nature communications

Article

https://doi.org/10.1038/s41467-024-52687-7

Genomic and phenotypic imprints of microbial domestication on cheese starter cultures

Received: 29 March 2024

Accepted: 16 September 2024

Published online: 05 October 2024

Check for updates

Vincent Somerville[1,2,3,4] ✉, Nadine Thierer[2], Remo S. Schmidt[2], Alexandra Roetschi[2], Lauriane Braillard[2], Monika Haueter[2], Hélène Berthoud[2], Noam Shani[2], Ueli von Ah[2], Florent Mazel[1,5] & Philipp Engel[1,5]

Domestication – the artificial selection of wild species to obtain variants with traits of human interest – was integral to the rise of complex societies. The oversupply of food was probably associated with the formalization of food preservation strategies through microbial fermentation. While considerable literature exists on the antiquity of fermented food, only few eukaryotic microbes have been studied so far for signs of domestication, less is known for bacteria. Here, we tested if cheese starter cultures harbour typical hallmarks of domestication by characterising over 100 community samples and over 100 individual strains isolated from historical and modern traditional Swiss cheese starter cultures. We find that cheese starter cultures have low genetic diversity both at the species and strain-level and maintained stable phenotypic traits. Molecular clock dating further suggests that the evolutionary origin of the bacteria approximately coincided with the first archaeological records of cheese making. Finally, we find evidence for ongoing genome decay and pseudogenization via transposon insertion related to a reduction of their niche breadth. Future work documenting the prevalence of these hallmarks across diverse fermented food systems and geographic regions will be key to unveiling the joint history of humanity with fermented food microbes.

Domestication is the process of modifying wild species through artificial selection to the benefit of a "domesticator", which is usually human[1,2]. This process was integral to the rise of complex human societies[3–5]. In particular, the domestication of crop plants[6] and livestock[7] over the past 12,000 years[4] enabled the emergence of social complexity and initiated large-scale anthropogenic changes in the earth's biosphere[8]. Plant and animal domestication is characterised by phenotypic and genetic changes from the wild ancestor, such as adaptation of body mass[9] or change in nutrient content[10], which overall contributed to the production of food surplus.

This oversupply of food was associated with the formalisation of food preservation strategies by microbial fermentation – a metabolic process that converts sugars into acids – to produce, for example, fermented vegetables, wine or cheese[11,12]. This leads to a decrease in pH, which reduces undesired microbial growth and prevents spoilage of stored food. These processes were likely adopted by humans many millennia ago[13]. For cheese, the earliest archaeological evidence can be found within the Neolithic (ca. 12th-5th millennium B.C.)[11,13] and likely became more common and important during the Copper, Bronze and Iron Age (ca. 5th-1st millennium B.C.)[14]. Fermented food products have diversified in a myriad of forms all over the globe[15,16], constituting healthy and tasty components of the human diet, which are key in many cultures and sustainable opportunities for the future of food in others[17]. This raises the fundamental question of whether, as for plants

[1]Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland. [2]Agroscope, Liebefeld, Switzerland. [3]Université Laval, Quebec, Canada. [4]McGill, Montréal, Canada. [5]These authors contributed equally: Florent Mazel, Philipp Engel. ✉e-mail: vincent.somerville@unil.ch

and animals, fermented food microbes have also been domesticated by humans, and if so, when and how this has happened[18,19].

It is commonly accepted that some fermented food microbes have been maintained through continuous passaging[20,21] and artificial selection for specific traits (e.g., shelf life or taste). Accordingly, we expect them to present typical "hallmarks" of domestication, i.e., genomic and phenotypic signatures associated with (microbial) domestication that distinguish them from their wild counterparts. In fact, a handful of microbial domestication cases have been documented empirically[22,23], mostly for eukaryotic microbes. For example, *Saccharomyces cerevisiae* used for bread making[24] and alcoholic fermentation[25,26], *Aspergillus oryzae* used for sake, soy sauce, and miso production[27] and *Penicillium camemberti* found on camembert cheese rind[28] show genomic and phenotypic characteristics that distinguish them from their wild counterparts[18,29,30].

Thus, while the genomic signatures of domestication are well-defined for eukaryotic microbes, we know less about the collective genomic and phenotypic consequences of bacterial domestication used for food fermentation[29,31]. For example, *Oenococcus oeni*, which is responsible for the malolactic conversion in winemaking, is thought to have rapidly diverged from its ancestor due to the emergence of hypermutator strains, but it remains unclear to what extent subsequent evolution was influenced by human domestication[32]. Similarly, strains of *Lacticaseibacillus paracasei* isolated from ripening cheese show signs of adaptation to milk but lack other hallmarks of domestication[33]. Altogether, this suggests that most suspected cases of domestication in fermented food bacteria lack a complete characterisation of domestication hallmarks previously observed in microbial eukaryotes like yeast.

To explain this discrepancy, we posit that, unlike yeasts (i.e., in beer, bread and wine), the previously studied bacteria have not undergone continuous passaging with rapid and iterative bursts of growth solely on the fermented foods but rather persisted in food-associated environments (e.g., grape skin for *Oenococcus oeni*). This reduces the strength with which artificial selection can act on evolution. In contrast, cheese starter cultures (starter cultures) are a promising candidate to test for domestication in fermentation. These cultures have been passaged and selected for thousands of years via backslopping (i.e., continuous re-inoculation of previous-day whey) and have been extensively selected for flavour and rapid acidification purposes[34]. Starter cultures are categorised by their usage temperature: mesophilic (below 42 °C) and thermophilic (above 42 °C). Mesophilic cultures, dominated by *Lactococcus lactis* or *L. cremoris*, show taxonomic and genetic adaptations to the dairy environment, namely the loss of genes encoding functions needed in animal and plant environments[29]. However, these species maintain comparatively large genomes[30] and occur in multiple environments[31] partially due to the presence of diverse plasmids[32,33]. Thermophilic starter cultures are dominated by three thermophilic bacteria, namely *Streptococcus thermophilus* and *Lactobacillus delbrueckii* subsp. *lactis* (hereafter only *L. delbrueckii*) and *Lactobacillus helveticus*[35]. Although previous genomic analyses have shown signs of genome decay in these species[22,36–40], systematic screening for domestication hallmarks in these microbial communities is still lacking[20,21].

In this study, we aimed to detect signs of domestication in thermophilic starter cultures and to date the potential domestication events. To this end, we collected both modern and historic (1970s) samples of 11 undefined starter cultures that are continuously passaged as undefined starter cultures (described in refs. 41–44) and used to make three cheese varieties in different regions in Switzerland (Fig. 1A). We characterised over 1000 samples phenotypically, and about 100 metagenomes and more than 100 bacterial isolates from historical and modern Swiss starter cultures were genetically characterised (Fig. 1B). By conducting species dating and genomic analysis of additional isolates from public databases, sampled in different geographic locations and dairy fermentations, we expanded on previously proposed hallmarks of domestication for eukaryotes to define five specific hallmarks of microbial community domestication for bacteria[29]: (i) phenotypic reliability over time, as a result of the selection for food preservation, (ii) simple and stable microbial diversity both at the species and strain-level, as a result of continuous passaging, (iii) evolutionary origin of focal species coinciding with the start of food preservation, (iv) gradual genome decay and (v) adaptation to the food environment by the gradual reduction of niche breadth. Collectively, our results suggest that thermophilic starter cultures have been domesticated by humans for millennia.

Results

Food preservation by rapid and reliable acidification as a result of lactose fermentation by thermophilic cheese starter cultures

The first hallmark of domestication predicts phenotypic stability over time as a result of continuous passaging in a stable environment (i.e., milk) and selection for a specific trait, especially food preservation but also ripening and flavouring, as determined by a characteristic aroma and favourable process parameters. The underlying phenotypic properties that give rise to food preservation (i.e., preventing the growth of undesired microbes) are the rapid and reliable (i) acidification of the environment and (ii) reduction of the amount of easily accessible nutrients (Fig. 1C).

We analysed over 1000 weekly routine quality-control measurements for three key phenotypes of food preservation via lactose fermentation, namely acidification, lactate production, and D- to L-lactate ratio (Fig. 1B, C). In routine starter culture quality control, acidification values above 110 acid base titration value (°Th), amount of detected lactate between 6 and 12 g/kg and D- to L-lactate ratio between 12 and 45 are generally regarded as efficient milk fermentation. Firstly, we titrated 10 ml of > 1000 samples and found that acidification potential was consistently (> 110 acid base titration value (°Th)) and reproducibly high over time with no significant temporal trend (Fig. 1D, mixed effect linear model time slope estimates with $p < .05$). Secondly, the amount of detected lactate and the ratio of the corresponding enantiomers, D- and L-lactate, slightly but not dramatically increased over time (Fig. 1E, F, mixed effect linear model time slope estimates with $p < .05$). The former suggests a stable nutrient depletion by accumulation of lactate as a final product. The latter suggests that the relative metabolic contribution of the two dominant community members is similar across time. This is essential as *S. thermophilus* ferments rapidly but only until pH 5, whereas *L. delbrueckii* commonly starts significantly fermenting from pH 5 downwards for a longer time.

In summary, these results show that starter cultures are phenotypically stable, measured by rapidly (within the quality control boundaries) and reliably (no significant trend over time, and the distribution of phenotypes within the range of what is considered efficient in the industry, in Fig. 1d) acidify milk, as expected when communities are selected for specific traits.

Cheese starter cultures are simple and stable microbial communities

The second hallmark of domestication predicts that the microbial diversity, both at the species and strain level are simple and stable as a result of the passaging in a contained and highly stable nutrient-rich environment. We tested this by selecting a total of 98 starter cultures for shotgun sequencing (6–11 Mio. reads per sample; see circles, triangles, and squares in Fig. 1B) and determined their taxonomic composition using a short read taxonomic profiler (mOTU2[45]). As expected, we found that most samples were dominated by only two species, *Streptococcus thermophilus* and *Lactobacillus delbrueckii* subsp. *lactis* (Fig. 2A). Yet, we noted two apparent signs of instability over time: (i) *L. helveticus* was only present in early samples of some

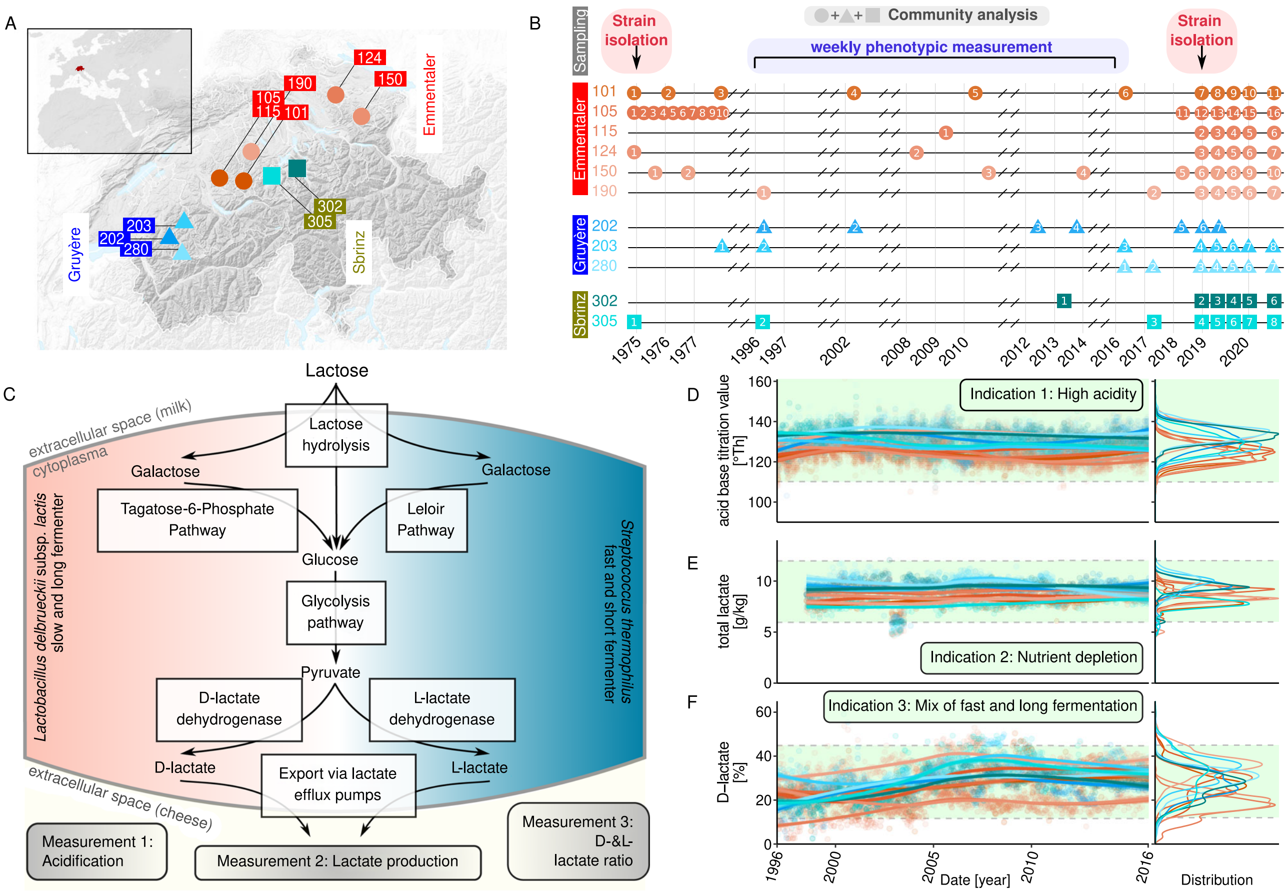

Fig. 1 | Phenotypic stability of thermophilic cheese starter cultures over 50 years of cheese production. A Origin of 11 continuously passaged starter cultures originating from different regions in Switzerland and different cheese varieties (the inset shows the location of Switzerland within Eurasia.) **B** Both modern and historic sampling time points for the community analysis (coloured circles, triangles, and squares), phenotypic analysis (blue background) and strain isolation (red background) of all starter cultures. **C** The general process of food preservation for thermophilic starter cultures consisting of stable (i) acidification, (ii) lactate production and (iii) D- to L-lactate ratio (ratio of slow fermenters (*L.delbrueckii*) and fast fermenters (*S.thermophilus*)). Phenotypic measurements of the starter cultures (**D**) total acid-base titration value, (**E**) total lactate and (**F**) percent D-lactate with respect to L-lactate. **D**–**F** The circles represent the individual measurements, and the lines represent the smoothed rolling means. All colours correspond to the starter culture colour introduced in (**A**), and the green shading represents the values expected in routine starter culture quality control. In (**E**), there was a systematic measurement error in the lactate measurement in 2003/4, which was corrected in 2004 and did not influence the rolling mean.

cheese starters, and (ii) the relative abundance of *S. thermophilus* and *L. delbrueckii* varied across samples of the same starter culture (Fig. 2A). The former observation suggests that *L. helveticus* may have been lost over time without changing the phenotypic properties of the starter cultures. The latter observation may be due to the fact that the precise sampling time points during the acidification process by the starter cultures were not controlled for and that *L. delbrueckii* growth is delayed relative to *S. thermophilus* during acidification[41] by their mutualistic interaction termed protocooperation[46,47]. Also, we cannot exclude that sample degradation through time may have biased the community composition (the older samples are typically not viable anymore). The stable D- to L-lactate ratio (Fig. 1F) corroborates that the integrated metabolic activity of the two species at the sampling timepoint remains stable over time. Altogether, this suggests that the species-level composition of the starter culture was remarkably stable over nearly 50 years of sampling, with no additional species identified in the metagenomic sequences.

To assess the within-species diversity, we mapped the metagenomic reads against a reference database containing isolate genomes of each of the three species and quantified the number of polymorphic sites detected in core genes (i.e., genes identified in all strains of a given species, see methods) in each sample. The proportion of polymorphic sites was similar over the samples, with around 0.11% (SD = 0.72 %) and 0.02 % (SD = 0.03 %) for *S. thermophilus* and *L. delbrueckii*, respectively (SFig. 1). This is comparatively low with respect to bacteria found in non-food fermentation systems like in the gut of animals (3% and 2–10% polymorphic sites within species in the gut microbiota of human[48] and honey bees[49], respectively, even for *Lactobacillus* species).

To determine the actual number of strains the detected within-species diversity corresponds to, we genotyped > 2000 colonies from the 11 starter cultures (Fig. 1B) and sequenced the genomes of 112 isolates. Using an all-against-all genomic distance analysis implemented in poppunk[50], we found that the sequenced genomes separate into 12 *S. thermophilus*, two *L. delbrueckii* and two *L. helveticus* sub-species clades (see "Methods"). Overall, the sub-species clades within the different species are very similar (min ANI: Sterm = 98.6 %, Ldel = 98.9 %, Supplementary Fig. 2[51]). These sub-species clades accounted for most of the SNPs detected by metagenomic sequencing (93% and 78% of the metagenomic SNPs from *S. thermophilus* and *L. delbrueckii*, respectively, Fig. 2B and Supplementary Fig. 3), suggesting that we have isolated and sequenced most of the sub-species diversity present in the analysed starter cultures. In the case of *L. delbrueckii* and *L. helveticus*, a single sub-species clade dominated in all analysed samples, while for *S. thermophilus*, one to four sub-species clades per sample were detected.

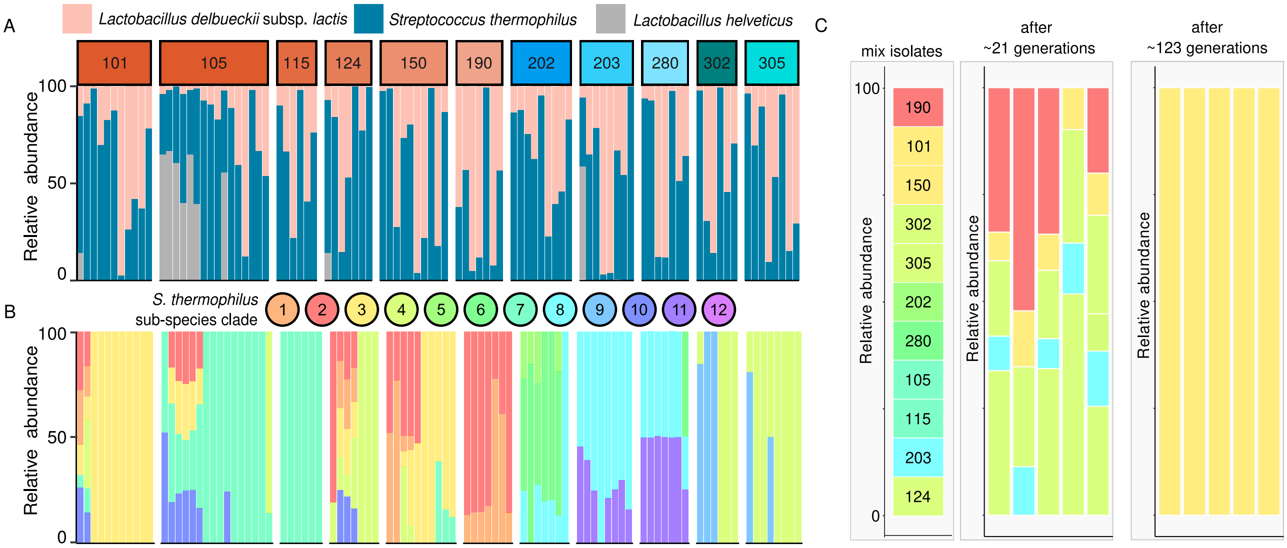

Fig. 2 | Simple and stable community composition both at the species- and sub-species-levels in cheese starter cultures. A Species-level composition across all samples. Within each starter culture, samples are ordered from older to more recent. **B** Compositional diversity across the different samples for the 12 different *S. thermophilus* sub-species clades identified (each colour is representative of a sub-species clade). **C** The propagation experiment involved 11 isolates of *S. thermophilus* and 11 isolates of *L. delbrueckii*, each being the most dominant sub-species clade from the 11 different starter cultures. Note that *L. delbrueckii* is not illustrated, as it only included one sub-species clade per starter culture. The results are shown after ~ 21 and ~ 123 passages for the five biological replicates.

Finally, to test if the milk environment in which these bacteria have been passaged cannot carry a larger amount of diversity at the sub-species level, we inoculated sterile milk with a diverse synthetic community containing all 11 dominant *L. delbrueckii* isolates (all from the same sub-species clade) and all 11 dominant *S. thermophilus* isolates (from seven different sub-species clades) each from one of the 11 starter cultures. We inoculated them at equal amounts and co-cultured them in five replicates for 27 passages (~ 123 generations) (Fig. 2C). After 21 generations, we found that five of the 11 original *S. thermophilus* strains were still present. However, after 123 generations, only one strain could be detected in the community (Fig. 2C), the most dominant strain in RMK 101 and 150). This experiment suggests that only a limited number of strains can co-exist due to passaging, as observed in natural settings, but that other local factors (not mimicked in our experiment) might contribute to the dominance of alternative strains in different starter cultures.

In summary, these results show that undefined starter cultures have a simple community composition at the species and strain levels that shows a high degree of stability, in particular at the species level, likely due to the continuous passaging in a stable, closed and nutrient-rich environment.

The onset of cheese starter strain diversification coincides with the origin of dairy fermentation in humans

The third hallmark of domestication predicts that the evolutionary origin of starter bacteria should coincide with or postdate the start and diversification of cheese making or dairy fermentation in prehistoric times, as referenced by the archaeological record. The actual timing of the origin of cheese making is difficult to estimate from archaeological evidence because cheese making tools (wooden vessels, cloth strainers, leather sacks) and residues are usually organic and unlikely to be well preserved over millennia[52,53]. The first archaeological evidence of cheese making based on organic matter found in ceramic stainers places the latest estimate of cheese-making origin around 7000 BC in the Near East[13] and around 6000 BC in Europe[11]. We thus hypothesise that the origin of the starter bacteria should similarly fall within the Neolithic, and possibly even into the Copper, Bronze, and Iron Ages, when important innovations in cheese making occurred[11]. To test this, we sought to date the evolutionary transition(s) from a non-dairy environment to a dairy environment using phylogenetic comparative methods. This approach first maps strain or species habitat preference (dairy/non-dairy) onto a phylogeny to identify the evolutionary transition(s) between habitats and then, in a second step, uses the molecular clock and the dates of historical samples as calibration points to estimate the age of this transition.

We assembled a genomic database of 234 strains from the three cheese starter bacteria as well as closely related species by combining our own dataset with publicly available data (complete list of genomes provided in Supplementary Data 1). We mapped the isolation source on whole genome phylogenies (for within-species phylogenetic relationships) and on rRNA phylogenies (for between-species relationships). Most strains from *S. thermophilus, L. delbrueckii* and *L. helveticus* were isolated from cheese starter cultures or other milk products, with a few isolates from faecal samples (Supplementary Fig. 4). Corroborating previous findings[54], ancestral niche reconstruction suggests that with a high likelihood of 99.99%, 99.99% and 76% the niche of the ancestors of all known strains of *S. thermophilus, L. delbrueckii* subsp. *lactis*, and *L. helveticus*, respectively, were already associated with dairy products (Maximum likelihood model, Supplementary Fig. 5). In addition, although the three closest related sister species to our focal species are of animal rather than dairy origin, they all encode enzymes to degrade milk (LacG or LacZ galactosidases, Fig. 3A–C). This suggests that these milk-adapted sister clades probably evolved with the appearance of milk in mammals roughly 200 million years ago[55], but only our focal species are found in the dairy environment. Therefore, we propose that the origin of cheese starter bacteria is located somewhere between the split from the sister species (stem age) and the most recent common ancestor (MRCA) of the strains within the species (crown age) of the three starter culture species (branches highlighted in black in Fig. 3A–C).

To estimate the age of the non-dairy-to-dairy transition, we took advantage of two independent molecular clocks of the stem and crown age. Stem ages were determined using the divergence time from the rRNA phylogeny. Assuming an rRNA substitution rate of 1 substitution per site per 100 million years[56], we estimated the divergence of *S. thermophilus*, *L. delbrueckii* subsp. *lactis* and *L. helveticus* from their sister taxon at around 432,000, 264,000 and 1,044,000 years ago, respectively (Fig. 3A–C). While this is more recent than most other

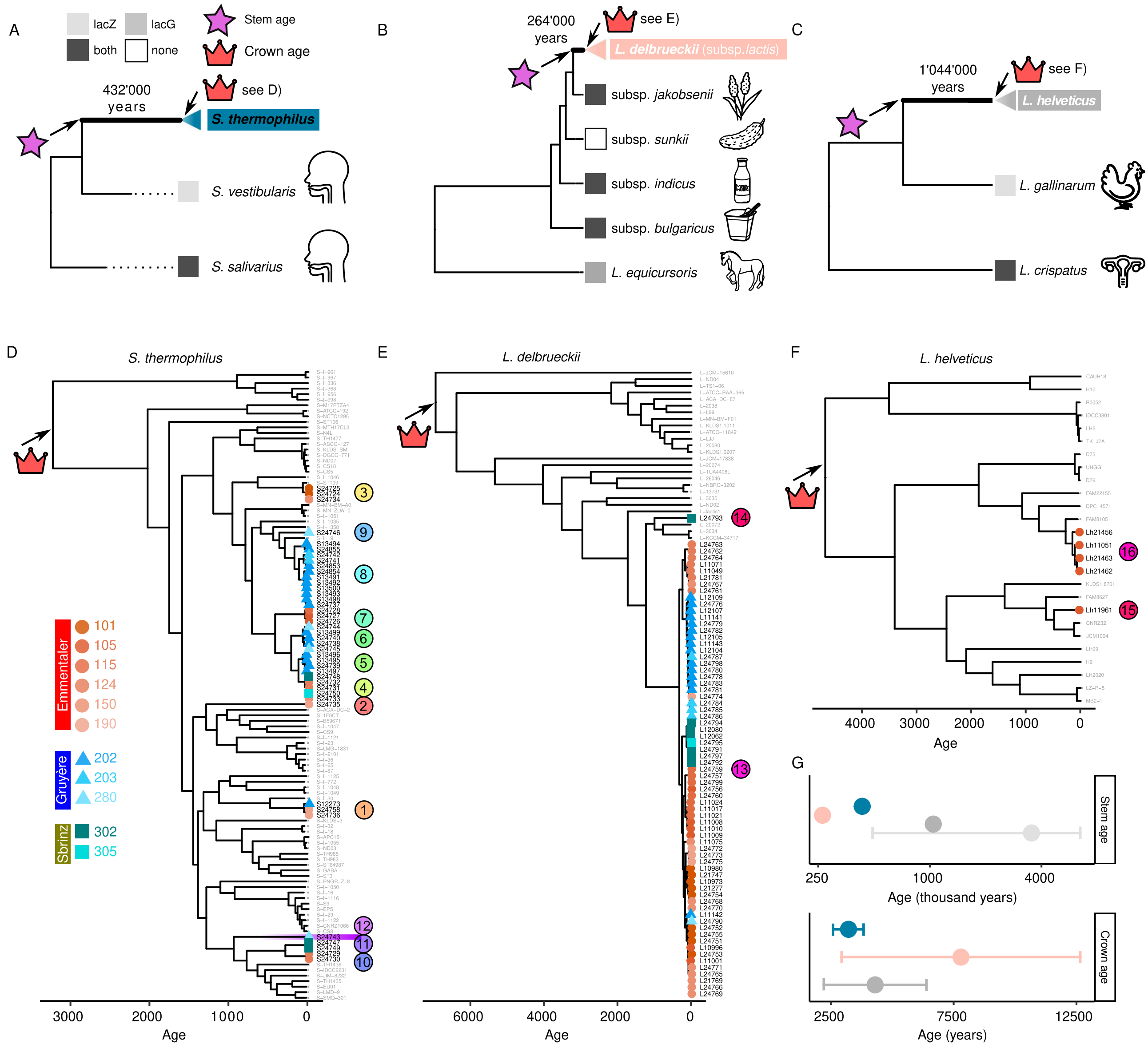

Fig. 3 | Phylogenetic dating of the origin of cheese starter bacteria. A–C Maximum likelihood rRNA operon (~5 kb) phylogeny of the focal species and their closest sister (sub)species. Tips are annotated with labels indicating the presence of lacZ or lacG genes within the genomes and the sampling origin of the different isolates (depicted by silhouettes). Stem age is depicted on the phylogeny with a star and annotated with its estimated age, along with a representation of sub-species diversity, depicted with a triangle that represents the phylogeny of panels (**D**–**F**). **D**–**F** Genome-wide phylogenies illustrating recent diversification of cheese starter sub-species clades for (**D**) *S. thermophilus*, (**E**) *L. delbrueckii* and (**F**) *L. helveticus* with the crown age depicted as a crown. Different sub-species clades are highlighted in circled numbers in different colours. These colours and numbers match sub-species clades described in Fig. 2B, C). **G** The stem age is calculated by the molecular clock of the rRNA operon for the different species and genera (also annotated near the stem nodes in panels (**A**–**C**)). The starter culture-related species are highlighted with the corresponding colours (*L. delbrueckii* = pink, *L. helveticus* = grey and *S. thermophilus* = blue). The light grey point and whiskers illustrate the mean and standard deviation of the stem age of all other species in the lactic acid bacteria family. The crown ages are inferred by a molecular clock rate based on the core genome phylogeny and historical samples as calibration points after removing putative horizontal transfer events. The whiskers in the crown age illustrate the confidence interval of the molecular clock dating.

species within the two genera (Fig. 3G), it is still three orders of magnitude older than the first report of dairy fermentation. Crown ages were determined using dated core-genome phylogenies of our focal species (Fig. 3D–F and Supplementary Fig. 6). Using a molecular clock based on the core genome phylogeny and historical samples as calibration points[57], we estimated a substitution rate of 1.1 SNPs per clonal core genome per year, which falls in the average range typically observed for other bacterial species[58,59]. We extrapolate that the crown age of *S. thermophilus*, *L. delbrueckii* subsp. *lactis* and *L. helveticus* were around 3221, 7798 and 4304 years ago, respectively (Fig. 3G). This method provides an approximate age range for the different clades rather than precise dating represented by the large estimated confidence interval errors, ranging from 2163 to 12,708 years ago. Nevertheless, it indicates that the origin of known strain-level diversity within species is roughly similar to the expected origin of dairy fermentation from archaeological records[11].

Cheese starter bacteria show genome decay by transposon expansion

The fourth hallmark of domestication predicts that cheese starter bacteria show signs of gene loss and genome decay[60]. This is expected when bacteria thrive in stable environments with extensive nutrient

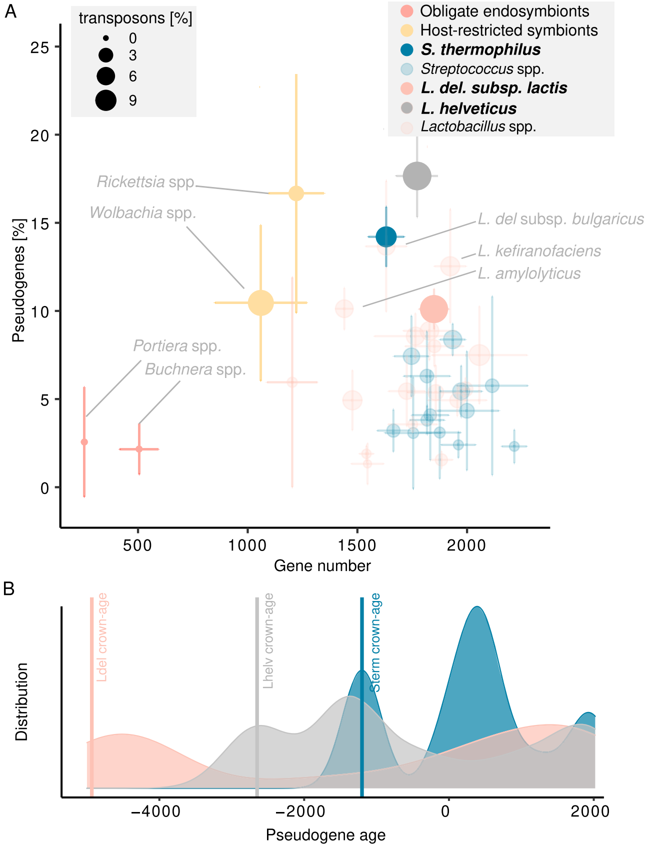

Fig. 4 | Dating of ongoing genome decay by pseudogenization. A The fraction of pseudogenes per genome vs. the gene number in the genomes of (i) two obligate endosymbionts (*Buchnera* spp. and *Portiera* spp.), and host-restricted symbionts (*Wolbachia* spp. and *Rickettsia* spp.), (ii) the three focal species (*S. thermophilus, L. delbrueckii* and *L. helveticus*), and (iii) all other *Lactobacillus* spp. and *Streptococcus* spp. for comparison. The size of the points indicates the percent of transposons detected in the genomes; the error bars indicate the standard deviation of the pseudogenes (y-axis) or the gene number (x-axis). Specifically for the three focal species, this is *S. thermophilus*: mean = 1657 genes, std = 47; *L. delbrueckii*: mean = 1868 genes, std = 33 and *L. helveticus* = 1772, std = 95. % of pseudogenes: *S. thermophilus*: mean = 14%, std = 1%; *L. delbrueckii*: mean = 10%, std = 0.6% and *L. helveticus* = 18, std = 2.3. **B** The smoothed density plot of the approximate date of the pseudogenes is estimated from the most recent common ancestor of all taxa containing a given pseudogene. Vertical lines indicate the MRCA of all sequenced strains within the species.

availability and relatively small population sizes[61]. This hallmark additionally predicts that the onset of genome decay must have started after the onset of domestication itself. To test this hallmark, we compared the genome size and the extent of genome decay (number of pseudogenes) of our focal cheese starter strains to their closest relative. To put these estimates into a broader context, we also included data on insect-associated bacterial endosymbionts that show extensive genome decay due to pseudogenization[62].

All three starter culture species had smaller genome sizes (Wilcoxon test, $p < 0.05$, Fig. 4A and Supplementary Fig. 7) and a higher number of pseudogenes (Wilcoxon, $p < 0.05$, Fig. 4A and Supplementary Fig. 8) than closely related species. We found that many of the detected pseudogenes in the three starter culture species were the result of insertion events of transposons (mean = 45%, std = 9%; Supplementary Fig. 9) belonging to 15 different IS element families (Supplementary Fig. 10). Three other *Lactobacillus* species had similar signs of genome decay, all of which were also food fermentation-associated species, namely *L. delbrueckii* subsp. *bulgaricus*, *L. kefiranofaciens* and *L. amylolyticus*. The observed genome decay of our focal genomes is in the range of host-restricted symbionts such as *Rickettsia* spp. and *Wolbachia spp.*, but not as extreme as for obligate endosymbionts of plant-sap feeding insects that have co-evolved with their host for millions of years. Overall, our findings suggest that all three starter culture species have experienced pseudogenization via transposon insertion during their evolutionary history.

If genome decay started at the onset of domestication as predicted from the hallmark, the timing of pseudogenization events should be more recent than the onset of domestication itself. To estimate the onset of genome decay, we reconstructed the evolutionary history of pseudogenization events by mapping the presence/absence of modern pseudogenes onto the strain phylogenies. For each pseudogene, we identified the most recent common ancestor (MRCA) of the genomes which contained the pseudogene or in which the gene was completely missing (assuming the pseudogene was lost, Supplementary Fig. 11). While we observed a few pseudogenes originating around the crown age (oldest peak, Fig. 4B), we found that most pseudogenes likely originated in the last millennia (intermediate peak, Fig. 4B) or even as recent as in the last century (most recent peak, Fig. 4B). The lag between the crown age (prehistoric) and the increase in pseudogenization in the last millennia, could arise if cheese making remained a spontaneous fermentation-like process for millennia before being more tightly controlled by backslopping. We conclude that the overall age similarity between the loss of functional genes (i.e., pseudogenization) and the cheese making development (within the last couple of millennia) for all three focal species suggests that the persistent bottleneck and selection pressures of cheese-making are associated with recent and ongoing genome decay.

Reduction of the niche breadth and adaptation to the cheese making environment

The fifth hallmark of domestication predicts a reduction of the niche breadth associated with the loss of non-essential functions due to the adaptation to the stable and nutrient-rich cheese environment. We first tested the ability of starter culture bacteria to grow on a wide range of carbon sources that are representative of diverse non-dairy environments. We found that *S. thermophilus* and *L. delbrueckii* can metabolise only 5 and 9, respectively, of the 92 tested carbon sources, while their closest non-dairy relatives could metabolise 12 (58% drop), 61 (85% drop), respectively, (Supplementary Figs. 12, 13).

We then sought to explore whether the pseudogenization of genes observed in these species could explain the loss of these metabolic capabilities. We tested whether cellular functions - in particular carbon metabolism - were enriched in pseudogenes. Pseudogenes were spread across many orthologous gene families (OGs). Of a total of 19,728 OGs in all three species, we identified 5639 (29%) containing at least one pseudogene. Three COG categories were overrepresented among the OGs containing pseudogenes, independently in each of the three starter culture species: carbohydrate (G), amino acid (E) and inorganic ion transport and metabolism (I) ($\chi2 > 0.3$, p-value < 0.05, Fig. 5, Supplementary Fig. 14). More specifically, the three KEGG metabolic modules (i) pentose phosphate pathway (ko00030), (ii) fructose and mannose metabolism (ko00051) and (iii) starch and sucrose metabolism (ko00500) were commonly pseudogenized (Supplementary Fig. 15).

To confirm that the identified pseudogenes are indeed non-functional, we carried out RNA-seq of cheese during the first 24 h of cheese making to look at their expression levels. We found that pseudogenized genes were generally less expressed than non-pseudogenized genes throughout the first 24 hours of fermentation (RNA-seq experiment, Supplementary Fig. 17).

In summary, these results show that recent pseudogenization affected genes involved in the degradation of carbohydrates that occur in plants but not in milk, suggesting that cheese making strains

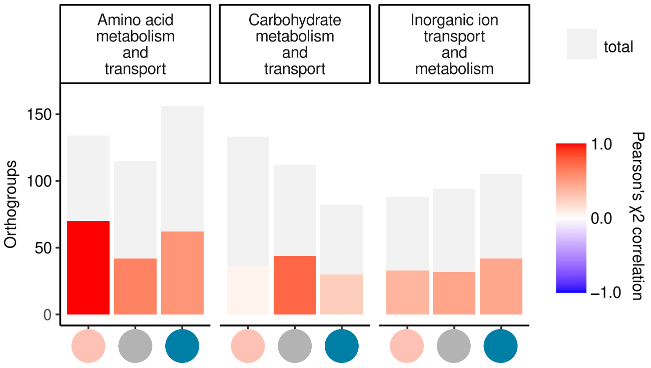

Fig. 5 | Enriched functions in pseudogenes. The total number of orthogroups (grey) and the orthogroups containing pseudogenes coloured if overrepresented (red) based on the Pearson chi-square test (scale indicated in the figure) for the three most overrepresented COG categories. The three different species (*L. delbrueckii* = pink, *L. helveticus* = grey and *S. thermophilus* = blue) are separated accordingly by colour.

have lost genes that appear non-essential today in a stable and predictable dairy environment but were likely more important in their ancestral niche.

Discussion

Overall, fermented foods represent a stable, microbe-rich environment where human-controlled continuous passaging and selection results in the establishment of defined and stable microbial communities with specific phenotypic properties. So far, apart from the domestication of some eukaryotic microbes (e.g., Saccharomyces cerevisiae used in beer, wine or bread), we know surprisingly little about the history of fermented food microbes, and whether they have been domesticated as extensively as cattle and crop plants remains an open question[29,31].

We can conclude from our study that thermophilic starter cultures show clear signs of domestication. (i) They are highly reliable in their acidification and lactose utilisation, likely because of the ongoing selection of the preservation properties (hallmark 1). (ii) They contain simple and stable microbial communities perhaps because of continuous passaging (hallmark 2). (iii) The timing of the origin of the strains is roughly similar to the emergence of dairy fermentation and the resulting anthropogenic selection pressure (hallmark 3). (iv) They show clear signs of recent and ongoing genome decay that can be expected from stable and nutrient-rich environments with relatively small population sizes and continuous passaging (hallmark 4). (v) They show a reduction of the niche breadth and adaptation to the cheese making environment, suggesting that their current niche is restricted to dairy fermentation batches (hallmark 5).

We acknowledge that the domestication hallmarks highlighted here could vary substantially across species and systems depending on the processes and routes of domestication[63]. We suggest that comparing and classifying various fermented food systems into distinct routes of domestication represents an exciting avenue for future research (as previously considered for eukaryotes[64]). In addition, the list of hallmarks we provide here is not necessarily complete. For example, recent studies have looked at the evolution of the pangenome of the different clades of *L. delbrueckii*[40,65] and *S. thermophilus*[39]. While these studies similarly conclude that these species have a distinct evolutionary history tied to dairy fermentation, it currently remains unclear if the pangenome is open and expanding[39]. Notably, the accessory genome contains numerous carbohydrate utilisation genes suggesting a broad nutrient range. Here, we suggest that the size of the pangenome is likely also a consequence of extensive and ongoing pseudogenization because of a substantial fraction of highly mobile genes in the accessory genome, namely active transposases, phages or phage defence genes[40,41,66,67].

Interestingly, phages targeting *S. thermophilus* have adapted to the specific dairy conditions similarly to their host, potentially playing an important role in shaping the remaining genetic diversity of the bacterial species[68]. In any case, the addition of more high-quality genomes and a broader sampling in dairy and non-dairy environments will enable us to understand the pangenome diversity better and explore how it relates to domestication history. Finally, including a wider diversity of cheese starter bacteria from all over the world will probably help push back the date of the onset of the diversification of cheese bacteria and represent an important research avenue.

Collectively, we suggest that thermophilic starter cultures have been domesticated by humans for millennia. Domestication probably started during the Neolithic era when cheese making emerged but likely gradually continued for millennia[14]. In our data, this is evidenced by the large difference in timing of origin between cheese starter members as well as the delay between the origin of the current strain diversity and the pseudogenization. While the exact timing of the transition from a spontaneous dairy fermentation to backslopping is impossible to pinpoint precisely, our estimates of starter culture origin using molecular data roughly agree with the archaeological record. Overall, we suggest that the bacteria have likely been in repeated cycles of selection for rapid and reliable acidification and also genetic drift through the repeat subsampling by backslopping.

Our study fills an important knowledge gap by addressing if, when and how microbes have evolved to anthropogenic usage and provide a conceptual framework that can be applied to other fermented food products. For example, from the pseudogenization analysis (Fig. 4), we pinpoint candidate species that might have been domesticated, in particular, *L. delbrueckii* subsp. *bulgaricus* which has previously been associated with a distinct evolutionary history[69] and genomic repertoire[40] tightly linked to yogurt production. Moreover, it will be instrumental in guiding decision makers in the cheese community to decide on optimal community and genomic diversity guidelines, as, for example, has already been done in beer brewing[70]. While rapid and reliable acidification is the primary role of starter cultures, many secondary roles, most notably phage defence and flavour production, play key roles in the selection of starter cultures today.

Key questions in domestication research ask whether, for a given domesticate, there was a single domestication event that was restricted to a particular geographic area that spread across regions or if there were multiple independent domestication events[12]. For cheese starter bacteria, this remains unknown. In the case of crop plants and animals, the most likely scenario seems to depend on the domesticate and the continent[1,64]. The origin of *S. cerevisiae* is likely Asian, but it is still unclear whether the domestication happened initially in Asia and the domesticate then spread or whether the wild ancestor spread first and then got domesticated several times in different places[24,71], as reviewed in ref. 29. In the case of cheese starter bacteria, data from across a larger diversity of continents and traditional cheese varieties will be key to differentiating between a single or a multiple-origin scenario and will probably help push back the date of the onset of diversification of cheese bacteria. In addition, the data we present here do not encompass closely related "wild" strains that might share a very recent common ancestor with starter culture strains. Discovering and characterising these wild relatives will be instrumental to reconstructing the evolutionary history of starter culture and testing the single vs. multiple origin scenarios as well as the degree of hybridisation between domesticated and wild strains. Previous studies have already suggested that these species likely do not have an original niche in the human gut[54,72–74] potentially having a plant origin. However, we speculate here that the ancestral niche of cheese starter bacteria might be the gut of milk-feeding (i.e., juvenile) dairy animals that were originally used as a source for rennet[75].

Methods

Community sampling

The starter cultures were continuously passaged at Agroscope (Liebefeld, Switzerland) as undefined starter cultures. The process of maintaining an undefined starter culture is explained extensively in ref. 41. In short, a stock culture is aliquoted into several samples, which are used on a weekly basis to fulfil the demand of cheese makers. The initial stock is maintained as a freeze-dried ampule at 4 °C and is only passaged and aliquoted if necessary. The number of passages is irregular and not documented. In general, the stock cultures were passaged more frequently from 1970 to approximately 2000, as frequently as every month since we have numerous samples dated one month apart. Today, the frequency of passaging is reduced to a minimum, which can be as rare as every 5 years. The historic samples were stored at 4 °C throughout the years, and DNA isolation was done collectively in 2019. The DNA isolation was done as previously described[76]. In short, the DNA was isolated with the EZ1 DNA Tissue kit on the BioRobot EZ1 robot (Qiagen, Hombrechtikon, Switzerland).

Strain isolation and bacterial counts

Strains were isolated in 1975 and 2019 (Fig. 1B). The isolates from 1975 were picked from MR11 plates (for both species) and stored in the strain collection of the Agroscope. The isolates from 2019 were isolated, as previously explained in Somerville et al. 2022[41]. In short, all 11 starter cultures were plated on two selective media SPY9.3[77] and MR11 (MRS adjusted to pH 5.4 according to ISO7889) for *S. thermophilus* and *L. delbrueckii*, respectively. Ninety-six colonies per species were randomly picked and cultured in liquid media for 24 h at 37 °C. For genotyping, DNA from 100 µL of culture was extracted using the EtNa DNA isolation method[78], and mini-satellite PCR for strain identification of *S. thermophilus* and *L. delbrueckii* was done as explained previously[41]. In short, the length of a mini-satellite region in the two species was evaluated by quantifying the PCR length (all primers in Supplementary Data 2) on a Fragment Analyser™ (Advanced Analytical Technologies, Ankeny, IA, USA). Moreover, colony-forming units (CFU/ml) were determined by serial dilution and plate counting with an Eddy Jet Spiral Plater and SphereFlash Automatic Colony Counter (both from IUL, Barcelona, Spain). A full description, including all primers used, can be found in the supplemental methods.

Metagenome and genome sample preparation and sequencing

Ninety-eight samples, including historic freeze-dried ampules, present working stocks and starter cultures, were prepared for shotgun metagenome sequencing. The DNA was isolated as previously explained[41], and Nextera flex libraries were prepared and subjected to HiSeq4000 150PE (Illumina) sequencing at the Genomic Technologies Facility in Lausanne, Switzerland. Further, a subset of genomic samples was sequenced on a minION (Nanopore) with a rapid barcoding kit.

Raw read analysis

The raw reads for both metagenomic and genomic samples were handled similarly. All adaptors and barcodes were removed with trim galore (v0.6.10)[79]. Reads mapping to the cow genome from the milk was removed with KneadData (v0.7.3) (https://bitbucket.org/biobakery/kneaddata). The reads were mapped with bwa mem (bwa-0.7.18-r1243)[80]. The genomes and metagenomes were assembled with SPAdes (v3.13.1)[81] for short reads and Flye for long reads (v2.9.3-b1797)[82]. Extensive genome polishing was done with Racon (v1.3.3)[83] and freebayes (v1.3.7)[84].

Genome analysis

The genome assemblies were submitted to NCBI and annotated with PGAP[85]. In addition, eggnog mapper was used to identify annotations of pseudogenes[86]. The pairwise ANI values were calculated with fastANI (v1.33)[87]. Additional genomes and their respective PGAP of the three focal species were downloaded from NCBI RefSeq (12.01.2020). Only completely assembled genomes were used as we have previously observed a substantial number of genes (and pseudogenes) not being assembled with Illumina only assemblies due to the repetitive nature of the tranposase-rich genomes[67]. Orthofinder (v2.3.1) was used to identify single-copy core genes within the species[88].

Metagenome analysis

The metagenomic raw reads were profiled for species abundance with mOTU2 (v0.3.2)[45]. In addition, the metagenomic reads were mapped against reference genomes of the three focal species with bwa mem (bwa-0.7.18-r1243)[89]. We performed a single nucleotide analysis (SNV) with freebayes-parallel (v1.3.7) on the genomes and metagenomes in comparison to random reference genomes[84]. The observed SNVs were filtered with vcftools (v0.1.16)[90] and SNPeffect (v4.3t)[91] to include only SNVs with a minimum allele frequency of 0.05, read coverage of 5 and in single-copy-genes (identified with OrthoFinder). The metagenomic SNVs were compared to the previously identified SNVs from the reference genomes. Sub-species clade frequency was calculated by averaging the allele frequency of all sub-species clade-specific SNVs (for details, see script).

Propagation experiment and measurements

For the propagation experiment, we created a pooled starting sample consisting of one random isolate of *L. delbrueckii* and one of *S.thermophilus* per starter culture (in total 22 isolates) (see selection in Supplementary Data 1). The starting sample was propagated in five replicates as described in Somerville et al. 2022[41]. In short, the samples were propagated to simulate the production of starter cultures. We conducted two passages per week. On the first day, 100 µl of the freeze-dried sample was inoculated into 10 ml autoclaved organic milk media (BM) and incubated for 18 h at 37 °C. For the second passage on the next day, the pre-culture was inoculated into 10 ml autoclaved BM and incubated for another 18 h at 37 °C. For the final step, 100 µl of the incubated samples were transferred into a freeze-dry ampule and stored at − 30 °C for at least 1 h. Thereafter, the samples were freeze-dried for 7 hours until dry. For pH measurements, we used the hydroplate system (PreSens, Germany). The pH was normalised with pH standards of pH 4 and 7. The measurements were done in four replicates for 30 h at 37 °C.

Species-level lactic acid bacteria phylogeny

All representative genomes from lactic acid bacteria species were downloaded from NCBI RefSeq. Moreover, from the genus *Streptococcus* and *Lactobacillus*, all genomes deposited on NCBI before 12.01.2020 were included. The annotations were screened for Galactosidase (lacZ and lacG). The phylogeny was reconstructed by concatenating the 16S, 23S and 5S rRNA sequence of the genomes, making a multi-sequence-alignment file with mafft-linsi (v7.526)[92] and calculating the phylogeny with RaxML (v8.2.12) and the "GTRCAT" model[93]. The plot was created in R with ggtree[94].

Strain-level dated whole genome phylogeny

The preliminary species tree was calculated with OrthoFinder (v2.3.1)[88]. Therefore, we back-translated the single copy core genes into the nucleotide space and created a core genome species tree and gene trees as described in ref. 95 using MAFFT (v7.526)[96] and RAxML (v8.2.12) with 100 bootstrap rounds and the GTRCAT[93]. Moreover, by including the gene tree and the species tree we calculated the clonal species tree with ClonalFrameML (v1.12)[97]. Therefore a dated phylogeny was calculated with BactDating (v1.1.2) and 10^7 Bayes repetitions using our historical samples as time calibration points[57]. From the subsequent phylogeny, we predicted sub-species clades for our genome isolates with poppunk (v1.2.0)[50]. In order to quantify the presence of sub-species clades in the metagenomes, we identified all sub-species clade-specific SNVs in the core genes. The dated phylogeny and the information from the isolation source (Supplementary Data 1) were used to reconstruct ancestral habitat reconstruction with the ace function from the ape package[98].

Transcriptome analysis

Samples for transcriptome analysis were collected throughout the first 24 hours of a regular gruyere-type cheese making process at the cheese pilot plant at Agroscope (Liebefeld, Switzerland). The samples were immediately stored in liquid nitrogen and the RNA extraction was carried out with the Qiagen EZ1 extraction robot and the RNA Tissue kit. Illumina libraries were prepared with the TruSeq Str-RNA Zero, and sequencing was performed on a 150 PE HiSeq 4000 (Illumina) at the Genomic Technologies Facility in Lausanne, Switzerland. The Illumina sequences were cleaned with trim galore (v0.6.10)[79] and sortmeRNA (v2.1)[99]. Mapping of reads to isolate reference genomes was performed with bwa mem (bwa-0.7.18-r1243)[89] and gene counting with HT-seq (v0.11.2)[100]. Further, sample and gene normalisation was performed with DESeq2[101].

Phenotypic assays of isolates

The carbohydrate utilisation profiles were determined using Biolog™ phenotypic microarray plates PM01 and PM02A. Samples were prepared according to the manufacturer's protocols. Plates were incubated in Omnilog™ for 48 h at 37 °C. The equipment records the contrast difference every 15 min to generate growth curves. This data was evaluated using Biologs data acquisition software™ and the opm package in ref. 102.

Phenotypic assays of starter cultures

The starter cultures were regularly monitored as part of the regular in-house quality control at Agroscope starting in 1996 (acidity, lactate ratio) and 2000 (total lactate) until 2020 (last measurement included in the manuscript). Specifically, (1) the sample acidity was measured weekly, (2) the total lactate and (3) the lactate ratio was measured monthly. The acidity of the cultures was determined as follows: 10 mL reconstituted skim milk was inoculated using 0.1 ‰ (10 µl) of the corresponding culture and incubated for 18 h at 38 °C. 1 drop of phenolphthalein was added to the sample and then titrated with 0.1 M NaOH till a visible colour change was detected. The recorded volume of 0.1 M NaOH in mL was multiplied by 10 and rounded to the whole number resulting in the determined °Th (or Clark degree). Total lactate (D- and L-Lactate) was analysed enzymatically according to the instruction protocol of the kit manufacturer (Boehringer, Manheim, Germany) using an automated spectrophotometric analyser (Gallery, Thermo, Switzerland). The proportion of L-lactic acid to total lactic acid was calculated as a percentage. This method has previously been published[103]. The linear mixed effects model was fitted into the phenotypic data with the lmer function in the lme4 package[104].

Statistics and reproducibility

No statistical methods were used to predetermine sample size, but rather all available samples were used. The experiments and lab work were, wherever possible, randomised and blinded. If not mentioned otherwise, the data analyses and statistics were done in R (R Core Team, 2020) and plotting was done using ggplot2[105]. The wilcox.test function from the rstatix package and all other tests (*t* test and Pearson chi-square) from the base package was used. All details and specific parameters can be retrieved in the available code (see "Code availability" section). Moreover, the version numbers for all tools used in this study are provided in the supplemental methods.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequence data and, if applicable, assembled genomes that support the findings of this study have been deposited in on NCBI with the BioProject codes: PRJNA717134 (Illumina only genomes), PRJNA717134, PRJNA1083966 (ONT and Illumina genomes), PRJNA1048529 (shotgun metagenomics) and PRJNA1157897 (RNA-seq). In addition, all data frames, intermediate figures and outputs are available on Zenodo (https://doi.org/10.5281/zenodo.10783580).

Code availability

The complete code is available on Git Hub (https://github.com/Freevini/Starter-Culture-diversity) or as a permanent version on Zenodo (https://doi.org/10.5281/zenodo.10783580).

References

1. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).
2. Purugganan, M. D. What is domestication? *Trends Ecol. Evol.* **37**, 663–671 (2022).
3. Driscoll, C. A. & Macdonald, D. W. O'Brien SJ. From wild animals to domestic pets, an evolutionary view of domestication. *Proc. Natl. Acad. Sci. USA* **106**, 9971–9978 (2009).
4. Lazaridis, I. et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
5. Weisdorf, J. L. From foraging to farming: Explaining the Neolithic Revolution. *J. Econ. Surv.* **19**, 561–586 (2005).
6. Yousef, E. A. A., Müller, T., Börner, A. & Schmid, K. J. Comparative analysis of genetic diversity and differentiation of cauliflower (Brassica oleracea var. botrytis) accessions from two ex situ genebanks. *PLoS ONE* **13**, e0192062 (2018).
7. Pitt, D. et al. Domestication of cattle: Two or three events? *Evol. Appl.* **12**, 123 (2019).
8. Ellis, E. C. Anthropogenic transformation of the terrestrial biosphere. *Philos. Trans. A Math. Phys. Eng. Sci.* **369**, 1010–1035 (2011).
9. Milla, R. et al. Phylogenetic patterns and phenotypic profiles of the species of plants and mammals farmed for food. *Nat. Ecol. Evol.* **2**, 1808–1817 (2018).
10. Meyer, R. S., DuVal, A. E. & Jensen, H. R. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *N. Phytol.* **196**, 29–48 (2012).
11. Salque, M. et al. Earliest evidence for cheese making in the sixth millennium BC in northern Europe. *Nature* **493**, 522–525 (2013).
12. Wilkin, S. et al. Dairy pastoralism sustained eastern Eurasian steppe populations for 5000 years. *Nat. Ecol. Evol.* **4**, 346–355 (2020).
13. Evershed, R. P. et al. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature* **455**, 528–531 (2008).
14. Greenfield, H. J. & Arnold, E. R. Go(a)t milk? New perspectives on the zooarchaeological evidence for the earliest intensification of dairying in south eastern Europe. *World Archaeol.* **47**, 792–818 (2015).
15. Marco, M. L. et al. The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on fermented foods. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 196–208 (2021).
16. Gänzle, M. The periodic table of fermented foods: limitations and opportunities. *Appl. Microbiol. Biotechnol.* **106**, 2815–2826 (2022).
17. Castro, G. R., Nadda, A. K., Nguyen, T. A., Sharma, S. & Bilal, M. *Nanomaterials for Bioreactors and Bioprocessing Applications*. (Elsevier, 2023).
18. Gibbons, J. G. & Rinker, D. C. The genomics of microbial domestication in the fermented food environment. *Curr. Opin. Genet. Dev.* **35**, 1–8 (2015).

19. Warinner, C. An archaeology of microbes. *J. Anthropol. Res.* **78**, 420–458 (2022).
20. Swanson, K. S. et al. The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of synbiotics. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 687–701 (2020).
21. Arias-Sánchez, F. I., Vessman, B. & Mitri, S. Artificially selecting microbial communities: If we can breed dogs, why not microbiomes? *PLoS Biol.* **17**, e3000356 (2019).
22. Douglas, G. L. & Klaenhammer, T. R. Genomic evolution of domesticated microorganisms. *Annu Rev. Food Sci. Technol.* **1**, 397–414 (2010).
23. Dunn, R. R., Wilson, J., Nichols, L. M. & Gavin, M. C. Toward a global ecology of fermented foods. *Curr. Anthropol.* **62**, S220–S232 (2021).
24. Peter, J. et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).
25. Gonçalves, M. et al. Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26**, 2750–2761 (2016).
26. Gallone, B. et al. Domestication and divergence of saccharomyces cerevisiae beer yeasts. *Cell* **166**, 1397–410 (2016).
27. Gibbons, J. G. et al. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus Aspergillus oryzae. *Curr. Biol.* **22**, 1403–1409 (2012).
28. Ropars, J. et al. Domestication of the emblematic white cheese-making fungus Penicillium camemberti and its diversification into two varieties. *Curr. Biol.* **30**, 4441–53 (2020).
29. Steensels, J., Gallone, B., Voordeckers, K. & Verstrepen, K. J. Domestication of industrial microbes. *Curr. Biol.* **29**, R381–R393 (2019).
30. Landis, E. A. et al. The diversity and function of sourdough starter microbiomes. *Elife* **10**, e61644 (2021).
31. Kelly, W. J., Ward, L. J. H. & Leahy, S. C. Chromosomal diversity in Lactococcus lactis and the origin of dairy starter cultures. *Genome Biol. Evol.* **2**, 729–744 (2010).
32. Marcobal, A. M., Sela, D. A., Wolf, Y. I., Makarova, K. S. & Mills, D. A. Role of hypermutability in the evolution of the genus Oenococcus. *J. Bacteriol.* **190**, 564–570 (2008).
33. Smokvina, T. et al. Lactobacillus paracasei comparative genomics: towards species pan-genome definition and exploitation of diversity. *PLoS ONE* **8**, e68731 (2013).
34. Roux, E. et al. The genomic basis of the Streptococcus thermophilus health-promoting properties. *BMC Genom.* **23**, 210 (2022).
35. Hill, C. & Paul Ross R. *Genetic Modification in the Food Industry.* (1998).
36. Bolotin, A. et al. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. *Nat. Biotechnol.* **22**, 1554–1558 (2004).
37. van de Guchte, M. et al. The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution. *Proc. Natl. Acad. Sci. USA* **103**, 9274–9279 (2006).
38. Schmid, M. et al. Comparative genomics of completely sequenced lactobacillus helveticus genomes provides insights into strain-specific genes and resolves metagenomics data down to the strain level. *Front. Microbiol.* **9**, 63 (2018).
39. Alexandraki, V. et al. Comparative genomics of streptococcus thermophilus support important traits concerning the evolution, biology and technological properties of the species. *Front. Microbiol.* **10**, 2916 (2019).
40. Baek, M. G., Kim, K. W. & Yi, H. Subspecies-level genome comparison of Lactobacillus delbrueckii. *Sci. Rep.* **13**, 3171 (2023).
41. Somerville, V. et al. Functional strain redundancy and persistent phage infection in Swiss hard cheese starter cultures. *ISME J.* **16**, 388–399 (2022).
42. Sieuwerts, S., de Bok, F. A. M., Hugenholtz, J. & van Hylckama Vlieg, J. E. T. Unraveling microbial interactions in food fermentations: from classical to genomics approaches. *Appl. Environ. Microbiol.* **74**, 4997–5007 (2008).
43. Sieuwerts, S. et al. Mixed-culture transcriptome analysis reveals the molecular basis of mixed-culture growth in Streptococcus thermophilus and Lactobacillus bulgaricus. *Appl Environ. Microbiol.* **76**, 7775–7784 (2010).
44. Iskandar, C. F., Cailliez-Grimal, C., Borges, F. & Revol-Junelles, A. M. Review of lactose and galactose metabolism in Lactic Acid Bacteria dedicated to expert genomic annotation. *Trends Food Sci. Technol.* **88**, 121–132 (2019).
45. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
46. Smid, E. J. & Lacroix, C. Microbe–microbe interactions in mixed culture food fermentations. *Curr. Opin. Biotechnol.* **24**, 148–154 (2013).
47. Sieuwerts, S. Microbial interactions in the yoghurt consortium: Current status and product implications. *SOJ Microbiol. Infect. Dis.* **4**, 01–05 (2016).
48. Fontana, F. et al. Multifactorial microvariability of the Italian raw milk cheese microbiota and implication for current regulatory scheme. *mSystems* **8**, e0106822 (2023).
49. Kamilari, E., Tsaltas, D., Stanton, C. & Ross, R. P. Metataxonomic mapping of the microbial diversity of Irish and Eastern mediterranean cheeses. *Foods* **11**, 2483 (2022).
50. Lees, J. A. et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).
51. Rodriguez-R, L. M. et al. An ANI gap within bacterial species that advances the definitions of intra-species units. *MBio* **15**, e0269623 (2024).
52. McGovern, P. E. & Hall, G. R. Charting a future course for organic residue analysis in archaeology. *J. Archaeol. Method Theory* **23**, 592–622 (2016).
53. Stott, A. W. et al. Direct dating of archaeological pottery by compound-specific 14C analysis of preserved lipids. *Anal. Chem.* **75**, 5037–5045 (2003).
54. Pasolli, E. et al. Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nat. Commun.* **11**, 2610 (2020).
55. Pickrell, J. How the earliest mammals thrived alongside dinosaurs. *Nature* **574**, 468–472 (2019).
56. Kuo, C. H. & Ochman, H. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol. Direct* **4**, 35 (2009).
57. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).
58. Didelot, X., Sarah Walker, A., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150–162 (2016).
59. Zhao, S. et al. Adaptive evolution within gut microbiomes of healthy People. *Cell Host Microbe* **25**, 656–67 (2019).
60. Goh, Y. J., Goin, C., O'Flaherty, S., Altermann, E. & Hutkins, R. Specialized adaptation of a lactic acid bacterium to the milk environment: the comparative genomics of Streptococcus thermophilus LMD-9. *Micro. Cell Fact.* **10**, S22 (2011).
61. Hottes, A. K. et al. Bacterial adaptation through loss of function. *PLoS Genet.* **9**, e1003617 (2013).
62. Ochman, H. & Moran, N. A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
63. Bogaard, A. et al. Reconsidering domestication from a process archaeology perspective. *World Archaeol.* **53**, 56–77 (2021).

64. Zeder, M. A. The domestication of animals. *J. Anthropol. Res.* **68**, 161–190 (2012).
65. Grizon, A. et al. Genomic characterization of wild lactobacillus delbrueckii strains reveals low diversity but strong typicity. *Microorganisms* **12**, 512 (2024).
66. Somerville, V. et al. Extensive diversity and rapid turnover of phage defense repertoires in cheese-associated bacterial communities. *Microbiome* **10**, 137 (2022).
67. Somerville, V. et al. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol.* **19**, 143 (2019).
68. Oechslin, F. et al. Fermentation practices select for thermostable endolysins in phages. *Mol. Biol. Evol.* **41**, msae055 (2024).
69. El Kafsi, H. et al. Lactobacillus delbrueckii ssp. lactis and ssp. bulgaricus: a chronicle of evolution in action. *BMC Genom.* **15**, 407 (2014).
70. Molinet, J. et al. Wild Patagonian yeast improve the evolutionary potential of novel interspecific hybrid strains for lager brewing. *PLoS Genet.* **20**, e1011154 (2024).
71. Duan, S. F. et al. The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat. Commun.* **9**, 2690 (2018).
72. Walter, J. Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. *Appl. Environ. Microbiol.* **74**, 4985–4996 (2008).
73. Duar, R. M. et al. Lifestyles in transition: evolution and natural history of the genus Lactobacillus. *FEMS Microbiol. Rev.* **41**, S27–S48 (2017).
74. De Filippis, F., Pasolli, E. & Ercolini, D. The food-gut axis: lactic acid bacteria and their link to food, the gut microbiome and human health. *FEMS Microbiol. Rev.* **44**, 454–489 (2020).
75. Fleischmann, W. *Das Molkereiwesen: Ein Buch für Praxis und Wissenschaft. Zugleich als vierter Theil zu Otto-Birnbaun's Lehrbuch der landwirthschaftlichen Gewerbe.* (F. Vieweg & Sohn, 1876).
76. Moser, A., Berthoud, H., Eugster, E., Meile, L. & Irmler, S. Detection and enumeration of Lactobacillus helveticus in dairy products. *Int. Dairy J.* **68**, 52–59 (2017).
77. Shani, N., Isolini, D., Marzohl, D. & Berthoud, H. Evaluation of a new culture medium for the enumeration and isolation of Streptococcus salivarius subsp. thermophilus from cheese. *Food Microbiol.* **95**, 103672 (2021).
78. Vingataramin, L. & Frost, E. H. A single protocol for extraction of gDNA from bacteria and yeast. *Biotechniques* **58**, 120–125 (2015).
79. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10 (2011).
80. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
81. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
82. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long error-prone reads using repeat graphs. **37**, 540–546 (2019).
83. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
84. Garrison, E. & Marth G. Haplotype-based variant detection from short-read sequencing. Preprint at *arXiv* (2012).
85. Tatusova, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
86. Rodríguez Del Río, Á. et al. Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **626**, 377–384 (2023).
87. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
88. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
89. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* (2013).
90. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
91. Baets, G. D. et al. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* **40**, D935–D939 (2012).
92. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
93. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
94. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
95. Ellegaard, K. M. & Engel, P. Genomic diversity landscape of the honey bee gut microbiota. *Nat. Commun.* **10**, 446 (2019).
96. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
97. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLOS Comput. Biol.* **11**, e1004041 (2015).
98. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).
99. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
100. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2014).
101. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
102. Vaas, L. A. I., Sikorski, J., Michael, V., Göker, M. & Klenk, H. P. Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS ONE* **7**, e34846 (2012).
103. Borshchevskaya, L. N., Gordeeva, T. L., Kalinina, A. N., & Sineokii, S. P. Spectrophotometric determination of lactic acid. *J. Anal. Chem.* **71**, 755–758 (2016).
104. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
105. Wilkinson, L. ggplot2: Elegant graphics for data analysis by WICKHAM, H. *Biometrics* **67**, 678–679 (2011).

Acknowledgements

We thank Meral Turgay and Elvira Wagner for the help with the minisatellite PCR, Xavier Didelot for the useful advice for calculating the molecular clock, Herwig Bachmann and Bas Teusink for the many discussions on these topics, Julien Marquis and Johann Weber from the Genomic Technologies Facility in Lausanne, Alban Ramette from the IFIK for the sequencing. The authors would like to thank SNF Doc.Mobility P1LAP3_199476 (V.S.), SNF Postdoc.Mobility P500PB_214419 (V.S.). This work was supported as a part of NCCR Microbiomes, a National Centre of Competence in Research, funded by the Swiss National Science

Foundation (grant number 180575). F.M. thanks Uni. of Lausanne, NCCR Microbiome, SNF and Antoine Guisan.

Author contributions

V.S.: Conceptualisation, formal analysis, Funding Acquisition, Visualisation, Writing – Review & Editing, U.v.A., F.M., and P.E.: Funding Acquisition, Writing, Conceptualisation – Review & Editing. N.T., A.R., L.B., M.H., H.B., N.S., and R.S.S.: Methodology.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52687-7.

Correspondence and requests for materials should be addressed to Vincent Somerville.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.