Sequencing of 185 *Streptococcus thermophilus* and identification of fermentation biomarkers

Wenjun Liu  
Inner Mongolia Agricultural University

Linjie Wu  
School of Mathematical Sciences and Center for Statistical Science, Peking University

Jie Zhao  
Inner Mongolia Agricultural University

Weicheng Li  
Inner Mongolia Agricultural University

Yu Wang  
Inner Mongolia Agricultural University

Huijuan Zheng  
Inner Mongolia Agricultural University

Tiansong Sun  
Inner Mongolia Agricultural University

Heping Zhang  
Inner Mongolia Agricultural University

Ruibin Xi  
School of Mathematical Sciences and Center for Statistical Science, Peking University

Zhihong sun (sunzihong78@163.com)  
Inner Mongolia Agricultural University

Article

**Keywords:** Streptococcus thermophilus, Whole genome sequencing, Antibiotic resistance, Growth score, Acidification

**Posted Date:** September 4th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-61428/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

*Streptococcus (S.) thermophilus* is an important dairy starter in the production of fermented dairy products has important significance, from natural fermentation in the past to industrial production today. While the genetic architecture underlying *S. thermophilus* traits and phenotypes is largely unknown. Here, we sequenced 185 *S. thermophilus* strains, which isolated from natural fermented dairy products of China and Mongolia and using comparative genomic and genome wide association study to provide novel point for genetic architecture underlying its traits and phenotypes. Genome analysis of *S. thermophilus* showed association of phylogeny with environmental and phenotypic features and revealed clades with high acid production potential or with substantial genome decay. A few *S. thermophilus* isolated from areas with high chloramphenicol emissions had a chloramphenicol-resistant gene *CatB8*. Most importantly, we defined a growth score and identified a missense mutation G1118698T located at the gene *AcnA* that were both predictive of acidification capability of *S. thermophilus*. Our findings provide novel insight in *S. thermophilus* genetic traits, antibiotic resistant and predictive of acidification capability which both may had huge help in culture starter screening.

Introduction

*Streptococcus (S.) thermophilus* is a predominant lactic acid bacterium (LAB) with rapid acidification capability and is a major dairy starter used in the milk fermentation and cheese production[1]. The starter strains of fermentation have vast economic value and produce a large amount fermented dairy products every year[2]. Traditional nomads have been making naturally fermented milk for thousands of years and these naturally fermented dairy products contain rich LAB resources[3].

However, previous researches mostly concentrate on genomics of industry strains of *S. thermophilus*. For example, one study showed that most of 47 industry *S. thermophilus* have similar genetic distance, indicated genome stability of industry strains[4]. The genetic architecture underlying natural *S. thermophilus* strains is largely unexplored.

In this study, we isolated and sequenced 185 *S. thermophilus* strains from natural dairy products in traditional pasture of different regions in China and Mongolia, to uncover the genetic evolution of wild strains and to explore the relationship between genotype and fermentation characteristics. Phylogenetic tree of *S. thermophilus* revealed clades with high acid production potential or significant genome decay. We also found 13 *S. thermophilus* which have a chloramphenicol-resistant gene *CatB8* isolated from areas with high chloramphenicol emissions. Besides, we defined a growth score and showed that the growth score could accurately predict acidification capability of *S. thermophilus*. With the growth score, we identified a missense mutation G1118698T by genome wide association study (GWAS) located at the gene *AcnA* that were significantly associated with the acidification. Our findings provided novel insights in *S. thermophilus* genetic traits and identified robust biomarkers screening of culture starter with high acidification capability.
Results

Phenotypic analysis and phylogenetic analysis of S. thermophilus

From natural fermented dairy products in China and Mongolia, we isolated 185 S. thermophilus and found 61 strains having potential high acid production capability (H-Acid) using a preliminary fermentation experiment. Among the 185 S. thermophilus, 65 were from China and 120 were from Mongolia (Fig. 1A). Whole genome sequencing (WGS) was performed for all 185 strains (mean coverage 374X, Table S1). After short read alignment to the reference genome CNRZ1066 of S. thermophilus, we identified 58,734 single nucleic acid polymorphisms (SNPs), 2,246 insertions and deletions (Indels, Table S2) and 450 copy number variations (CNVs, Table S3). As expected, ribosomal RNA (rRNA) and transfer RNA (tRNA) had the lowest mutation rates, followed by protein coding regions, pseudogenes and intergenic regions (Fig. 1B). De novo assembly showed that the genome sizes of S. thermophilus ranged from 1.72 to 2.60 Mb and the numbers of genes ranged from 1,704 to 2,158 (Fig. 1C-D).

Pan-genome analysis of our 185 isolates as well as 32 S. thermophilus available from the NCBI[5] totally identified 7,629 genes including 827 core genes and 315 soft core genes (Fig. 1E, Table S4-S6). The core and soft-core genes were mainly enriched in metabolism pathways like metabolic and biosynthesis of amino acids, which are essential for bacterial growth (Fig. S1A). Glycolysis/gluconeogenesis pathway and starch and sucrose metabolism pathways were enriched in shell genes (shared by 15%-95% samples). These two pathways are involved in carbohydrate utilization. Diversity in these pathways indicates that S. thermophilus strains might have undergone adaptive evolution in carbohydrate utilization pathways. Besides, the cloud genes (shared by less than 15% samples) were mainly enriched in quorum sensing and beta-lactam resistance pathways. 53% of cloud genes were homologous to genes in other Streptococcus strains and 8% to genes in Lactobacillales and Lactococcus, two bacterial species often found in natural fermented dairy products[6] (Fig. S1B). This indicated that a portion of cloud genes came from horizontal gene transfer (HGT).

The phylogenetic tree of the 217 S. thermophilus revealed four large clades and was consistent with multi locus sequence typing (MLST) (Fig. 2) The within-clade average nucleotide identity (ANI) was significantly larger than the between-clade ANI and the principle component analysis also showed that the four clades are well-separated (Fig. S2A-B). Clade A, including strains from China, Mongolia and NCBI database, was closer to the roots of phylogenetic trees. In the other three clades, strains isolated from China and Mongolia also showed clear aggregation. Clade D was basically composed of strains isolated from yoghurts of Mongolia, the strains in Clade C were mainly from yoghurts of Mongolia and Xiniiang China. The clade B was mostly isolated from goat yoghurts in China and yoghurts in Mongolia. Genetic distance between isolates were significantly correlated with the geographical location of sampling sites (Fig. S2C). The dairy product type was also significantly correlated with the phylogenetic clades (Fisher’s exact test, p-value = 5⋅10^{-3}, Table S7).
Interestingly, the H-Acid isolates and NCBI stains were significantly enriched in clade A and B (Fisher’s exact test, p-value = 5.1\cdot 10^{-7}, Fig. 2). Note that strains from NCBI were extensively used in industrial production and generally had high dairy fermentation capability. Furthermore, the cell-wall protease gene PrtS, a gene known to be associated with rapid growth and acidification rates at bacteria in milk[7], was significantly enriched in clade A (Fisher’s exact test, p-value = 1.3\cdot 10^{-28}, Table S8). These data implied that strains in clade A and B might have better fermentation potential than the strains in clade C and D.

**Genome decay and antibiotic resistance of S. thermophilus**

*S. thermophilus* in clade D had significantly fewer number of genes (Fig. 3A), smaller genome sizes and more copy number losses than other clades, indicating that clade D might have undergone considerable genome decay (Fig. 3B-C and Fig. S3A). We found 131 genes that were prevalent in clade A-C (frequency > 0.5) but were significantly depleted in clade D (Fisher’s test, Benjamini–Hochberg adjusted p-value < 0.05, Table S9). These genes were significantly enriched in the pathways including quorum sensing, beta-lactam resistance, and ABC transporters (Table S10). Further, clade D had significantly less quorum sensing genes compared with clade A and B (Fig. 3D). Many of the depleted quorum sensing genes were *Blp* bacteriocins related genes. By comparing with the *Blp* protein family in *Streptococcus pneumoniae* (Table S11), we found that *blpB*, *blpM*, *blpH* and *blpR* genes were significantly lost in strains from clade D comparing with other clades (Fig. 2, Fisher’s test p-value < 0.001). *BlpB* protein, a transport accessory protein, is essential for the secretion of antimicrobial compounds. *BlpM* is a bacteriocin-encoding gene which directly influence the production of bacteriocins. Both *BlpH* and *BlpR* are members of two-component regulatory system, which allow bacteria to sense and respond to changes in different environment conditions. The knockout of *blpB*, *blpH* and *blpR* genes in *S. thermophilus* reduced production of bacteriocins compared with the wild type[8]. These suggest that strains in clade D may have lower production capacity of bacteriocins than strains in other clades. We found that 104 of the 131 depleted genes in D had homologs in 10 other streptococci species and these genes were also enriched in quorum sensing pathway, especially the bacteriocins cluster, followed by ABC transporters (Table S12). This phenomenon was consistent with previous research, which found a striking level of genome decay in *S. thermophilus* compared with other streptococci[9]. Besides, 76 in 131 genes were identified in *Streptococcus salivarius* subspecies *salivarius*, implying that instead of being acquired by strains in clade A-C, the depleted genes in clade D were probably lost in strains from clade D in their adaptation to the milk niche.

Due to the widespread misuse of antibiotics, the problems caused by bacterial resistance have received wide attention. By comparing pan-genome with Comprehensive Antibiotic Resistance Database[10], we found 77 genes related with resistance of 31 different antibiotics (Table S13). Most antibiotic-related genes were associated with efflux (25 of 77, 32.47%) and target alteration (35 of 77, 45.45%). 16 strains in clade C had a gene homologous to the glycopeptides antibiotics resistance protein ARO3002945 (*VanH*) [11]. These 16 strains with *VanH* were mostly isolated from Xinjiang Autonomous Region in China. Three antibiotics resistance-related genes, including two antibiotic efflux genes (ARO3000614 and ARO3004054) and one antibiotic target alteration gene ARO3004253 (*VanU*) [12], are more likely being
lost in clade D instead of being acquired by A-C. Besides, we also found that 13 strains in clade A had a gene homologous to the chloramphenicol resistance protein AR03002680 (CatB8, Fig. 2, Table S14). Most strains with CatB8 were isolated from Hongyuan prairie in Sichuan and Gannan prairie in Gansu, two high chloramphenicol emission provinces of China [13]. We identified nearby transposon sequences around CatB8 (~1 kb and ~5 kb) for 12 out of 13 strains with the CatB8 gene (Table S15), indicating that the CatB8 gene was probably acquired by lateral gene transfer. CatB8, chloramphenicol acetyltransferase, inactivates chloramphenicol by acetylation[14]. The acquisition of CatB8 might confer *S. thermophilus* resistance to chloramphenicol. In fact, we applied droplet digital PCR (ddPCR) for strains with CatB8 and found a significantly higher expression of CatB8 in M17 cultures with 8 µg/ml chloramphenicol than cultures without chloramphenicol (Wilcoxon's test, p-value = 0.0025, Fig. S3B, Table S16), indicating that the bacteria responded to the exposure of chloramphenicol by elevating the expression of CatB8. These data suggested that the misuse of chloramphenicol might be closely related with the antibiotic resistant *S. thermophilus* and attention should be paid in the screening of potential starter to avoid the spread of resistance genes.

**Growth score and acidification of *S. thermophilus***

Acidification is the most important characteristic of *S. thermophilus* as a starter in fermentation of dairy products. Rapid acidification can shorten fermentation time of yoghurt production and reduce the production costs. Screening of *S. thermophilus* strains with rapid acidification is one of most concerned problems in fermentation dairy enterprises. However, currently biomarkers for rapid acidification is still lacking. Theoretically, acidification is closely related with the growth rate of *S. thermophilus*. As most bacteria, *S. thermophilus* has a single circular genome. During replication, DNA sequences passed the replication fork should have two copies and those to be replicated should have only single copy. Thus, because millions of cells at different replication stages were used in WGS of *S. thermophilus*, a genomic region's read depth should be negatively correlated with its distance to the replication origin and the strength of this correlation should reflect the growth rate of *S. thermophilus*. In our WGS data, the adjusted read depths (Methods) were indeed negatively correlated with the distance to the replication origin for all isolates. We defined a growth score as the negative value of Spearman's correlation between them (Fig. 4A-B, Fig S4A, Table S17). Clade A and B strains had significantly larger growth score than clade C and D (Fig. S4B). The growth score was significantly larger in H-Acid strains (Fig. 4C) and in the strains with the PrtS gene (Fig. 4D). These data implied that the growth score might provide an accurate marker for the acidification capability of *S. thermophilus*.

Using the growth score, we performed Genome-Wide Association Study (GWAS) and stability selection to screen for genomic variations that might be related with the acidification capability of *S. thermophilus* (Methods). We found that 7 SNPs were significantly correlated with growth score by GWAS (Wilcoxon test, p-value < 10^{-5} or Bonferroni adjusted p-value < 0.05, Table S18). Among the 7 SNPs, the missense SNP A764991G located at the gene Asnc, which promotes the growth of the *S. thermophilus* in milk by regulating aspartic acid metabolism[15, 16], had the highest selection probability and minimum p-value (1.4⋅10^{-8}) in GWAS analysis. The missense SNP G1118698T located at the gene AcnA, which encodes
aconitate hydratase A, also had high selection probability and was very significant in GWAS analysis (p-value = 4.5 \times 10^{-6}, Fig. S4C-D). AcnA involves in succinate and citrate production, contribute to acidification. 27 S. thermophilus with this mutation were all from clade B and D and were all in the non-H-Acid group (Fig. S4E).

To confirm that the proposed growth score and the SNP G1118698T were associated with the acidification capability of S. thermophilus, we randomly selected 85 strains and performed fermentation experiments (Table S19). We evaluated the acidification capability of S. thermophilus by the acidity and the acid production speed (Methods). The acidity and acid production speed were significantly higher in strains with the PrtS gene, and higher in clade A and B (Fig. S4F-G). The growth score was significantly positively correlated with acid production speed (Pearson's correlation 0.4, p-value = 1.2 \times 10^{-4}, Fig. 4E) and acidity (Pearson's correlation 0.35, p-value = 1.2 \times 10^{-3}, Fig. 4F). Similarly, the SNP G1118698T was significantly associated with acidity and the acidification speed (Fig. 4G). These data suggested that the proposed growth score and the SNP G1118698T could serve as reliable biomarkers for screening S. thermophilus isolates with high acidification speed.

**Discussion**

We reported so far the largest WGS data of S. thermophilus isolated from natural fermentation dairy products. This large amount of data allowed us to systematically investigate the genomics landscape of S. thermophilus. We found that S. thermophilus had four large clades. Strains in clade A seemed to have high industry application potential, while strains in clade D might have undergone considerable genome decay through gene loss. We also identified novel biomarkers for the acidification capability of S. thermophilus. The data and novel discoveries in this paper provided valuable resources for understanding the evolution and genomics of S. thermophilus and for the industry application of S. thermophilus.

Compared with many other streptococci species, S. thermophilus lives in a rather stable environment. Previous researches[17] discovered that many virulence-related genes were lost or became pseudogenes in S. thermophilus. Here, we revealed that in clade D strains of S. thermophilus had significantly smaller genomes than strains in other clades, indicating that the genome decay might be an ongoing process of S. thermophilus. This large genome decay was possibly due to the adaptation of S. thermophilus to their stable niche of milk. In fact, mathematical simulation[18] showed that stable environments often lead to smaller genomes than environments with greater variability. Many quorum-sensing genes, especially blp genes, were lost in clade D strains. By analyzing genomes of three S. thermophilus strains (LMG18311, CNRZ1066, and LMD-9), previous research[8] found that, although identified in all three strains, the blp gene clusters were only full functional in LMD-9. It was thus plausible that these blp genes conferred little or no survival advantages to S. thermophilus, and then the genes in this pathway gradually became inactive and eventually lost clade D S. thermophilus strains.

The traditional method for evaluating the acidification capability of S. thermophilus strains is very labor-intensive, time-consuming and costly. With the advancement of sequencing technologies, WGS becomes
very efficient and cost-effective. The growth score defined in this paper can be calculated only using WGS data and thus provided a very convenient and cost-effective surrogate to the traditional evaluation method. With this score, one can easily screen hundreds of *S. thermophilus* strains. In addition, since fast growth is often a desirable property for industrial bacteria, the growth score defined here might also serve as a robust criterion for evaluating other industrial bacteria and thus has a large application potential in industry. One disadvantage of this growth score is that it currently can only be used for monoculture bacteria. In industrial applications, multiple species of bacteria are often simultaneously used. WGS of the mixture of the different bacteria would cause read mapping ambiguity to different reference genomes and thus the growth score cannot be directly applied. However, we could only consider the genomic sequences that are unique to each species and generalize the growth score using these unique sequences for the mixed sequencing data.

**Materials And Methods**

**Variant calling, assemble and annotation**

Genomic DNA was sequenced using an Illumina HiSeq 4000 platform (Illumina, San Diego, CA) generating 150-bp paired-end reads with an average insert size of 350 bps. All 185 *S. thermophilus* sequencing data were mapped to reference genome CNRZ1066 by BWA-mem[19] with default parameters. SNPs and Indels were called by GATK UnifiedGenotyper[20] and annotated by SnpEff[21]. SNPs having Indels within its 10 bp neighborhood were filtered. CNVs were called by CNV-BAC[22]. We performed *de novo* assembly using SOAPdenovo2[23] (k-mer = 71). The contigs were than annotated by Prokka[24]. Roary[25] was used for the pan-genome analysis. Core genes were defined as genes shared by all strains, soft core genes shared by at least 95% strains, shell genes shared by 15%-95% strains and cloud genes shared by less than 15% strains. The origin of gene sequences in pan-genome were identified by comparing with nr database using blastp[26]. We used the species with highest bitscore and longest alignment length, identity > 40% and e-value < 10^{-6} as the final origin for each gene.

**Phylogenetic Analyses**

We used the *Streptococcus salivarius* CP013216 as an outgroup strain in the phylogenetic analysis. We aligned the outgroup strain genome, the 32 *S. thermophilus* genomes in the NCBI database, as well as the 185 assembled *S. thermophilus* sequences to the reference genome CNRZ1066 using the algorithm MumMer[27]. Neighbor-Joining tree was first generated using MEGA7[28] with default parameters. Then, we used ClonalFrameML[29] with NJ tree and alignment sequences to reconstruct the tree to remove influence of recombination. Among the genes that were prevalent in clades A-C (frequency > 0.5 in at least one of clades A-C) but less prevalent in clade D (frequency < 0.5 in clade D), we used Fisher’s exact test to identify genes significantly depleted genes in clade D. The genes with Benjamini–Hochberg adjusted p-value < 0.05 and odds ratio > 1.5 were selected. This gave use 158 genes.

**Proteolysis and antibiotic resistance genes**
We compared annotated genes to the reference sequence of proteolysis genes from NCBI database using blastp[26]. We kept the alignments with e values less than $10^{-5}$ and bit scores larger than 30. Antibiotic resistance genes were identified by comparing annotated genes with the sequences in the Comprehensive Antibiotic Resistance database[10].

**Calculation of growth score and GWAS analysis**

We first normalized the read depth by considering local GC content and the mappability of short reads by BIC-seq2[30]. The adjusted read depth was calculated in 1000 bp bins as the ratio between the observed read count in the bin and the expected read count given by BIC-seq2. The replication origin of the reference CNRZ1066 was obtained from the DoirC database[31]. For each strain, we calculated the Spearman correlation between the bin's adjusted read depth and its distance to the replication origin. For the GWAS analysis, we first performed a principle component analysis (PCA) based on SNPs and Indels with allele frequencies within (0.05, 0.95) and genes whose occurrence frequencies were in (0.05, 0.95). We then performed a linear regression using the growth score as the response variable and the first two PCA components as the covariates and calculated the residuals of the linear regression for each strain. This step was to remove potential confounding factors (such as hidden population structure) that might influence the growth score. Finally, we performed Wilcoxon's rank test to identify nonsynonymous SNPs and Indels that were significantly correlated with the growth score residuals. For the stability selection, we first filtered the SNPs by controlling the false discovery rate less than 0.05. This gave us 690 SNPs. We then performed stability selection[32] using the lasso regression.

**Fermentation experiment**

In the preliminary acidification experiments, the *S. thermophilus* were inoculated into reconstituted skimmed milk. After 12h fermentation, titratable acidity was measured. Strains with curd time less than 12h and titratable acidity above 55 °T were defined as high acid production capability (H-Acid). The rest strains were defined as non-H-Acid group. Thus, we distinguish the 185 *S. thermophilus* strains into two groups preliminary. To test the acidification capability of *S. thermophiles*, strains from frozen stock were reactivated at 37°C in M17 Broth (Oxoid) and subcultured twice at 24 h before use. Milk was prepared by adding 6% sucrose to 11.5 % reconstituted skimmed milk, which was then sterilized at 95 °C for 10 min and cooled to 42 °C before inoculation (about 6 log10 cfus mL$^{-1}$ for each strain). Fermentation allowed to proceed at 42 °C until fermentation completed. The fermentation experiment was performed in triplicate. PH and titratable acidity (TA, °T) were measured in triplicate to evaluate fermentation progress. The pH was evaluated by pH meter (Mettler Toledo, Switzerland). Titratable acidity was measured using the method described in National Standards of the People's Republic of China. Each sample (5.0 g) was mixed with 4.5 ml of distilled water and titrated with 0.1N NaOH in the presence of 0.5% phenolphthalein indicator to an end point of faint pink color.

**Minimum Inhibitory Concentration of Chloramphenicol**
We selected and reactivated twelve *S. thermophilus* isolates at 37°C in M17 Broth (Oxoid) before use. A wide range of chloramphenicol concentration (spanning across a wide concentration range from 0.125 µg/mL to 64 µg/mL achieved by ten-fold dilution) were prepared before use. The minimum inhibitory concentration (MIC) was determined according to ISO Standard 10932:2010. Briefly, bacterial suspensions were diluted by 1000-fold (~3×10^5 cfu/mL) and tested against each chloramphenicol concentration. The MIC was recorded after incubating the bacterial cells for 48 h at 37 °C in strictly anaerobic conditions.

**Chloramphenicol resistance gene expression checked by droplet digital PCR**

The chloramphenicol resistance gene was quantified using QX100 droplet digital PCR (ddPCR, Bio-Rad), with the gene specific primer (Strep-F: 5'-AATGTTTAGCAATGACGGAAGCC-3', Strep-R: 5'-TTCACCAATGTAAATCCCACCAC-3'). Quantification was performed using ddPCR as follows: initially, a final volume of 20 µL reaction mixture containing 2 µL cDNA, 10uL ddPCR Supermix for EvaGreen (Bio-Rad), 0.2 µL forward primer (20mM), 0.2 µL reverse primer (20mM) and 7.6 µL ddH2O were per-mixed; Each 20 µL reaction with 70 µL of droplet generation oil (Bio-Rad) was used to generate droplets; Droplets were generated by a droplet generator (Bio-Rad). The generated droplets with foil seal were then placed on a conventional PCR Thermocycler. After PCR, the PCR plate was loaded on the droplet reader (Bio-Rad), which automatically reads the droplets from each well of the plate. Analysis of the ddPCR data was performed with QuantaSoft analysis software (Bio-Rad) that accompanied the droplet reader.

**Availability of data and materials**

The data for the 185 isolated *S. thermophilus* has been deposited in the NCBI database under the BioProject ID: PRJNA594100. Supplementary tables see https://github.com/XiDsLab/ST185.

**Declarations**

**Competing interests**

All authors declare that there were no conflicts of interest during completion of the current research.

**Author Contributions**

Z.S., H.Z., T.S. and R.X. conceived and designed the experiments. W.L., Y.W. and H.Z. performed the experiments. L.W., J.Z. and W.L. analyzed the data. R.X. supervised all data analysis. R.X., W.L., L.W. and J.Z. drafted the manuscript. All authors read and approved the final manuscript.

**Acknowledgments**
We thank Professor Narisu and Yujun Cui for their suggestions. This research was supported by the National Natural Science Foundation of China (Grant no. 31622043, 3140066, 31771954, 11971039 and 71532001), China Agriculture Research System (Grant No. CARS-36), and National Key Basic Research Project of China (2016YFC0207705).

References

1. Ravyts, F., De Vuyst, L., and Leroy, F. (2012). Bacterial diversity and functionalities in food fermentations. Eng Life Sci 12, 356–367.
2. Douillard, F.P., and de Vos, W.M. (2014). Functional genomics of lactic acid bacteria: from food to health. Micr Cell Fact 13 Suppl 1, S8.
3. Song, Y., Sun, Z., Guo, C., Wu, Y., Liu, W., Yu, J., Menghe, B., Yang, R., and Zhang, H. (2016). Genetic diversity and population structure of Lactobacillus delbrueckii subspecies bulgaricus isolated from naturally fermented dairy foods. Sci Rep 6, 22704.
4. Rasmussen, T.B., Danielsen, M., Valina, O., Garrigues, C., Johansen, E., and Pedersen, M.B. (2008). Streptococcus thermophilus core genome: comparative genome hybridization study of 47 strains. Appl Environ Microbiol 74, 4703–4710.
5. Sayers, E.W., Agarwala, R., Bolton, E.E., Brister, J.R., Canese, K., Clark, K.,Connor, R., Fiorini, N., Funk, K., Hefferson, T., et al. (2019). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 47, D23-D28.
6. Macori, G., and Cotter, P.D. (2018). Novel insights into the microbiology of fermented dairy foods. Curr Opin Biotech 49, 172–178.
7. Hols, P., Hancy, F., Fontaine, L., Grossiord, B., Prozzi, D., Leblond-Bourget, N., Decaris, B., Bolotin, A., Delorme, C., Ehrlich, S.D., et al. (2005). New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics. Fems Microbiol Rev 29, 435–463.
8. Fontaine, L., Boutry, C., Guedon, E., Guillot, A., Ibrahim, M., Grossiord, B., and Hols, P. (2007). Quorum-sensing regulation of the production of blp bacteriocins in Streptococcus thermophilus. J Bacteriol 189, 7195–7205.
9. Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D., et al. (2004). Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nat Biotechnol 22, 1554–1558.
10. Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., and Sharma, A.N. (2016). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res, gkw1004.
11. Handwerger, S., Pucci, M.J., Volk, K.J., Liu, J.P., and Lee, M.S. (1992). The Cytoplasmic Peptidoglycan Precursor of Vancomycin-Resistant Enterococcus-Faecalis Terminates in Lactate. Journal of Bacteriology 174, 5982–5984.
12. Depardieu, F., Bonora, M.G., Reynolds, P.E., and Courvalin, P. (2003). The vanG glycopeptide resistance operon from Enterococcus faecalis revisited. Mol Microbiol 50, 931–948.

13. Zhang, Q.Q., Ying, G.G., Pan, C.G., Liu, Y.S., and Zhao, J.L. (2015). Comprehensive Evaluation of Antibiotics Emission and Fate in the River Basins of China: Source Analysis, Multimedia Modeling, and Linkage to Bacterial Resistance. Environ Sci Technol 49, 6772–6782.

14. Schwarz, S., Kehrenberg, C., Doublet, B., and Cloeckaert, A. (2004). Molecular basis of bacterial resistance to chloramphenicol and florfenicol. Fems Microbiol Rev 28, 519–542.

15. Kolling, R., Gielow, A., Seufert, W., Kucherer, C., and Messer, W. (1988). Asnc, a Multifunctional Regulator of Genes Located around the Replication Origin of Escherichia-Coli, Oric. Mol Gen Genet 212, 99–104.

16. Arioli, S., Monnet, C., Guglielmetti, S., Parini, C., De Noni, I., Hogenboom, J., Halami, P.M., and Mora, D. (2007). Aspartate biosynthesis is essential for the growth of Streptococcus thermophilus in milk, and aspartate availability modulates the level of urease activity. Appl Environ Microb 73, 5789–5796.

17. Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D., et al. (2004). Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nature Biotechnology 22, 1554–1558.

18. Bentkowski, P., Van Oosterhout, C., and Mock, T. (2015). A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. Genome Biol Evol 7, 2344–2351.

19. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

20. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kemytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303.

21. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80–92.

22. Wu, L., Wang, H., Xia, Y., and Xi, R. (2020). CNV-BAC: Copy Number Variation Detection in Bacterial Circular Genome. Bioinformatics 36, 3890–3891.

23. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18.

24. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.

25. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693.

26. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421.
27. Delcher, A.L., Salzberg, S.L., and Phillippy, A.M. (2003). Using MUMmer to identify similar regions in large sequence sets. Current protocols in bioinformatics, 10.13. 11-10.13. 18.
28. Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33, 1870–1874.
29. Didelot, X., and Wilson, D.J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS computational biology 11, e1004041.
30. Xi, R.B., Lee, S., Xia, Y.C., Kim, T.M., and Park, P.J. (2016). Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res 44, 6274–6286.
31. Luo, H., and Gao, F. (2019). DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. Nucleic Acids Res 47, D74-D77.
32. Meinshausen, N., and Buhlmann, P. (2010). Stability selection. J R Stat Soc B 72, 417–473.

Figures

![Figure 1](image-url)
Descriptive statistics of 185 S. thermophilus. (A) Geographical distribution of 185 samples. (B) Boxplots of mutation rates for different genomic elements (SNPs and Indels). Conservative regions such as rRNA and tRNA showed smaller mutation rates. (C-D) The distribution of genome sizes and number of genes of the S. thermophilus strains (E) The number of genes for different gene type in pan-genome. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 2

The phylogenetic tree as well as various genotypes and phenotypes of the S. thermophilus. The sampling locations, the dairy products from which the S. thermophilus was isolated and the pollution level of chloramphenicol are also shown.
Figure 3

Genome decay. (A-D) The boxplots of the number of genes, genome sizes, lengths of deletion regions and number of genes in quorum sensing for isolates from different clades, where p-values were calculated by Wilcoxon’s test.
Figure 4

Growth score and acidification. (A) Density plot of growth score (B) The smoothed adjusted read depth for all isolates (right) and the raw adjusted read depth for three example isolates (left). (C) Boxplots of growth scores of H-Acid and Non-H-Acid isolates. (D) Boxplot of growth scores in strains with or without the PrtS gene. (E-F) Scatter plot of acidity and speed of acid production versus the growth score in 85 strains. (G) Boxplots of acidity (left) and speed of acid production (right) in strains with or without the SNP G1118698T.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx
- FigureS4.pdf
- FigureS1.pdf
- FigureS3.pdf
- FigureS2.pdf
- DatasetS1.xls
- DatasetS2.xls
- DatasetS3.xls
- DatasetS4.xls
• DatasetS5.xls
• DatasetS6.xls