Characterization of the genomic diversity and gene content of a lactobacilli collection

Romane Junker, Victoria Chuat, Florence Valence, Michel-Yves Mistou, Hélène Chiapello

Characterization of the genomic diversity and gene content of a lactobacilli collection

Romane Junker[1], Victoria Chuat[2], Florence Valence[2], Michel-Yves Mistou[1] and Hélène Chiapello[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, 78350, Jouy-en-Josas, France
[2] UMR1253 STLO, CIRM-BIA, INRAE, Institut Agro, F35000, Rennes, France

Bactéries d'Intérêt Alimentaire

Introduction

- The *Lactobacillus* genus comprises 261 species displaying a great diversity of genotypes, phenotypes and habitats, some of them exhibiting key importance in food, biotechnology and therapeutic applications. The initial taxonomy of lactobacilli, mostly based on phenotyping traits and chemotaxonomic criteria such as DNA-DNA hybridization, was recently drastically revised using a comparative genomics approach, leading to the creation of 23 novel genera [1].
- In this context, the International Center for Microbial Resources dedicated to food associated bacteria at INRAE (CIRM-BIA) has recently decided to explore and characterize the intra-species genomic and functional diversity of a collection of 250 food associated strains from 21 species reflecting the three major lifestyle categories (free, commensal, nomadic) known for this group. For the CIRM-BIA, the challenge is also to find fast identification methods of closely related species based on specific protein markers.
- In order to analyze this dataset, we designed a bioinformatics workflow allowing the fine characterization of the quality of genomic data, the assembly and annotation of genomes, and the analysis of genomic diversity of the 250 *Lactobacillus* strains, both at inter and intra-species levels

Can a phylogenetic tree representative of the lactobacilli diversity be generated from our dataset ? Which of the distance or maximum likelihood trees is most consistent with the revised classification of lactobacilli ?

Method

Reproducible science approach

Workflow Snakemake [2] — Analysis performed on RMarkdown — Git-repository manager — INRAE bioinformatics facility

Assembly
- fastp filtering of low quality reads [3]
- Unicycler assembly of reads by SPAdes and correction [4]
- Quast evaluation of the assemblies quality by comparison with reference assemblies [5]

Annotation
- Prokka fast annotation of assemblies [6]
- Platon detection of contigs likely to be plasmids [7]

Genome comparison
- dRep [8]

- rapid primary Mash comparison based on kmers
- secondary Average Nucleotide Identity (ANI) comparison on sets of genomes that have at least 90% Mash ANI
- dereplication of the dataset
- ANI comparison on dereplicated genomes

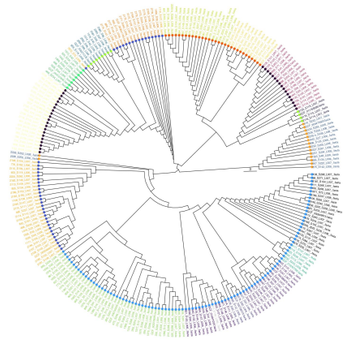
Core genome analysis at the genus level
- SCARAP [9]

- infer the pangenome of a random subset of 30 seed genomes
- build a profile HMM database of "candidate core orthogroups" that are present in at least 25 seed genomes
- identify the candidate core genes in the full set of genomes by searching all proteins of all genomes against the database of candidate core genes.
- identify the core genes from the candidates by imposing a minimum percentage presence cutoff (98%) in the full set of genomes.
- perform a core genome alignment

Tree computation
- ape::BioNJ Neighbor-Joining method using the Mash or ANI distance matrix [10] → Trees A and B
- FastTree Maximum Likelihood method using the core genome protein super alignment (LG + CAT model) with calculation of the Shimodaira-Hasegawa (SH) support [11] → Trees C and D

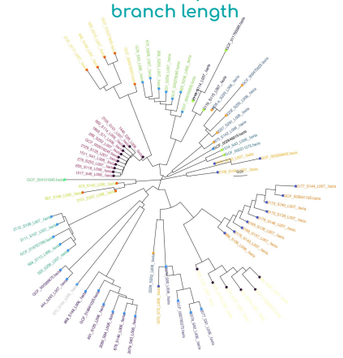
Results

A — Mash distance tree



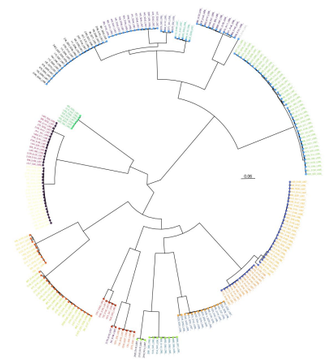
Most strains within the same species belong to the same cluster. This is less frequent at the new genus level (which is expected with the Mash distance).

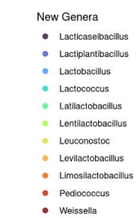
B — ANI distance tree on dereplicated dataset with branch length



The dereplicated dataset is composed of 76 genomes. The number of retained assemblies depends on the species: this can be explained by a difference in intra-species diversity in our dataset.

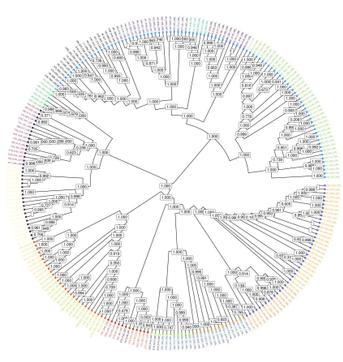
C — Maximum Likelihood tree with branch length





Clusters similar to the ones in Mash distance tree are present at the species level, but the core genome phylogenetic tree also allows the identification of the genera of the new lactobacilli taxonomy.

D — Maximum Likelihood tree with SH support



The obtained topology is highly robust up to the species level.

Conclusion - Perspectives

- The Mash distance tree is relevant at the intra-species level while the maximum likelihood tree from the core genome super-alignment is consistent with the new taxonomy of lactobacilli.
- To further characterize the dataset, a functional study of the strains is planned, in particular to investigate the content in glycoside hydrolases of the strains in connection with the fermentations they perform.
- The pangenome study will allow the identification of gene families specific to a given species and therefore potential species markers (e.g. to differentiate between *plantarum* and *pentosus* species whose 16s RNAs are very similar).

Acknowledgements We are grateful to the ENS Paris-Saclay for the funding of this internship and to INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing help and computing and storage resources.

References
[1] Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M., Harris, H. M., Mattarelli, P., ... & Lebeer, S. (2020). A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *International journal of systematic and evolutionary microbiology*, *70*(4), 2782-2858.
[2] Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520-2522
[3] Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890.
[4] Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, *13*(6), e1005595.
[5] Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.
[6] Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.
[7] Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., & Goesmann, A. (2020). Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial genomics*, *6*(10).
[8] Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal*, *11*(12), 2864-2868.
[9] Wittouck, S., Wuyts, S., Meehan, C. J., van Noort, V., & Lebeer, S. (2019). A genome-based species taxonomy of the Lactobacillus Genus Complex. *Msystems*, *4*(5), e00264-19.
[10] Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526-528.
[11] Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), e9490.