pairs should be made after careful consideration of their coverage in diversity with respect to the studied environment (67). 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.

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 another environments, such as the human gut. Classification accuracy performed on sequences from type strains included in the Human Intestinal Tract database (HITdb) showed that the DAIRYdb performed comparably well to the RDP trainset v16 and significantly better than Silva and Greengenes (Figure 8) (69). Yet, the way and ability to recognize the basic unit for taxonomy of prokaryotes depends on the resolution power of the observational methods actually available (52). 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* (90). Although a separate full species status was proposed (91), persistent contention prevented a full ratification by the taxonomic committees (90). Increasing sequence read lengths will make it possible to cover three HVRs or even the entire 16S, thus significantly improving taxonomic annotation accuracy at species rank by using a manually curated database like the DAIRYdb.

Although it can be considered a significant progress to obtain over 90% of accurate species classification based on short 16S fragments, quality of dairy products is often influenced by different strains of the same species. 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 (85), on the other hand, the intraspecies variability of the 16S lacks sufficient resolution to clearly discern between strains and subspecies within the

same species (92). Nevertheless, recent powerful bioinformatics tools, such as Oligotyping (93) or Minimal Entropy Decomposition (MED) (94), can be applied to distinguish between ecologically relevant amplicon sequence variants (ASVs) within OTUs assigned to a same species. 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 (85, 89, 95).



**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 (69). On sequences from origin other than dairy products, the DAIRYdb performs significantly better than Silva and Greengenes, but not better than the RDP trainset v16.

## 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.

While universal 16S databases, such as Silva, RDP and Greengenes cover a broad biodiversity allowing to capture the maximal biodiversity available in a system, the enormous number of sequences in those databases lead to conflicting taxonomy annotation at genus and species ranks and ambiguous annotations or blanks due to competing sequences increase accordingly (69). Moreover, the size of the database can be a deterrent for researchers to improve the quality of taxonomic annotation of the sequences therein. Most of the detected OTUs in NGS analyses diverge from authoritative reference sequences from type strains either due to sequencing biases or missing cultivated representative strains. Beside reference sequences, many environmental sequences are annotated by the universal databases Silva, RDP and Greengenes based on different taxonomic classification standards, e.g., Bergey's Manual, the List of Prokaryotic Names with Standing Nomenclature (LPSN), International Nucleotide Sequence Database Collaboration (INSDC). These different curation strategies lead to annotation conflicts between the databases and disagreement between microbiome studies, which are not biologically explained, rather a consequence of the database used for annotation.

Different to available universal databases, DAIRYdb achieved correct taxonomy annotation for ~90% of species names on single HVRs and HVR pairs with sequences present in dairy samples (70). In fact, the DAIRYdb significantly reduced conflicting miss-annotated sequences and facilitated manual curation, while covering the inspected biodiversity. 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 (96). 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 (54, 68, 69). 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 (54, 68).

A certainly valid argument against manually curated databases is their lack of reproducibility (54). However, annotation accuracy achieved with the DAIRYdb significantly outperformed all universal databases tested here, as well as the RDP trainset v16, which was shown to have the best performance among the universal databases (54). The training sets of the universal databases contained sequences with missing taxonomic labels at uncertain classifications at any taxonomic rank with increasing number of blanks at species levels (54). The consequences are numerous unclassified OTUs with no biological meaning.

We therefore propose the manually curated DAIRYdb as

a gold standard 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 (67, 97).

## 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 Midi-Pyrenees (Bioinfo Genotoul) and INRA MIGALE bioinformatics platform (<http://migale.jouy.inra.fr>) for providing computing and storage resources.

## Bibliography

1. Teresia M Porter and Mehrdad Hajibabaei. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol Ecol*, 27(2):313–338, Jan 2018. doi: 10.1111/mec.14478.
2. Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Lacey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evgenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingling Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciółek, Nicholas A. Bokulich, Joshua Leffler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrod Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, Rob Knight, Jose L. Agosto Rivera, Lisa Al-Moosawi, John Alverdy, Katherine R. Amato, Jason Andras, Largus T. Angenent, Dionysios A. Antonopoulos, Amy Apprill, David Armitage, Kate Ballantine, Jiri Bárta, Julia K. Baum, Alison Berry, Ashish Bhatnagar, Monica Bhatnagar, Jennifer F. Biddle, Lucie Bittner, Bazartseren Boldgiv, Eric Bottos, Donal M. Boyer, Josephine Braun, William Brazelton, Francis Q. Brearley, Alexandra H. Campbell, J. Gregory Caporaso, Cesar Cardona, Jo Lynn Carroll, S. Craig Cary, Brenda B. Casper, Trevor C. Charles, Haiyan Chu, Danielle C. Claar, Robert G. Clark, Jonathan B. Clayton, Jose C. Clemente, Alyssa Cochran, Maureen L. Coleman, Gavin Collins, Rita R. Colwell, Mónica Contreras, Benjamin B. Crary, Simon Creer, Daniel A. Cristol, Byron C. Crump, Duoying Cui, Sarah E. Daly, Liliana Davalos, Russell D. Dawson, Jennifer Defazio, Frédéric Delsuc, Hebe M. Dionisi, Maria Gloria Dominguez-Bello, Robin Dowell, Eric A. Dubinsky, Peter O. Dunn, Danilo Ercolini, Robert E. Espinoza, Vanessa Ezenwa, Nathalie Fenner, Helen S. Findlay, Irma D. Fleming, Vincenzo Fogliano, Anna Forsman, Chris Freeman, Kristina Guyton, Sarah Jane Haig, Vanessa Hale, Maria Alexandra Garcia-Amado, David Garshelis, Robin B. Gasser, Gunnar Gerdt, Molly K. Gibson, Isaac Gifford, Ryan T. Gill, Tugrul Giray, Antje Gittel, Peter Golyshtin, Donglai Gong, Hans Peter Grossart, Kristina Guyton, Sarah Jane Haig, Vanessa Hale, Ross Stephen Hall, Steven J. Hallam, Kim M. Handley, Nur A. Hasan, Shane R. Haydon, Jonathan E. Hickman, Glida Hidalgo, Kirsten S. Hofmøckel, Jeff Hooker, Stefan Hulth, Jenni Hultman, Embriette Hyde, Juan Diego Ibáñez-Álamo, Julie D. Jastrow, Aaron R. Jex, L. Scott Johnson, Eric R. Johnston, Stephen Joseph, Stephanie D. Jurburg, Diogo Jurelevicius, Anders Karlsson, Roger Karlsson, Seth Kauppinen, Colleen T.E. Kellogg, Suzanne J. Kennedy, Lee J. Kerkhof, Gary M. King, George W. Klint, Anson V. Koehler, Monika Krezalek, Jordan Kuennenman, Regina Lamendella, Emily M. Landon, Kelly Lanede Graaf, Julie LaRoche, Peter Larsen, Bonnie Laverock, Simon Lax, Miguel Lentino, Iris I. Levin, Pierre Liancourt, Wenju Liang, Alexandra M. Linz, David A. Lipson, Yongqin Liu, Manuel E. Magris, Mariana Lozada, Catherine M. Spirito, Walter P. McCormack, Aurora MacRae-Crerar, Magda Magris, Antonio M. Martin-Platero, Manuel Martin-Vivaldi, L. Margarita Martinez, Manuel Martinez-Bueno, Ezequiel M. Marzinelli, Olivia U. Mason, Gregory D. Mayer, Jamie M. McDavitt-Irwin, James E. McDonald, Krista L. McGuire, Katherine D. McMahon, Ryan McMinds, Mónica Medina, Joseph R. Mendelson, Jessica L. Metcalf, Folker Meyer, Fabian Micheloni, Kim Miller, David A. Mills, Jeremiah Minich, Stefano Mocali, Lucas Moitinho-Silva, Anni Moore, Rachael M. Morgan-Kiss, Paul Munroe, David Myrold, Josh D. Neufeld, Yingying Ni, Graeme W. Nicol, Shaun Nielsen, Jozef I. Nissimov, Kefeng Niu, Matthew J. Nolan, Karen Noyce, Sarah L. O'Brien, Noriko Okamoto, Ludovic Orlandu, Yádra Ortiz Castellano, Olayinka Osulale, Wyatt Oswald, Jacob Parnell, Juan M. Peralta-Sánchez, Peter Petraitis, Catherine Pfister, Elizabeth Pilon-Smits, Paola Piombino, Stephen B. Pointing, F. Joseph Pollock, Caitlin Potter, Bharath Prithiviraj, Christopher Quince, Asha Rani, Ravi Ranjan, Subramanya Rao, Andrew P. Rees, Miles Richardson, Ulf Riebesell, Carol Robinson, Karl J. Rockne, Selena Marie Rodriguez, Forest Rohwer, Wayne Roundstone, Rebecca J. Safran, Naseer Sangwan, Virginia Sanz, Matthew Schrenk, Mark D. Schrenzel, Nicole M. Scott, Rita L. Seger, Andaine Seguinorlando, Lucy Seldin, Lauren M. Seyler, Baddr Shakhsher, Gabriela M. Sheets, Congcong Shen, Yu Shi, Hakdong Shin, Benjamin D. Shogan, Dave Shutter, Jeffrey Siegel, Steve Simmons, Sara Sjöling, Daniel P. Smith, Juan J. Soler, Martin Sperling, Peter D. Steinberg, Brent Stephens, Melita A. Stevens, Safiyh Taghavi, Vera Tai, Karen Tait, Chia L. Tan, Neslihan Taş, D. Lee Taylor, Torsten Thomas, Ina Timling, Benjamin L. Turner, Tim Ulrich, Luke K. Ursell, Daniel Van Der Lelie, William Van Treuren, Lukas Van Zwieten, Daniela Vargas-Robles, Rebecca Vega Thurber, Paola Vitaglione, Donald A. Walker, William A. t. Shi Wang, Tao Wang, Tom Weaver, Nicole S. Webster, Beck Wehrle, Pamela Weisenhorn, Sophie Weiss, Jeffrey J. Werner, Kristin West, Andrew Whitehead, Susan R. Whitehead, Linda A. Whittingham, Eske Willerslev, Allison E. Williams, Stephen A. Wood, Douglas C. Woodhams, Yeqin Yang, Jesse Zaneveld, Irtaze Zarraonandia, Qikun Zhang, and Hongxia Zhao. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017. ISSN 14746687. doi: 10.1038/nature24621.

3. Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesc D'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kozoska, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Ryo-Llonch, Hugo Sarmiento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weisenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Ocean plankton. Structure and function of the global ocean microbiome. *Science* (New York, N.Y.), 348(6237):1261359, 2015. ISSN 1095-9203. doi: 10.1126/science.1261359.
4. Mary Ann Moran. The global ocean microbiome. *Science*, 350(6266), 2015. ISSN 0036-8075. doi: 10.1126/science.aac8455.
5. Marco Meola, Anna Lazzaro, and Josef Zeyer. Bacterial composition and survival on Sahara dust particles transported to the European Alps. *Frontiers in Microbiology*, 6(DEC):1–17, 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.01454.
6. David A. Pearce, K. A. Hughes, T. Lachlan-Cope, S. A. Harangozo, and A. E. Jones. Biodiversity of air-borne microorganisms at Halley station, Antarctica. *Extremophiles*, 14(2): 145–159, 2010. ISSN 14310651. doi: 10.1007/s00792-009-0293-8.
7. A Lazzaro, A Wismer, M Schneebeli, I Erny, and J Zeyer. Microbial abundance and community structure in a melting alpine snowpack. *Extremophiles*, 19(3):631–642, 2015. ISSN 1433-4909 (Electronic) 1431-0651 (Linking). doi: 10.1007/s00792-015-0744-3.
8. Brent C. Christner, Mark L. Skidmore, John C. Priscu, Martyn Tranter, and Christine M. Foreman. Bacteria in subglacial environments. *Psychrophiles: From Biodiversity to Biotechnology*, pages 51–71, 2008.
9. Bhagya. R. Yeluri Jonnala, Paul L. H. McSweeney, Jeremiah J. Sheehan, and Paul D. Cotter. Sequencing of the cheese microbiome and its relevance to industry. *Frontiers in Microbiology*, 9:1020, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.01020.
10. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–21, Jun 2012. doi: 10.1038/nature11209.
11. Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nat Rev Genet*, 13(4):260–70, Mar 2012. doi: 10.1038/nrg3182.
12. Allyson L Byrd, Yasmine Belkaid, and Julia A Segre. The human skin microbiome. *Nat Rev Microbiol*, 16(3):143–155, Mar 2018. doi: 10.1038/nrmicro.2017.157.
13. N Fierer, D Nemergut, R Knight, and J M Craine. Changes through time: integrating microorganisms into the study of succession. *Res Microbiol*, 161(8):635–642, 2010. ISSN 1769-7123 (Electronic) 0923-2508 (Linking). doi: 10.1016/j.resmic.2010.06.002.
14. Marco Meola, Anna Lazzaro, and Josef Zeyer. Diversity, resistance and resilience of the bacterial communities at two alpine glacier forefields after a reciprocal soil transplantation. *Environmental microbiology*, 16(6):1918–1934, 2014. ISSN 14622920. doi: 10.1111/1462-2920.12435.
15. Ashley Shade, J Gregory Caporaso, Jo Handelsman, Rob Knight, and Noah Fierer. A meta-analysis of changes in bacterial and archaeal communities with time. *The ISME Journal*, 7(8):1493–1506, 2013. ISSN 1751-7362. doi: 10.1038/ismej.2013.54.
16. C R Woese and G E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–90, Nov 1977.
17. G E Fox, E Stackebrandt, R B Hespell, J Gibson, J Maniloff, T A Dyer, R S Wolfe, W E Balch, R S Tanner, L J Magrum, L B Zablen, R Blakemore, R Gupta, L Bonen, B J Lewis, D A Stahl, K R Luehrs, K N Chen, and C R Woese. The phylogeny of prokaryotes. *Science*, 209(4455):457–63, Jul 1980.
18. N. R. Pace. A Molecular View of Microbial Diversity and the Biosphere. *Science*, 276(5313): 734–740, 1997. ISSN 00368075. doi: 10.1126/science.276.5313.734.
19. Anna Klindworth, Elmar Pruesse, Timmy Schwaer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 41(1):1–11, 2013. ISSN 03051048. doi: 10.1093/nar/gks808.
20. Matteo Ramazzotti and Giovanni Bacci. Chapter 5 - 16s rRNA-based taxonomy profiling in the metagenomics era. In Muniraj Nagarajan, editor, *Metagenomics*, pages 103 – 119. Academic Press, 2018. ISBN 978-0-08-102268-9. doi: <https://doi.org/10.1016/B978-0-08-102268-9.00005-7>.
21. Richard Christen. Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes Environ*, 23(4):253–68, 2008.
22. Hilde Vinje, Kristian Hovde Liland, Trygve Almøy, and Lars Snipen. Comparing K-mer based methods for improved classification of 16S sequences. *BMC Bioinformatics*, 16(1):1–13, 2015. ISSN 14712105. doi: 10.1186/s12859-015-0647-4.
23. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990. ISSN 0022-2836. doi: [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
24. Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Na?ve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007. ISSN 00992240. doi: 10.1128/AEM.00062-07.
25. Susan M. Huse, Les Dethlefsen, Julie A. Huber, David Mark Welch, David A. Relman, and Mitchell L. Sogin. Exploring microbial diversity and taxonomy using ssu rRNA hypervariable tag sequencing. *PLOS Genetics*, 4(11):1–10, 11 2008. doi: 10.1371/journal.pgen.1000255.
26. Suparna Mitra, Mario Stårk, and Daniel H. Huson. Analysis of 16s rRNA environmental sequences using megan. *BMC Genomics*, 12(3):S17, Nov 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-S3-S17.
27. Johan Bengtsson-Palme, Martin Hartmann, Karl Martin Eriksson, Chandan Pal, Kaisa Thorell, Dan Göran Joakim Larsson, and Rolf Henrik Nilsson. Metax2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour*, 15(6):1403–14, Nov 2015. doi: 10.1111/1755-0998.12399.
28. Matteo Ramazzotti, Luisa Berni, Claudio Donati, and Duccio Cavalieri. riboframe: An improved method for microbial taxonomy profiling from non-targeted metagenomics. *Frontiers in Genetics*, 6:329, 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00329.
29. Guy Allard, Feargal J. Ryan, Ian B. Jeffery, and Marcus J. Claesson. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, 16(1):324, 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0747-1.
30. Panu Somervuo, Sonja Koskela, Juho Pennanen, R. Henrik Nilsson, and Otsa Ovaskainen. Unbiased probabilistic taxonomic classification for dna barcoding. *Bioinformatics*, 32(19): 2920–2927, 2016. doi: 10.1093/bioinformatics/btw346.
31. Robert Edgar. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, page 074161, 2016. doi: 10.1101/074161.
32. Mohamed Mysara, Peter Vandamme, Ruben Props, Frederiek-Maarten Kerckhof, Natalie Leys, Nico Boon, Jeroen Raes, and Pieter Monsieurs. Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiology Ecology*, 93(4):fix029, 2017. doi: 10.1093/femsec/fix029.
33. Douglas J. Sherman. Humidor : Microbial community classification of the 16 s gene by training cigar strings with convolutional neural networks. 2017.
34. João F Matias Rodrigues, Thomas S B Schmidt, Janko Tackmann, and Christian von Mering. Mapseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics*, 33(23):3808–3810, 2017. doi: 10.1093/bioinformatics/btx517.
35. Kristian Hovde Liland, Hilde Vinje, and Lars Snipen. microclass: an R-package for 16S taxonomy classification. *BMC Bioinformatics*, 18(1):172, 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1583-2.
36. Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres, Gerhard G. Thallinger, David J. Van Horn, and Carolyn F. Weber. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009. ISSN 00992240. doi: 10.1128/AEM.01541-09.
37. Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, 7(5): 335–6, May 2010. doi: 10.1038/nmeth.f303.
38. Frédéric Escudé, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, and Géraldine Pascal. Frogs: Find, rapidly, otus with galaxy solution. *Bioinformatics*, 34(8):1287–1294, 2018. doi: 10.1093/bioinformatics/btx791.
39. Alejandra Escobar-Zepeda, Arturo Vera Ponce De León, and Alejandro Sanchez-Flores. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, 6(DEC):1–15, 2015. ISSN 16648021. doi: 10.3389/fgene.2015.00348.
40. C Camacho, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, and T L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 10.1186/1471-2105-10-421.
41. Johan Bengtsson-Palme, Rodney T Richardson, Marco Meola, Christian Wurzbacher, Émile D Tremblay, Kaisa Thorell, Kärt Kanger, K Martin Eriksson, Guillaume J Bilodeau, Reed M Johnson, Martin Hartmann, and R Henrik Nilsson. Metax2 database builder: Enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. *Bioinformatics*, page bty482, 2018. doi: 10.1093/bioinformatics/bty482.
42. Monika Balvočiūtė and Daniel H. Huson. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics*, 18(S2):114, 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3501-4.
43. Frank Oliver Glöckner, Pelin Yilmaz, Christian Quast, Jan Gerken, Alan Beccati, Andreea Ciuprina, Gerrit Bruns, Pablo Yarza, Jörg Peplies, Ralf Westram, and Wolfgang Ludwig. 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology*, (February):0–1, 2017. ISSN 18734863. doi: 10.1016/j.jbiotec.2017.06.1198.
44. Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196, 2007. ISSN 03051048. doi: 10.1093/nar/gkm864.
45. C Quast, E Pruesse, P Yilmaz, J Gerken, T Schwaer, P Yarza, J Peplies, and F O Glockner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, 41(Database issue):D590–6, 2013. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gks1219.
46. Aidan C Parte. Lpsn—list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res*, 42(Database issue):D613–6, Jan 2014. doi: 10.1093/nar/gkt1111.
47. Pelin Yilmaz, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schwaer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. The SILVA and “all-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research*, 42(D1):643–648, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1209.
48. J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kalam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(SUPPL. 1):141–145, 2009. ISSN 03051048. doi: 10.1093/nar/gkn879.
49. Guy Cochrane, Ilene Karsch-Mizrachi, Toshihisa Takagi, and International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. *Nucleic Acids Res*, 44(D1):D48–50, Jan 2016. doi: 10.1093/nar/gkv1323.
50. T Z DeSantis, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 72(7):5069–5072, 2006. ISSN 0099-2240 (Print) 0099-2240 (Linking). doi: 10.1128/AEM.03006-05.
51. Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetverinov, Deanna M. Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M. Fingerman, Lewis Y. Geer, Wolfgang Helmsberg, Yuri Kapustin, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrahi, James Ostell, Anna Panchenko, Lon Phan, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Yanli Wang, W. John Wilbur, Eugene Yashchenko, and

- Jian Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, 2011. doi: 10.1093/nar/gkq1172.
52. Ramon Rosselló-Móra. Towards a taxonomy of bacteria and archaea based on interactive and cumulative data repositories. *Environ Microbiol*, 14(2):318–34, Feb 2012. doi: 10.1111/j.1462-2920.2011.02599.x.
53. Konstantinos T. Konstantinidis, Ramon Rosselló-Móra, and Rudolf Amann. Uncultivated microbes in need of their own taxonomy. *ISME Journal*, 11(11):2399–2406, 2017. ISSN 17517370. doi: 10.1038/ismej.2017.113.
54. Robert Edgar. Taxonomy annotation and guide tree errors in 16s rna databases. *PeerJ*, 6:e5030, June 2018. ISSN 2167-8359. doi: 10.7717/peerj.5030.
55. Mark Yandell and Daniel Ence. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*, 13(5):329–42, Apr 2012. doi: 10.1038/nrg3174.
56. Yuichi Hongoh, Hiroe Yuzawa, Moriya Ohkuma, and Toshiaki Kudo. Evaluation of primers and pcr conditions for the analysis of 16s rna genes from a natural environment. *FEMS Microbiol Lett*, 221(2):299–304, Apr 2003.
57. Andreas Sundquist, Saharnaz Bigdeli, Roxana Jalili, Maurice L Druzin, Sarah Waller, Kristin M Pullen, Yasser Y El-Sayed, M Mark Taslimi, Serafim Batzoglou, and Mostafa Ronaghi. Bacterial flora-typing with targeted, chip-based pyrosequencing. *BMC Microbiol*, 7: 108, Nov 2007. doi: 10.1186/1471-2180-7-108.
58. R I Amann, B J Binder, R J Olson, S W Chisholm, R Devereux, and D A Stahl. Combination of 16s rna-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol*, 56(6):1919–25, Jun 1990.
59. S el Fantroussi, L Verschuere, W Verstraete, and E M Top. Effect of phenylurea herbicides on soil microbial communities estimated by analysis of 16s rna gene fingerprints and community-level physiological profiles. *Appl Environ Microbiol*, 65(3):982–8, Mar 1999.
60. G Muzer, E C de Waal, and A G Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16s rna. *Appl Environ Microbiol*, 59(3):695–700, Mar 1993.
61. William A Walters, J Gregory Caporaso, Christian L Lauber, Donna Berg-Lyons, Noah Fierer, and Rob Knight. Primer prospector: de novo design and taxonomic analysis of bar-coded polymerase chain reaction primers. *Bioinformatics*, 27(8):1159–61, Apr 2011. doi: 10.1093/bioinformatics/btr087.
62. Carlos W Nossa, William E Oberdorf, Liying Yang, Jørn A Aas, Bruce J Paster, Todd Z Desantis, Eoin L Brodie, Daniel Malamud, Michael A Poles, and Zhiheng Pei. Design of 16s rna gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol*, 16(33):4135–44, Sep 2010.
63. W T Liu, T L Marsh, H Cheng, and L J Forney. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16s rna. *Appl Environ Microbiol*, 63(11):4516–22, Nov 1997.
64. B J F Keijser, E Zaura, S M Huse, J M B M van der Vossen, F H J Schuren, R C Montijn, J M ten Cate, and W Crielaard. Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res*, 87(11):1016–20, Nov 2008. doi: 10.1177/154405910808701104.
65. Jeffrey J Walker and Norman R Pace. Phylogenetic composition of rocky mountain endolithic microbial ecosystems. *Appl Environ Microbiol*, 73(11):3497–504, Jun 2007. doi: 10.1128/AEM.02656-06.
66. G.C. Baker, J.J. Smith, and D.A. Cowan. Review and re-analysis of domain-specific 16s primers. *Journal of Microbiological Methods*, 55(3):541 – 555, 2003. ISSN 0167-7012. doi: <https://doi.org/10.1016/j.mimet.2003.08.009>.
67. Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. The madness of microbiome: Attempting to find consensus "best practice" for 16s microbiome studies. *Appl Environ Microbiol*, 84(7), Apr 2018. doi: 10.1128/AEM.02627-17.
68. Irene LG Newton and Guus Roeselers. The effect of training set on the classification of honey bee gut microbiota using the naïve bayesian classifier. *BMC Microbiology*, 12(1): 221, Sep 2012. ISSN 1471-2180. doi: 10.1186/1471-2180-12-221.
69. Jarmo Ritari, Jarkko Salojärvi, Leo Lahti, and Willem M. de Vos. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics*, 16(1):1056, 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-2265-y.
70. Robert C. Edgar. Accuracy of taxonomy prediction for 16s rna and fungal its sequences. *PeerJ*, 6:e4652, April 2018. ISSN 2167-8359. doi: 10.7717/peerj.4652.
71. Jody Hey. The mind of the species problem. *Trends in Ecology & Evolution*, 16(7):326 – 329, 2001. ISSN 0169-5347. doi: [https://doi.org/10.1016/S0169-5347\(01\)02145-0](https://doi.org/10.1016/S0169-5347(01)02145-0).
72. Ramon Rosselló-Móra. *DNA-DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation*, pages 23–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-31292-5.
73. Marie Christine Montel, Solange Buchin, Adrien Mallet, Céline Delbes-Paus, Dominique A. Vuitton, Nathalie Desmasures, and Françoise Berthier. Traditional cheeses: Rich and diverse microbiota with associated benefits. *International Journal of Food Microbiology*, 177 (May):136–154, 2014.
74. Françoise Irlinger, Séverine Layec, Sandra Hélinck, and Eric Dugat-Bony. Cheese rind microbial communities: diversity, composition and origin. *FEMS Microbiology Letters*, 362 (2):1–11, 2015. doi: 10.1093/femsle/fnu015.
75. Elmar Pruesse, Jörg Peplies, and Frank Oliver Glöckner. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts252.
76. Alexey M. Kozlov, Jiajie Zhang, Pelin Yilmaz, Frank Oliver Glöckner, and Alexandros Stamatakis. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, 44(11):5022–5033, 2016. ISSN 13624962. doi: 10.1093/nar/gkw396.
77. Elisa Salvetti, Sandra Torriani, and Giovanna E. Felis. The genus lactobacillus: A taxonomic update. *Probiotics and Antimicrobial Proteins*, 4(4):217–226, Dec 2012. ISSN 1867-1314. doi: 10.1007/s12602-012-9117-8.
78. Sander Wuyts, Stijn Wittouck, Ilke De Boeck, Camille N Allonsius, Edoardo Pasolli, Nicola Segata, and Sarah Lebeer. Large-scale phylogenomics of the lactobacillus casei group highlights taxonomic inconsistencies and reveals novel clade-associated features. *mSystems*, 2(4), 2017. doi: 10.1128/mSystems.00061-17.
79. Elisa Salvetti, Hugh M. B. Harris, Giovanna E. Felis, and Paul W. O'Toole. Comparative genomics reveals robust phylogroups in the genus lactobacillus as the basis for reclassification. *Applied and Environmental Microbiology*, 2018. doi: 10.1128/AEM.00993-18.
80. The bacterial diversity metadatabase.
81. Martin Hartmann, Charles G. Howes, Kessy Abarenkov, William W. Mohn, and R. Henrik Nilsson. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, 83(2):250–253, 2010. ISSN 01677012. doi: 10.1016/j.mimet.2010.08.008.
82. Pablo Yarla, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, 2014. ISSN 1740-1526. doi: 10.1038/nrmicro3330.
83. S T Cowan. Sense and nonsense in bacterial taxonomy. *Journal of General Microbiology*, 67:1–8, 1971.
84. Mincheol Kim, Hyun-Seok Oh, Sang-Cheol Park, and Jongsik Chun. Towards a taxonomic coherence between average nucleotide identity and 16s rna gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 64(2):346–351, 2014.
85. Alexander Koepfel, Elizabeth B Perry, Johannes Sikorski, Danny Krizanc, Andrew Warner, David M Ward, Alejandro P Rooney, Evelynne Brambila, Nora Connor, Rodney M Ratcliff, Eviatar Nevo, and Frederick M Cohan. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci U S A*, 105(7):2504–9, Feb 2008. doi: 10.1073/pnas.0712205105.
86. Fiona Fouhy, Adam G. Clooney, Catherine Stanton, Marcus J. Claessens, and Paul D. Cotter. 16s rna gene sequencing of mock microbial populations- impact of dna extraction method, primer choice and sequencing platform. *BMC Microbiology*, 16(1):123, Jun 2016. ISSN 1471-2180. doi: 10.1186/s12866-016-0738-z.
87. Noha Youssef, Cody S. Sheik, Lee R. Krumholz, Fares Z. Najjar, Bruce A. Roe, and Mostafa S. Elshahed. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rna gene-based environmental surveys. *Applied and Environmental Microbiology*, 75(16):5227–5236, 2009. doi: 10.1128/AEM.00592-09.
88. Patrick D. Schloss and Sarah L. Westcott. Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rna gene sequence analysis. *Applied and Environmental Microbiology*, 77(10):3219–3226, 2011. doi: 10.1128/AEM.02810-10.
89. Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, 11(12): 2639–2643, 12 2017. doi: 10.1038/ismej.2017.119.
90. J.P. Burton, R.M. Chanyi, and M. Schultz. Chapter 19 - common organisms and probiotics: Streptococcus thermophilus (streptococcus salivarius subsp. thermophilus). In Martin H. Floch, Yehuda Ringel, and W. Allan Walker, editors, *The Microbiota in Gastrointestinal Pathophysiology*, pages 165 – 169. Academic Press, Boston, 2017. ISBN 978-0-12-804024-9. doi: <https://doi.org/10.1016/B978-0-12-804024-9.00019-7>.
91. Karl Heinz Schleifer, Mathias Ehrmann, Uli Krusch, and Horst Neve. Revival of the species streptococcus thermophilus (ex orla-jensen, 1919) nom. rev. *Systematic and Applied Microbiology*, 14(4):386 – 388, 1991. ISSN 0723-2020. doi: [https://doi.org/10.1016/S0723-2020\(11\)80314-0](https://doi.org/10.1016/S0723-2020(11)80314-0).
92. E. Stackebrandt and J. Ebers. Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today*, 8:6–9, 2006.
93. Eren A. Murat, Maignien Lois, Sul Woo Jun, Murphy Leslie G., Grim Sharon L., Morrison Hilary G., and Sogin Mitchell L. Oligotyping: differentiating between closely related microbial taxa using 16s rna gene data. *Methods in Ecology and Evolution*, 4(12):1111–1119, 2013. doi: 10.1111/2041-210X.12114.
94. A Murat Eren, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis, and Mitchell L Sogin. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J*, 9(4):968–79, Mar 2015. doi: 10.1038/ismej.2014.195.
95. Michelle A. Berry, Jeffrey D. White, Timothy W. Davis, Sunit Jain, Thomas H. Johengen, Gregory J. Dick, Orlando Sarnelle, and Vincent J. Denef. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in Freshwater lakes. *Frontiers in Microbiology*, 8(MAR):1–7, 2017. ISSN 1664302X. doi: 10.3389/fmicb.2017.00365.
96. Jeffrey J Werner, Omry Koren, Philip Hugenholtz, Todd Z DeSantis, William A Walters, J Gregory Caporaso, LARGUS T Angenent, Rob Knight, and Ruth E Ley. Impact of training sets on classification of high-throughput bacterial 16s rna gene surveys. *ISME J*, 6(1): 94–103, Jan 2012. doi: 10.1038/ismej.2011.82.
97. Overcoming hurdles in sharing microbiome data. *Nat Microbiol*, 2(12):1573, Dec 2017. doi: 10.1038/s41564-017-0077-3.