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

Marco Meola, Etienne Rifa, Noam Shani, Céline Delbès, Hélène Berthoud, Christophe Chassard

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

Marco Meola[a*] , Etienne Rifa[b*], Noam Shani[a], Céline Delbès[b], Hélène Berthoud[a], Christophe Chassard[b]

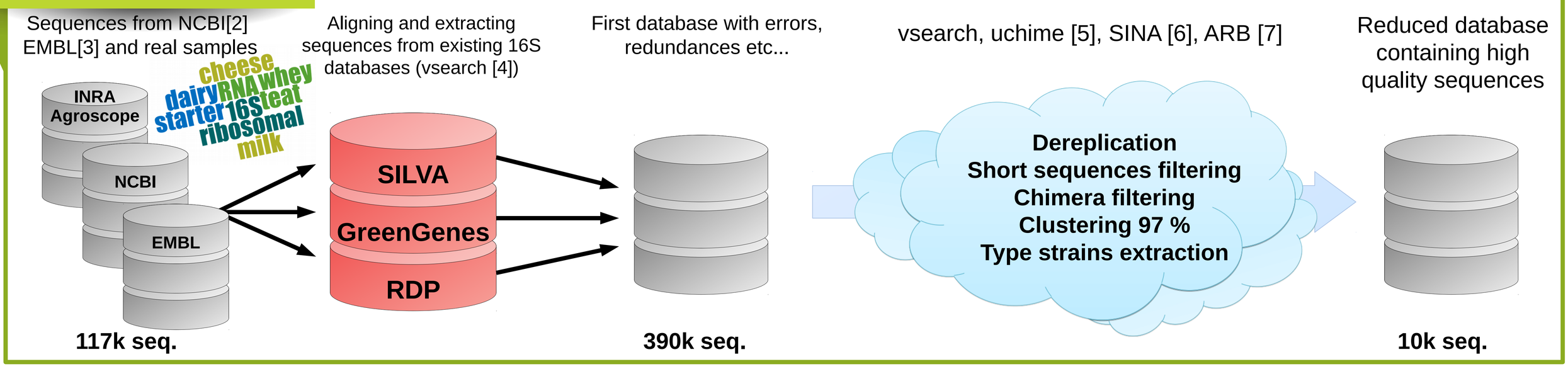
[a] Agroscope, Research Group Fermenting Organisms, Schwarzenburgstrasse 161; CH-3003 Bern; Suisse
[b] Université Clermont Auvergne, INRA, VetAgro Sup, UMRF, 20 côte de Reyne, Aurillac, 15000, France
* M. Meola and E. Rifa contributed equally to this work.

Published by BMC Genomics
2019, july
DOI: 10.1186/s12864-019-5914-8

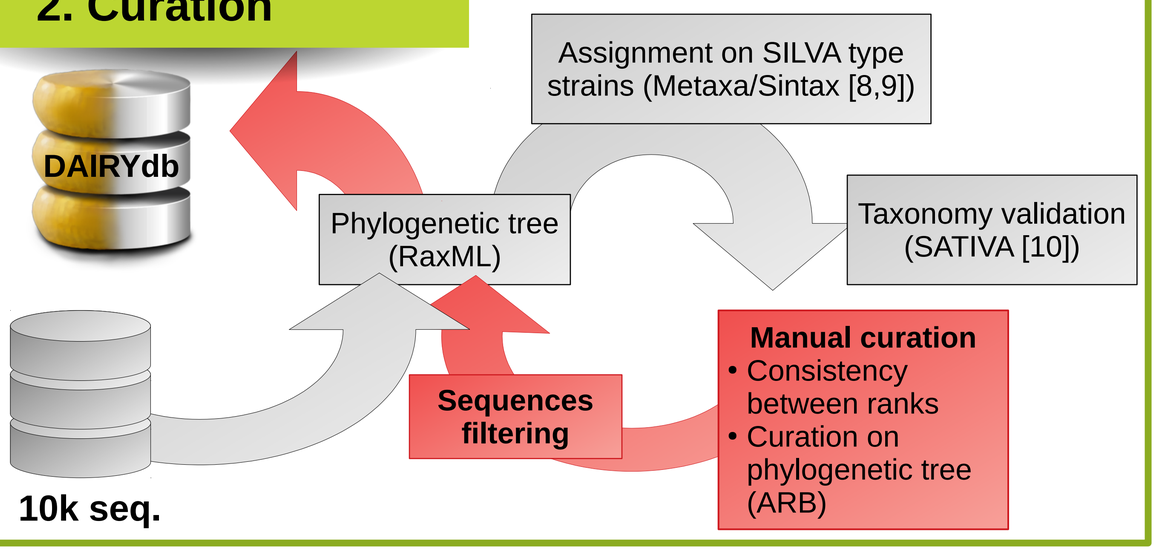
Context

Taxonomic assignment of amplicon sequence variants (ASV) is a key step in the metabarcoding data analysis for microbial community studies of cheese and dairy products. The taxonomic assignment depends mainly on the quality of the reference database. Current databases such as SILVA, RDP or Greengenes (GG) are complete and produce reliable assignments but do not allow to obtain an accurate taxonomic assignment of ASVs at species rank [1]. These databases also contain partial sequences of 16S rRNA with incomplete or erroneous taxonomy which impacts the assignment accuracy. Only a few percentage of their representative sequences present assignation at species rank.
DAIRYdb is a database built from complete and good quality 16S rRNA sequences from bacteria commonly found in cheese and dairy products. Thanks to the different bioinformatic treatments, we obtained a reduced database containing about 10,000 sequences. This allowed us to manually correct their taxonomy and then increase the assignment accuracy of ASVs from dairy products.

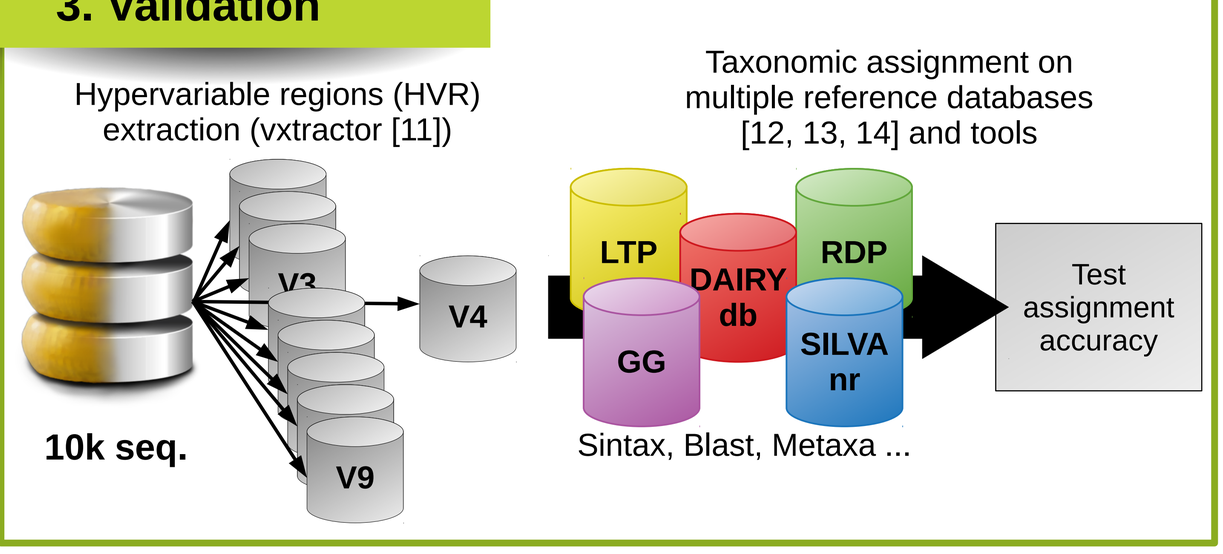
1. Contruction



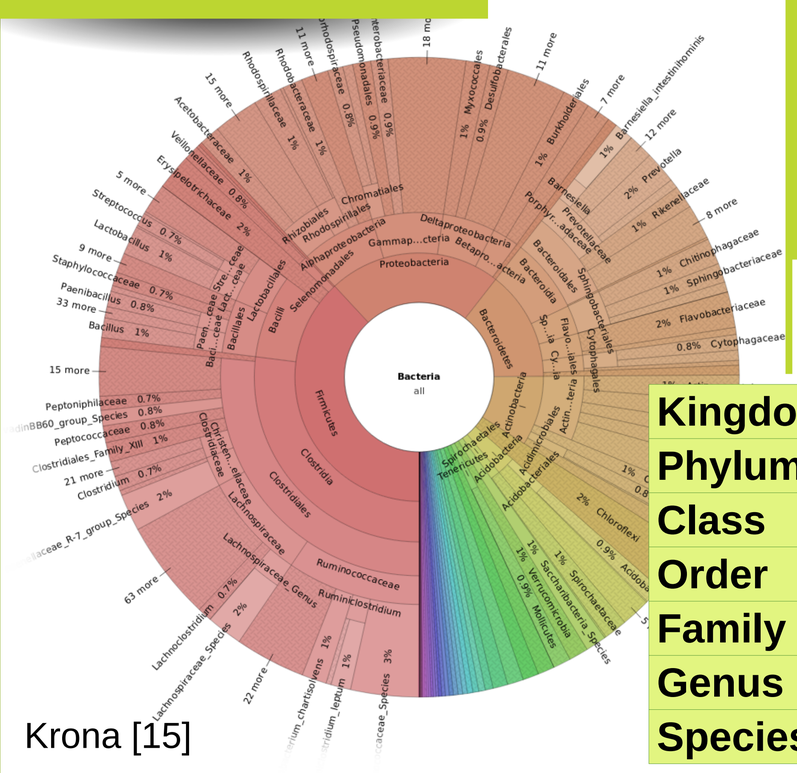
2. Curation



3. Validation



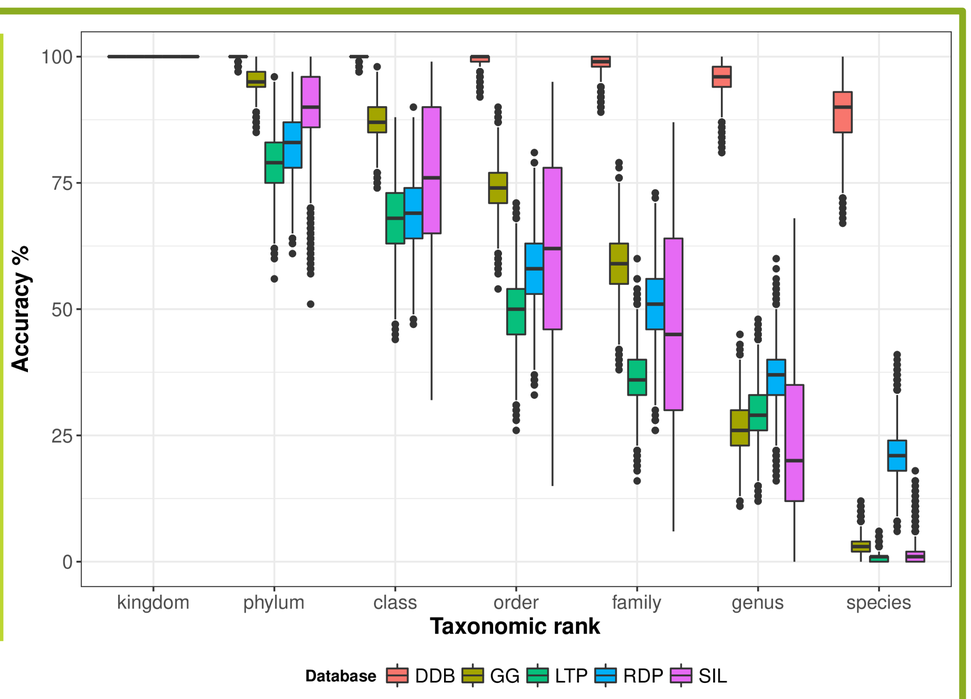
4. Curation



- DAIRYdb is composed of over 10000 sequences
- High diversity
- 38 % Firmicutes, 22 % Proteobacteria, 14 % Bacteroidetes, 9 % Actinobacteria

Kingdom	2
Phylum	46
Class	133
Order	244
Family	462
Genus	1754
Species	4027

- Assignment on each HVR
- DAIRYdb significantly improves taxonomic assignment at each rank.



Sintax assignment accuracy boxplot on every HVR (V1 to V9) tested on 5 different databases.

Conclusions

- DAIRYdb shows better taxonomic assignment accuracy up to species rank than standard databases.
- An environment specific reference database with authoritative full-length 16S sequences covering the diversity of dairy products.
- The reduced number of representative sequences, decreases risk of conflict (high identity and different taxonomy).

References

[1] Ritari, J. et al. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 16, 1056.;
[2] Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 https://www.ncbi.nlm.nih.gov/nucleotide/ ; [3] ENA EMBL [Internet]. European Nucleotide Archive, European Bioinformatic Institute (UK), 2017; https://www.ebi.ac.uk/ena ; [4] Torbjørn Rognes et al. (2016). VSEARCH: a versatile open source tool for metagenomics.; [5] Edgar, R.C. et al. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194–2200. ;
[6] Pruesse, E. et al. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829. ;
[7] Ludwig, W. et al. (2004). ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371. ; [8] Bengtsson, J. et al. (2011). Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. Antonie Van Leeuwenhoek 100, 471–475. ;
[9] Edgar, R. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv 74161. ; [10] Kozlov, A.M. et al. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res 44, 5022–5033.; [11] Hartmann, M. et al. (2010). 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 83, 250–253. ; [12] Quast, C. et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, D590–D596. ; [13] Cole, J.R. et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37, D141–D145. ; [14] DeSantis, T.Z. et al. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72, 5069–5072. ; [15] Ondov, B.D. et al. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385.

UMRF
20, côte de Reyne
15000 AURILLAC
https://www6.ara.inra.fr/umrf
04.71.45.64.10

Available on git repository at :
https://forgemia.inra.fr/umrf/dairydb