Supplementary Information for

Sequencing of 185 Streptococcus thermophilus and identification of fermentation biomarkers

Wenjun Liu[1,†], Linjie Wu[2,†], Jie Zhao[1,†], Weicheng Li[1], Yu Wang[1], Huijuan Zheng[1], Tiansong Sun[1], Heping Zhang[1], Ruibin Xi[2,*], Zhihong Sun[1,*]

1.Key Laboratory of Dairy Biotechnology and Engineering, Ministry of Education; Key Laboratory of Dairy Products Processing, Ministry of Agriculture and Rural Affairs; Inner Mongolia Key Laboratory of Dairy Biotechnology and Engineering; Inner Mongolia Agricultural University, Hohhot 010018, China.

† These authors contributed equally to this work.

* Corresponding author: Ruibin Xi, Zhihong Sun.

Email: ruibinxi@math.pku.edu.cn, sunzhihong78@163.com.

This PDF file includes:

Supplementary Information Text

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(1) with default parameters. SNPs and Indels were called by GATK Unifiedgenotyper(2) and annotated by SnpEff(3). SNPs having Indels within its 10bp neighborhood were filtered. CNVs were called by CNV-BAC(4). We performed *de novo* assembly using SOAPdenovo2(5) (k-mer =71). The contigs were than annotated by Prokka(6). Roary(7) 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 than15% strains.

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(8). Neighbor-Joining tree was first generated using MEGA7(9) with default parameters. Then, we used ClonalFrameML(10) 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(11). 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(12).

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(13). 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(14). 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(15) 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℃ 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℃ 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 (~3×10^5cfu/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.

Supplementary Figures

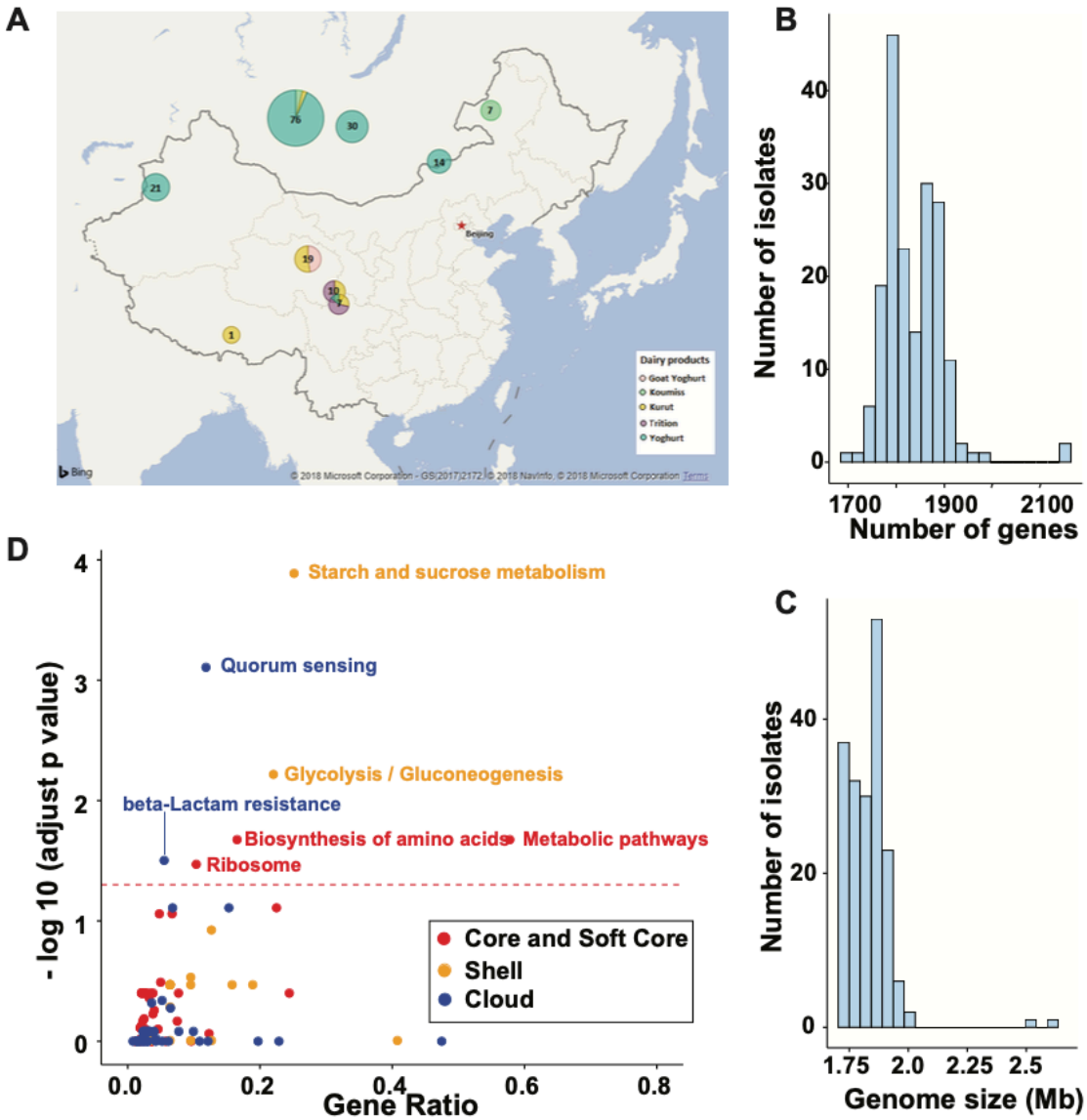

Fig. S1. Sample distribution and pan-genome analysis. (**A**) Distribution of 185 natural fermented dairy products. (**B-C**) Number of genes and genome size of assembled sequences. (**D**) Enriched KEGG pathway of pan-genome.

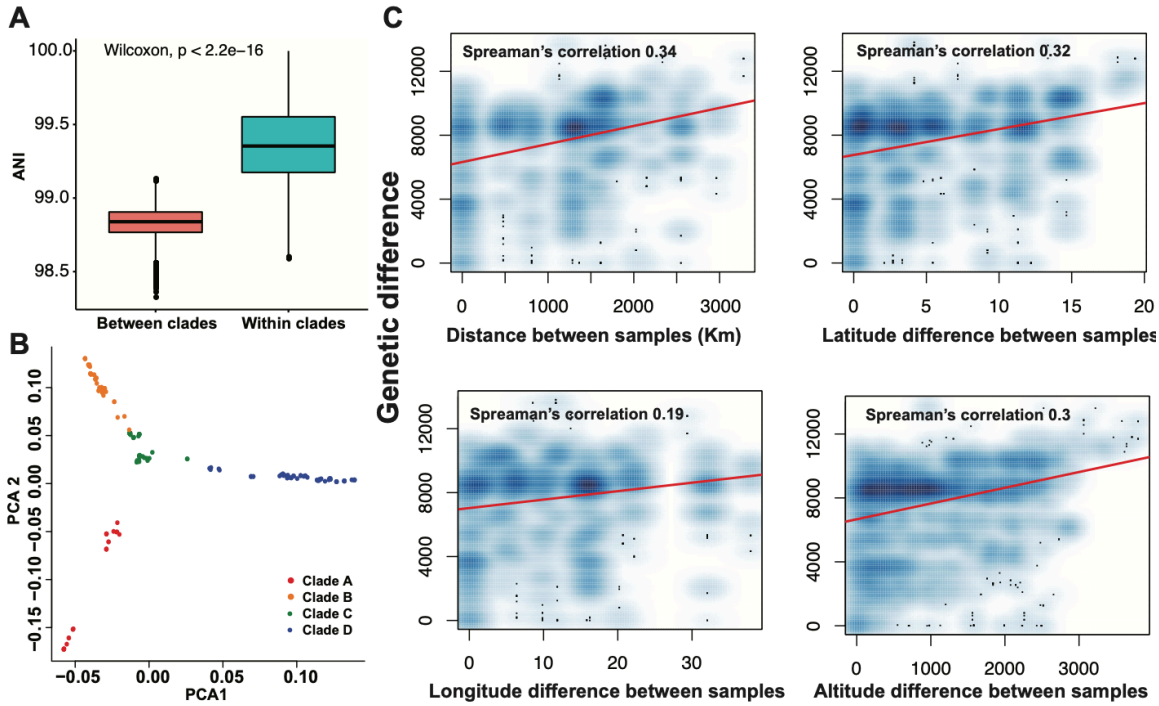

Fig. S2. Genetic difference between 185 isolates. (**A**) Average nucleotide identity (ANI) between clades are significantly smaller than within clades. (**B**) Dimension reduction by principle component analysis for SNPs of 185 isolates. (**C**) Genetic differences (measured by the number of different SNPs between isolates) are related with geographical distance, latitude, longitude and altitude of sampling sites.

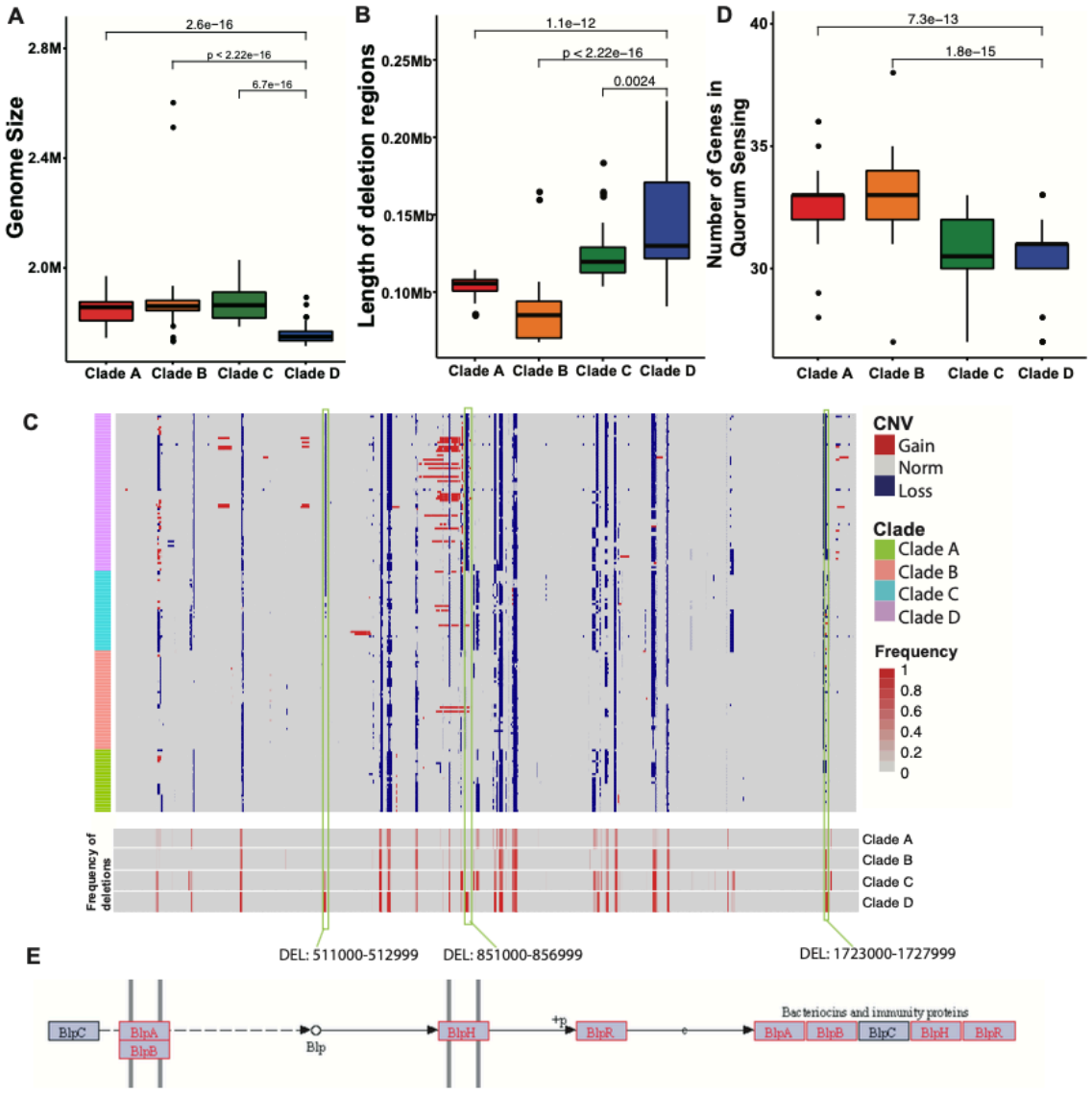

Fig. S3. Genome decay of clade D (**A-B**) The boxplots of genome sizes and lengths of deletion regions for isolates from different clades. (**C**) CNVs for each strain (top) and the frequency of deletions for each clade (bottom). (**D**) Boxplots of numbers of genes in quorum sensing pathway for isolates in different clade. (**E**) Many genes in the Blp bacteriocin pathway are depleted in clade D (marked in red).

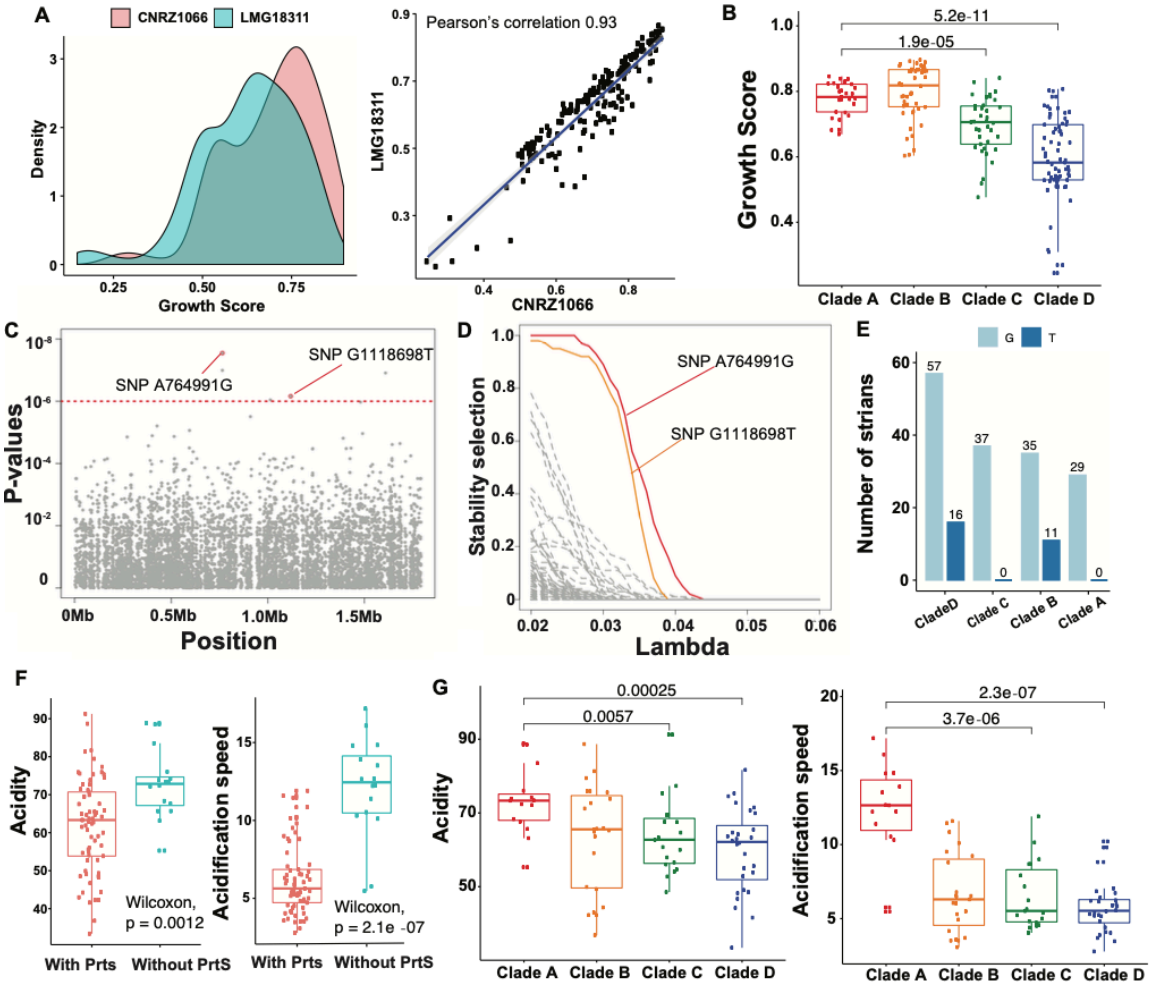

Fig. S4. Growth score and acidification of Streptococcus thermophilus. (**A**) Growth score calculated with different references showed high similarity. (**B**) Boxplots of growth scores in different clades. (**C**) The Manhattan plot of the GWAS analysis. (**D**) Selection path from the stability selection. SNP A764991G and G1118698T had the highest selection probabilities. (**E**) Number of strains with/without the SNP G1118698T in different clades. (**F**) Boxplots of acidity and acidification speed between isolates with or without the gene *PrtS*. (**G**) Boxplots of acidity and acidification speed in different clades.

SI References

1. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
2. A. McKenna *et al.*, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).
3. P. Cingolani *et al.*, 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 (2012).
4. L. Wu, H. Wang, Y. Xia, R. Xi, CNV-BAC: Copy Number Variation Detection in Bacterial Circular Genome. *bioRxiv* (2019).
5. R. Luo *et al.*, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
6. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
7. A. J. Page *et al.*, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).
8. A. L. Delcher, S. L. Salzberg, A. M. Phillippy, Using MUMmer to identify similar regions in large sequence sets. *Current protocols in bioinformatics*, 10.13. 11-10.13. 18 (2003).
9. S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**, 1870-1874 (2016).
10. X. Didelot, D. J. Wilson, ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology* **11**, e1004041 (2015).
11. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
12. B. Jia *et al.*, CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*, gkw1004 (2016).
13. R. B. Xi, S. Lee, Y. C. Xia, T. M. Kim, P. J. Park, 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 (2016).
14. H. Luo, F. Gao, DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. *Nucleic Acids Res* **47**, D74-D77 (2019).
15. N. Meinshausen, P. Buhlmann, Stability selection. *J R Stat Soc B* **72**, 417-473 (2010).