PLOS ONE

RESEARCH ARTICLE

Whole-Genome Sequencing and Comparative Genome Analysis of *Bacillus subtilis* Strains Isolated from Non-Salted Fermented Soybean Foods

Mayumi Kamada[1¤], Sumitaka Hase[1], Kazushi Fujii[2], Masato Miyake[1], Kengo Sato[1], Keitarou Kimura[3], Yasubumi Sakakibara[1]*

1 Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan, **2** Department of Biological Sciences, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan, **3** Division of Applied Microbiology, National Food Research Institute, 2-1-12 12 Kannondai, Tsukuba, Ibaraki 305-8642, Japan

¤ Current address: Department of Clinical Systems Onco-Informatics, Graduate School of Medicine, Kyoto University, 54 Shogoin-Kawaharacho, Sakyo-ku, Kyoto, 606-8507 Japan
* yasu@bio.keio.ac.jp

Citation: Kamada M, Hase S, Fujii K, Miyake M, Sato K, Kimura K, et al. (2015) Whole-Genome Sequencing and Comparative Genome Analysis of *Bacillus subtilis* Strains Isolated from Non-Salted Fermented Soybean Foods. PLoS ONE 10(10): e0141369. doi:10.1371/journal.pone.0141369

Editor: Dario Cantu, University of California Davis, UNITED STATES

Received: June 15, 2015

Accepted: October 6, 2015

Published: October 27, 2015

Data Availability Statement: All sequence data used in this manuscript are available from DNA Data Bank of Japan (DDBJ) with accession number DRA003017 under DDBJ BioProject PRJDB3484.

Funding: This work was supported by a Grant-in-Aid for KAKENHI (Grant-in-Aid for Scientific Research) on Innovative Areas No. 221S0002 and Scientific Research (A) No. 23241066 from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Bacillus subtilis is the main component in the fermentation of soybeans. To investigate the genetics of the soybean-fermenting *B. subtilis* strains and its relationship with the productivity of extracellular poly-γ-glutamic acid (γPGA), we sequenced the whole genome of eight *B. subtilis* stains isolated from non-salted fermented soybean foods in Southeast Asia. Assembled nucleotide sequences were compared with those of a natto (fermented soybean food) starter strain *B. subtilis* BEST195 and the laboratory standard strain *B. subtilis* 168 that is incapable of γPGA production. Detected variants were investigated in terms of insertion sequences, biotin synthesis, production of subtilisin NAT, and regulatory genes for γPGA synthesis, which were related to fermentation process. Comparing genome sequences, we found that the strains that produce γPGA have a deletion in a protein that constitutes the flagellar basal body, and this deletion was not found in the non-producing strains. We further identified diversity in variants of the *bio* operon, which is responsible for the biotin auxotrophism of the natto starter strains. Phylogenetic analysis using multilocus sequencing typing revealed that the *B. subtilis* strains isolated from the non-salted fermented soybeans were not clustered together, while the natto-fermenting strains were tightly clustered; this analysis also suggested that the strain isolated from "Tua Nao" of Thailand traces a different evolutionary process from other strains.

Introduction

Microbial organisms have a huge influence on the environment and human health. Especially, fermentation by microorganisms plays an important role in food processing, not only in the

Competing Interests: The authors have declared that no competing interests exist.

preservation of foods but also the biological enrichment of food substrates with vitamins, protein, essential amino acids, and essential fatty acids, thus increasing the nutritional value [1]. Moreover, fermentation also enhances the health-promoting effects of soybeans. Fermented soybeans contain significantly more isoflavone genistein than non-fermented soybeans [2], and it has been reported as a chemopreventive agent against cancer [3]. *Bacillus subtilis* is the main component in the alkaline fermentation of soybeans without salt, which is common in East and South-east Asia and in West Africa as a seasoning or side dishes. *B. subtilis* is the best-characterized gram-positive bacterium and often is used as a model organism. About 30 groups from the USA, Japan, Korea, and Europe sequenced and annotated the whole genome of the laboratory standard strain *B. subtilis* subsp. *subtilis* 168 [4]. *B. subtilis* (natto) strain BEST195 whose genome was sequenced by third-generation sequencing technology [5] is a closely related organism of *B. subtilis* 168 and produces "natto", which is a non-salted fermented soybean food that is mainly consumed in Japan.

In the process of natto production, *B. subtilis* (natto) synthesizes some useful products for human health and industry, such as poly-γ-glutamic acid (γPGA), which is the major constituent of viscous material and is a useful polymer for biomedical and industrial applications [6, 7]. Actually, γPGA has been used to purify water in some developing countries [8], and the experimental study using mice reported that γPGA is effective against atopic dermatitis that is a chronic inflammatory skin disease [9]. On the other hand, *B. subtilis* 168 is incapable of the production of γPGA. However, not every *B. subtilis* that ferment soybeans does not produce γPGA, and vice versa; that is, a favorable phenotype for fermentation cannot be predicted only by the production of γPGA [10]. Therefore, unraveling the genetics of *B. subtilis* strains that can be used in the production of non-salted fermented foods is of high interest, and it would be helpful for the efficient production of useful material produced in the fermentation process.

In Southeast Asia, there are some non-salted fermented soybean foods similar to natto, including "Chungkuk Jang" in Korea, "Kinema" in Nepal, "Tua Nao" in northern Thailand, "Pepoke" in Myanmar and "Mac Tua Nao" in northern Laos. These foods are made in the same way as natto, but after fermentation, they are made into a paste. There are many ways to eat, some of fermented soybean pastes are placed in the sun to make them like cracker, and then used as seasoning. To characterize *B. subtilis* strains that are used to produce fermented soybean foods, ninety *B. subtilis* strains have been isolated from fermented soybean foods, including the above foods, and molecular biological investigations have been performed in terms of biotin requirement, productivity of γPGA protease and amylase, phage type and inheritance of insertion sequence (IS) [11]. The IS appeared to be widely distributed among *B. subtilis* strains isolated from non-salted types of fermented soybeans and a relatively small fraction of *B. subtilis* from salted types of fermented soybeans. IS elements are considered to be related to genetic competence of *B. subtilis* and genetic instability of γPGA production in BEST195 [12, 13]. However, no relationship between IS element and γPGA production was apparent in the experimental study [11], and the relevance of IS presence to natto fermentation is still unknown.

To understand the genetics and diversity of the strains producing non-salted fermented soybean foods in more depth, we sequenced the whole genomes of eight strains isolated from non-salted fermented soybean foods in six countries (Korea, Myanmar, Nepal, Thailand, Laos, and Japan) and performed comparative genome analyses with *B. subtilis* subsp. *subtilis* BEST195, which is a starter strain used for natto production and has been sequenced completely in our previous work [5]. In this paper, we describe the assembly and the annotation of the genomes of eight strains and variant analysis focusing on some important features of the soybean-fermenting strains. We also performed phylogenetic analysis using multilocus sequence typing (MLST). Using the assembled genome sequences and identified nucleotide changes from the

BEST195 genome, we found that differences in biotin synthesis and a nucleotide deletion in a flagella motor protein potentially affected the production of γPGA. Additionally, the MLST results indicated that the non-salted soybean-fermenting strains were not classified as a single group.

Materials and Methods

Bacterial strains and genomic DNA extraction

B. subtilis strains used in this study and source materials where they were isolated were listed in Table 1. Strains, Miyagino, Takahashi, and Naruse are three major natto starter strains used in Japan. The BEST195 strain whose genome was used as a reference genome is the Miyagino strain. Strains NARUSE and TAKAHASHI were purchased from Naruse Fermentation Chemical Laboratory K.K. (Tokyo, Japan) and Yuzo Takahashi Laboratory (Yamagata, Japan), respectively. The strain NAFM5 is a derivative of Miyagino whose plasmids were removed [14]. Five strains isolated from the non-salted fermented soybean foods of Asia were used: strain KorC1 was isolated from "Chungkuk Jang" in Korea, strain LaoA1 was isolated from "Mac Tua Nao" in Laos, strain MyaA2 was isolated from "Chine Pepoke" in Myanmar, strain ThaB was isolated from "Tua Nao" in Thailand, and strain NepD5 was isolated from "Kinema" in Nepal. These five *B. subtilis* strains were isolated as described previously [11]. Three of these five strains and a strain (LaoA3) isolated from the same source as LaoA1 were tested in terms of the production ability of γPGA in the previous study [11], and these results were listed in Table 1. Genomic DNA of *B. subtilis* was isolated from Luria Broth culture according to a routine biochemical isolation procedure [15].

We used *B. subtilis* (natto) BEST195 (GenBank accession number AP011541.2) [5] as the reference genome. For comparison analysis and phylogenetic analysis, some relative strains were used, and their genomes were accessed from the following GenBank accession numbers: *B. subitlis* subsp. *subtilis* 168 (NC 000964) [16], *B. subtilis* subsp. *spizizenii* W23 (CP002183) [17], and *B. amyloliquefaciens* LL3 (CP002634) [18].

Sequencing

Whole-genome shotgun sequencing was individually performed on the eight *B. subtilis* strains KorC1, LaoA1, MyaA2, ThaB, NepD5, NAFM5, NARUSE, and TAKAHASHI. Libraries for all strains were prepared using Paired-End Sample Prep Kit and Multiplexing Sample Preparation Oligonucleotide Kit (Illumina Inc., San Diego, CA, USA). DNA was sheared with a Covaris instrument (Covaris Inc., Woburn, MA, USA) to 500 bases, and fragmented DNAs were

Table 1. The detail of fermented soybean foods used as source and phenotypic characters of each strain.

Sample (original strain name in [11])	Fermented soybeans	Appearance	Country	γPGA production in [11]
KorC1 (NFRI8338)	Chungkuk Jang	Raw paste	Korea	untested
LaoA1 (NFRI8302)	Mac Tua Nao	Raw paste	Laos	untested (LaoA3 is No)
MyaA2 (NFRI8316)	Chine Pepoke	Semi-dried block	Myanmar	No
ThaB (NFRI8347)	Tua Nao	Sun-dried chips	Thailand	Yes
NepD5 (NFRI8292)	Kinema	Sun-dried block	Nepal	No
NAFM5	Natto	Raw paste	Japan	untested
NARUSE	Natto	Raw paste	Japan	untested
TAKAHASHI	Natto	Raw paste	Japan	untested

LaoA3 is a strain isolated from the same source as LaoA1.

doi:10.1371/journal.pone.0141369.t001

checked by an Agilent Bioanalyzer DNA 7500 kit. DNA fragments were enriched using a 10-cycle PCR for strains NARUSE and TAKAHASHI, and an 18-cycle PCR for the other strains. The amplified libraries were sequenced on a Genome Analyzer IIx (Illumina Inc., San Diego, CA, USA) instrument, generating 56-bp paired-end reads for strains NepD5, NAFM5, NARUSE, and TAKAHASHI and 58-bp paired-end reads for the other strains. To use only high-quality reads, short reads with a Phred quality below Q30 were filtered out using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Additionally, to remove Illlumina-specific sequencing errors, we applied Trowel [19], which is a *k*-mer spectrum-based error correction method for Illumina reads, with default parameters.

Reference-assisted genome assembly and annotation

To obtain an accurate assembly and utilize the reference genome sequence, we performed the following three steps to assemble genomes. First, short reads were assembled using the *de novo* assembly tool SPAdes version 3.1.0 [20] with *k*-mer = (15, 21, 25, 31, 35, 39, 45). Assembled contigs were then aligned to the reference genome, and consensus sequences were extracted as super-contigs using AMOScmp package [21]. Finally, super-contigs were connected using SSPACE version 3.0 [22], which is a stand-alone scaffolder of pre-assembled contigs using paired-read data. The connected super-contigs were used as draft genomes for each strain. Assembly results were evaluated on the basis of N50, maximum contig length, and the number of contigs was calculated with QUality ASsesment Tool (QUAST) version 2.2 [23]. The gene prediction program Glimmer version 3.02 [24] for the prokaryote genome was applied to each draft genome with trans_table = 11. tRNAs were annotated with tRNAscan-SE version 1.3.1 [25], and rRNA was annotated using RNAmmer version 1.2 [26].

Comparison of assembled genomes with *B. subtilis* BEST195 and 168

To identify orthologous genes to *B. subtilis* BEST195 and 168, we used the reciprocal best hit (RBH) method with BLASTx [27]. In this method, a gene *i* in species *A* is an RBH of gene *j* in species *B* if a query of species *A* with gene *i* yields gene *j* as the top hit with more than 80% identity, and a reciprocal query of species *B* with gene *j* yields gene *i* as the top hit with more than 80% identity. Here, we accounted for the alignment length of each gene and checked whether or not a gene *i*(*j*) is aligned to gene *j*(*i*) with more than 80% of its sequence length. We aligned each assembled draft genome to the reference genome (BEST195) using NUCmer, delta-filter and show-coords, which are modules in MUMmer version 3.23 [28], allowing local alignment. An image of aligned regions was generated using an original script and DNAplotter version 10.2 [29]. Note that from the above image, we can see only the result of local alignment with binary expression whether or not scaffolds in the draft genomes are aligned to the BEST195 genome.

Analysis of insertion sequence (IS)

For each strain, we identified genes corresponding to transposases of the ISs, that is, IS*4Bsu1*, IS*Bma2*, IS*643*, IS*256Bsu1*, IS*Lmo1*, and putative transposase via BLASTx against BEST195 with more than 80% identity. The scaffolds of the draft genomes containing identified IS genes were aligned to the BEST195 genome using MUMmer. To investigate and illustrate positions of a transposase of IS against the BEST195 genome, the scaffolds of the draft genomes were ordered and oriented using ABACAS [30]. After eliminating scaffolds that did not align to the BEST195 genome, the ordered and oriented scaffolds were concatenated into a sequence and aligned with the BEST195 genome using Murasaki [31].

Mapping and variants call

The short reads from each strain were aligned to the BEST195 genome using Burrows-Wheeler alignment (BWA) tool version 0.7.10 [32]. We sorted mapped reads and removed unmapped reads using SAMtools version 0.1.18 [33] and also removed duplicated reads using Picard tools (version 1.119; http://broadinstitute.github.io/picard/). Indel realignment and SNP/INDEL detection were then performed on each strain separately using the Genome Analysis Toolkit (GATK) version 3.2-2 [34] with the parameters -ploidy 1 and -glm BOTH to use a model for identifying SNPs and INDELs at the same time in a non-diploid organism. The detected variants were filtered by VariantFilteration in GATK with the following filter expressions: "DP < 10 || QUAL < 100.0" and "QD < 5.0".

Impacts for each detected variant were predicted at four levels, high, moderate, low, and modifier, using SNP effect predictor (SnpEff) version 3.4 [35]. First, we focused on variants predicted as high-, moderate-, and low-impact effects. High impact indicates that a variant is assumed to have a disruptive impact on the protein, such as frame shift, loss of start codon, and gain of stop codon. Moderate impact indicates a non-disruptive variant that might change the protein effectiveness, such as non-synonymous coding and codon change/insertion/deletion. Low impact indicates that a variant is assumed to be mostly harmless or unlikely to change protein behavior, such as a synonymous variation.

To analyze detected variants statistically, we calculated a genetic variant score for each gene based on the variant data with predicted impact effects. The genetic variant score of a gene *i* for a strain X is defined by

$$\text{gene}_i^{(X)} = \log\left(C_{high}^{(X)} N_{high_i}^{(X)} + C_{mod}^{(X)} N_{mod_i}^{(X)} + C_{low}^{(X)} N_{low_i}^{(X)} + 1\right),$$

where N_{high_i}, N_{mod_i}, and N_{low_i} are the number of variants in gene *i* annotated as high-, moderate- and low-effect impact, respectively, and *C* is a positive constant for each effect defined as follows:

$$C_{low} = 1, C_{mod} = MAX_{low}/2, C_{high} = (C_{mod} + MAX_{mod})/2,$$

where MAX_{low} and MAX_{mod} are the maximum number of low and moderate effects in each strain.

Two statistical analyses using the genetic variant vector, a principal component analysis (PCA), and hierarchical clustering based on the Euclidean distance and furthest neighbor methods were performed with R statistical package. For PCA, we did not normalize the variant matrices to norm 0 and variance 1 before performing PCA because the scores that calculated the above equation are comparable and no further preprocessing was necessary. For a detailed genetic analysis, we listed genes with variants with high and moderate impacts. Gene Ontology (GO) annotations for BEST195 genes were performed using Blast2GO version 3.0.7 [36] with a local BLASTx search against the nr database. The GO terms for listed genes with variants were counted and summarized at level 2 using Bioconductor R packages [37]. The level is defined on the GO graph in goProfiles, which is an R package for the statistical analysis of function profiles. Level 1 corresponds to three categories of GO terms: biological process, molecular function, and cellar component, and level 2 is a group of their directly connected children. For the variant analysis focusing on important genes, genome sequences for each strain were obtained by substituting the BEST195 genome sequence based on the detected variants using a script. Multiple sequence alignments and visualizations of the results were performed with CLC Sequence Viewer 7.5 (CLC Inc., Aarhus, Denmark). Plots of variant positions were generated using Gnuplot (version 4.4; http://gnuplot.sourceforge.net/).

MLST analysis

MLST characterizes bacterial isolates on the basis of sequence polymorphism within internal fragments of seven housekeeping genes. In this study, MLST based on the sequences of internal fragments of the *glpF*, *ilvD*, *pta*, *purH*, *pycA*, *rpoD*, and *tpiA* genes was carried out on strains KorC1, LaoA1, MyaA2, ThaB, NepD5, NAFM5, NARUSE, TAKAHASHI, *B. subtilis* BEST195, 168, W23, and *B. amyloliquefaciens* LL3. The MLST sequences of strains KorC1, LaoA1, MyaA2, ThaB, NepD5, NARUSE, TAKAHASHI, BEST195, and 168 were identified using the *B. subtilis* MLST website (PubMLST; http://pubmlst.org/bsubtilis/) developed by Keith Jolley and sited at the University of Oxford [38], and those of other strains were obtained from this database. The obtained sequences were concatenated into one sequence for each strain using a script. The alignment of these sequences was performed with ClustalW using the Molecular Evolutionary Genetic Analysis (MEGA) 6.0 software [39]. The genetic distance between the sequences was calculated using the Kimura 2-parameter model [40], and the phylogenetic tree was constructed using the neighbor-joining algorithm with MEGA 6.0 software. Branch quality was assessed by the bootstrap test using 1500 replicates. *B. subtilis* W23 and *B. amyloliquefaciens* LL3 strains were used as the outgroups.

De novo assembly with unmapped reads

As a result of mapping, some reads were left unmapped to the reference genome. These reads can be regarded as fragments of an extrachromosomal DNA such as a plasmid or differences from the reference genome. Thus, we extracted unmapped reads of each strain using SAMtools and assembled them with the de novo assembler SPAdes with *k*-mer = (15, 21, 25, 31, 35, 39, 45). First, we filtered scaffolds assembled by unmapped reads using BLASTn against BEST195 genes, and we removed scaffolds that had high similarity. Then, a BLASTn search against the nr database was performed for the remaining scaffolds that only targeted *Bacillus subtilis* group (taxid: 653685) with more than 80% identity. For identification of plasmids, we checked the results of BLAST search in terms of sequence similarity and alignment length.

Data deposition of nucleotide sequence

All whole-genome shotgun sequence reads have been deposited in the Read Archive at DNA Data Bank of Japan (DDBJ) with accession number DRA003017 under DDBJ BioProject PRJDB3484.

Results

Genome sequencing

Next-generation sequencing using the Illumina GAII platform was carried out for the eight *B. subtilis* strains, KorC1, MyaA2, LaoA1, ThaB, NepD5, NAFM5, NARUSE, and TAKAHASHI and an average of 54 million reads were obtained. We used only reads in which any base has more than Q30, which means that a base is inferred using base call accuracies of more than 99.9%, and we filtered out the others from datasets. As a result of filtering, an average of 13.6 million reads (about 25% of the original sequencing output) were left, and the average approximate sequence coverage for the whole genome was 190 and the minimum approximate sequence coverage was 68.06 for the NAFM5 strain. Although filtering with Q30 resulted in a pronounced decrease in the number of reads, we thought that these figures were enough for comparative genome analysis based on previous reports [41, 42]. The details of statistics of sequencing data for each strain are shown in Table A in S1 File.

Moreover, we corrected reads employing the Trowel [19] error-correction algorithm. The total number of reads does not change after error correction by Trowel, but the distribution of read length changes because Trowel trims bases off the read ends with the minimum quality value in a given dataset to achieve high-quality results. The statistics of the corrected reads are also shown in Table A in S1 File.

Reference-assisted assembly and annotation, and comparison with the reference genome

The statistics of the reference-assisted assembly of the eight strains are shown in Table 2. The draft genomes of the eight strains have between 50 and 222 contigs. Based on N50, the maximum contig size, and the number of contigs, strain LaoA1, with the highest sequence coverage, had a good-quality assembly result. In contrast, the assembled genome of strain MyaA2 had a small N50 and maximum contig size in spite of the high sequence coverage. Although strain NAFM5 had the lowest sequence coverage, the assembly result was comparable to that of others. This demonstrates the effectiveness of the reference-assisted assembly because strain NAFM5 is a plasmid-less derivative of the Miyagino (=BEST195) that was used as the reference BEST195 genome. For all eight strains, the total nucleotide length and GC contents of the assembled genomes were close to those of BEST195, which are 4.1Mb and 43.5%, respectively.

Fig 1 shows aligned regions of the BEST195 genome with each draft genome using MUMmer. In our previous study about whole-genome resequencing of BEST195 [5], we discussed gap (incomplete) regions in the first draft genome sequence that previously had been sequenced using very short reads generated from Illumina GAII [43], and we showed that these regions were attributed to GC bias and repetitive sequences. It is known that these regions are difficult to assemble with only short reads. The previous gap regions are indicated as black lines in Fig 1. Many of the unaligned regions (white regions in Fig 1) correspond to the previous gap regions, and regions including transposases of ISs and rRNA genes, which are also difficult to assemble because of the characteristics of repetitive sequences and patterns. When a sequence differs from the reference sequence or *de novo* assembly in the first step does not work well, the regions cannot be covered even though we used the reference-assisted process. Therefore, there is a high possibility that unaligned regions are regions of repetitive sequences that could not be assembled with short reads or deleted regions in the strains.

The numbers of predicted CDS, tRNA, and rRNA genes for the draft genomes are also shown in Table 2. Between 4450 and 5121 protein genes were predicted for the eight strains using Glimmer. One copy of the rRNA cluster (5s, 23s, and 16s rRNA), and 52, 50, and 26 tRNA genes were predicted for three Japanese strains, NAFM5, NARUSE, and TAKAHASHI, respectively. For strains KorC1 and NepD5, 2 copies of the 5s rRNA gene, 1 copy of the 23s and 16s rRNA genes, and 66 and 67 tRNA genes were predicted, respectively. For the other

Table 2. Statics of the reference-assisted assembly and the predicted genes of the eight *B. subtilis* strains.

Statistic	KorC1	LaoA1	MyaA2	ThaB	NepD5	NAFM5	NARUSE	TAKAHASHI
N50	108,025	857,804	69,655	72,647	152,266	114,073	82,070	39,436
total length (bp)	4,077,248	4,202,373	4,036,238	4,111,804	4,035,069	4,007,525	4,079,454	4,045,209
GC content (%)	43.20	43.14	43.28	43.21	43.33	43.42	43.32	43.09
No. contigs	105	50	148	130	83	95	111	222
Predicted CDS	4592	4668	4457	4620	4450	4559	4647	5121
Predicted tRNA	66	72	77	60	67	52	50	26
Predicted rRNA	4	6	6	6	4	3	3	3

doi:10.1371/journal.pone.0141369.t002

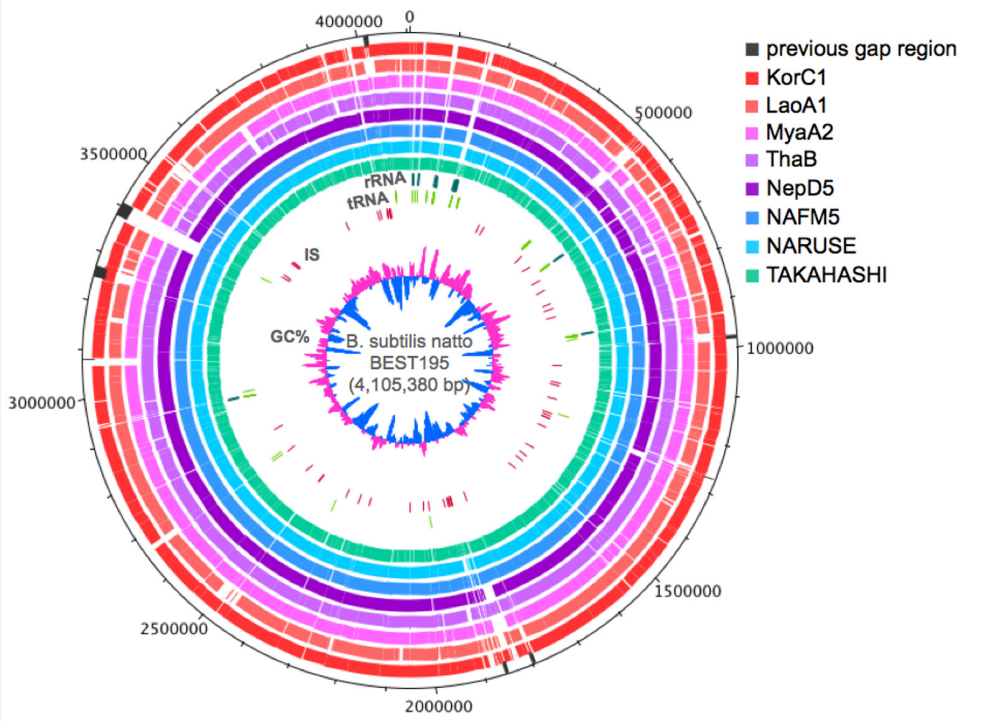

Fig 1. Pairwise sequence alignment between *Bacillus subtilis* BEST195 and the eight *B. subtilis* strains. Each draft genome was aligned to BEST195 using MUMmer. The outside black lines correspond to incomplete regions in the previous BEST195 genome [43], the colored regions in the second to ninth rings from the outside show that scaffolds in the draft genome were aligned to the BEST195 genome. The inner circle at the center displays the G+C content (window size = 10,000bp; step size = 200). This genetic map was generated using DNAPlotter.

doi:10.1371/journal.pone.0141369.g001

strains, LaoA1, MyaA2, and ThaB, 4 copies of the 5s rRNA gene, 1 copy of the 23s and 16s rRNA genes, and 72, 77, and 60 tRNA genes were predicted, respectively. Compared with the numbers of the rRNA cluster and tRNA genes in the BEST195 genome, which are 10 copies (total 30 rRNA genes) and 87, those of numbers in the draft genomes were small. This is attributed to the fact that tRNA and rRNA genes of BEST195 are included in unaligned regions in Fig 1, and these regions could not be assembled from short reads.

Using the RBH method with BLASTx, we investigated orthologous genes to BEST195 and 168. The non-Japanese strains KorC1, LaoA1, MyaA2, ThaB, and NepD5 have orthologs to 83.07, 77.10, 82.22, 82.49, and 87.14% of the BEST195 genes, respectively. The Japanese strains NAFM5, NARUSE, and TAKAHASHI have orthologs to 90.18, 87.55, and 73.64% of the BEST195 genes, respectively. The reason why strain TAKAHSHI had the smallest number of orthologous genes to BEST195 is that strain TAKAHASHI had a lot of small scaffolds in the assembled draft genome, resulting in many predicted genes with a small length. When we considered only the alignment length for the draft genome genes, strain TAKAHASHI had almost the same number of orthologous genes to BEST195 as the other Japanese strains (94.12, 94.14, and 93.94% for NAFM5, NARUSE, and TAKAHASHI strain, respectively). We performed the same procedure against *B. subtilis* 168, and we found that only strain LaoA1 had more orthologous genes to 168 than BEST195, which was 85.76% of the 168 genes. This suggests that strain LaoA1 is related more closely to 168 than BEST195.

Insertion sequence

We investigated the transposases of five types of ISs, that is, IS*4Bsu1*, IS*Bma2*, IS*643*, IS*256Bsu1*, and IS*Lmo1*, and a putative transposase. Performing the BLASTx search with the predicted genes in the draft genomes against BEST195, many genes had similarity with transposases of the ISs. The number of IS and details are shown in Table B in S1 File. As we mentioned in the above section, because the regions including transposase are difficult to be assembled with short reads, we did not consider the alignment length. Therefore, although BLASTx hits with a short alignment length were included, strain ThaB had the highest number of genes similar to transposases of BEST195, and strain NepD5 also had many genes with similarity to the transposases. It was experimentally reported in [11] that the IS frequency was high in strains that were used for fermented soybeans in Thailand and Nepal.

The scaffolds in the draft genomes containing genes similar to transposases were aligned to the gnome sequence of BEST195. Then, scaffolds that could not be aligned to the BEST195 genome could be regarded as nonexistent transposases in BEST195.

Here, we provide one example of these transposases in Fig 2. The coding region in strain NepD5 with length of 1155 bp had similarity to a transposase of IS*256Bsu1* with length of 1369 bp, and it was aligned to a position in BSNT_06845 that encodes a hypothetical protein. Although the function of BSNT_06845 has not been identified yet, *yddC* and *yddD*, which are located upstream and downstream of BSNT_06845, are annotated as transcriptional regulator and mobile element region, respectively. In Fig 2, we can also see another example of the insertion of a transposase in the NepD5 genome. BSNT_06855 is a transposase of IS*4Bsu1*, and it was aligned to a coding region of strain NepD5. This coding region with length 927 bp has similarity to a gene encoding a hypothetical protein of *B. subtilis*. The examples shown in Fig 2 exhibit frequent occurrence of transposition of IS in the soybean-fermenting *B. subtilis* genome, and there is a possibility that the above insertion of transposases is related to a difference of phenotypic trait between strain NepD5 and BEST195.

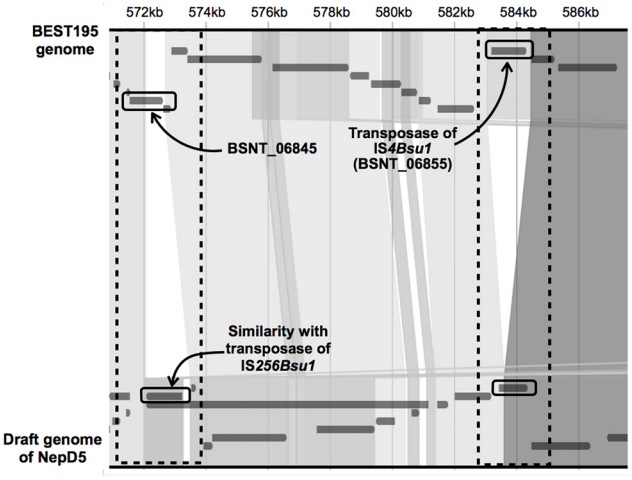

Fig 2. Insertion of transposase into the draft genome of strain NepD5 and the *B. subtilis* BEST195 genome. The NepD5 draft genome is aligned to BEST195 using Murasaki. The similar depth connections of both genomes indicate that they have conserved sequences, and the white regions indicate non-conserved sequences. In the NepD5 draft genome, the sequence similar to the transposase of IS*256Bsu1* is inserted into a gene (BSNT_06845) of BEST195 (left black dashed line), and in the BEST195 genome, the transposase of IS*4Bsu1* is inserted into a coding region of the NepD5 genome (right black dashed line).

doi:10.1371/journal.pone.0141369.g002

Mapping and variant calling

Filtered-corrected short reads sequenced from the eight *B. subtilis* strains were mapped to the BEST195 genome sequence using BWA. An average of 86.2% of reads were mapped to the BEST195 genome with an average of 193.6-fold coverage across the entire genome. Based on the mapping result, SNPs and INDELs were detected for all strains using GATK, and effect impacts of each variant on a genome were estimated by SnpEff. The statistics of the mapping and variant calls for each strain are summarized in Table C in S1 File.

To score and vectorize detected variants (see Material and Methods), we performed PCA and hierarchical clustering analysis, and the results are shown in Fig 3. The principal components obtained in PCA were transformations of variant score vectors by a linear combination that was chosen to maximize the variance of the score vectors of all eight strains. As shown in Fig 3-A, the first principal component (contributing rate: 51%) indicates a feature of the non-Japanese strains, and the second principal component (contributing rate: 19%) can be regarded as a feature to distinguish strain LaoA1 from the other non-Japanese strains. The Japanese strains converged and formed a small cluster. For the first principal component, principal scores are high for genes BSNT_09336, BSNT_09102, and BSNT_09338, which mean that these genes contribute to the first principal component. Although they all are annotated as

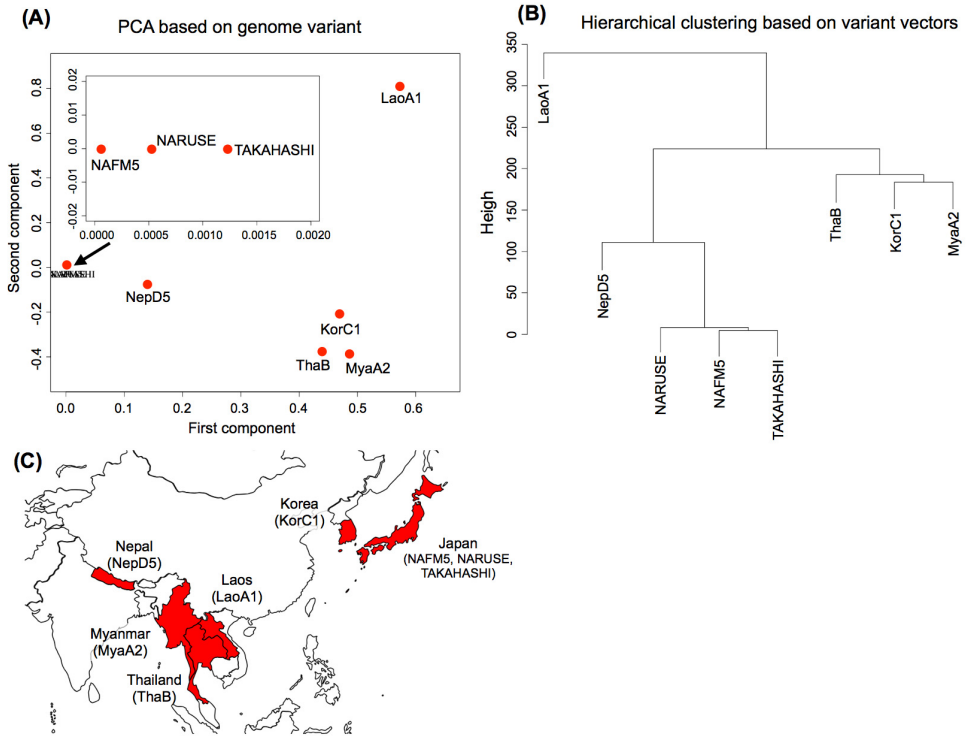

Fig 3. The results of PCA and hierarchical clustering based on variant vectors. (A) Biplot of principal component analysis based on variant vectors. The dots show the eight *B. subtilis* strains, and the upper left image is an enlarged image focused on the three Japanese strains located near (0, 0). The fist principal component features the non-Japanese strains, and the second principal component can be regarded as a feature to distinguish strain LaoA1 and the other non-Japanese strains. (B) Hierarchical clustering of the eight *B. subtilis* strains based on the Euclidean distance between variant scores of each strain using the furthest neighbor method. The different cluster indicates that strains have different variant score patterns. (C) Geographical location of each country.

doi:10.1371/journal.pone.0141369.g003

genes encoding hypothetical proteins and their function is not identified yet, BSNT_09337 and BSNT_09100, which are located near these genes, are annotated with GO terms related to membrane and DNA binding, respectively. Genes BSNT_08913, BSNT_06665, and BSNT_08139 highly contributed to the second principal component. The functions of these genes are also unclear, but BSNT_08139 is annotated with a GO term of motor activity. BSNT_06664, which is located upstream of BSNT_06665, is annotated with a GO term of transferase activity, *comGA* (BSNT_08912) encodes a competence protein GA, and BSNT_08914 is annotated with a GO term related to adenyl nucleotide binding.

We also performed a hierarchical clustering based on variant vectors to investigate differences in gene variants. As the clustering result in Fig 3 shows, the Japanese strains were grouped into a single cluster as in PCA, and strain NepD5 was in another cluster close to the Japanese strains. Strains ThaB, KorC1, and MayA2 fell into a cluster, and strain LaoA1 was in a cluster independent of any other strains. It is obvious from Fig 3-C that genetic differences of the non-Japanese strains are not associated with their geographical locations.

Next, to investigate functional differences across the eight strains, we focused on genes with variants predicted to have high- and moderate-impact effects that have effects on protein functions. A total of 2233, 1416, 1147, and 1214 genes were listed for strains LaoA1, KorC1, MyaA2, and ThaB, respectively, and 365 genes were listed for strain NepD5. On the other hand, for the Japanese strains, 6 and 23 genes were identified as genes with influential variants in strains NARUSE and TAKAHASHI, respectively, and no genes with high and moderate variants were found in strain NAFM5. These listed genes were investigated in terms of GO annotations along with the reference genome. For GO terms associated with "biological process", the number of genes with GO terms of "metabolic process", "cellular process", and "single-organism process" were high for both Japanese and non-Japanese strains. Details of the GO term analysis are shown in Table D in S1 File. Although many analyses can be done with the detected genes with variants, to discuss in depth about the most interesting aspect of *B. subtilis*, in the section below, we investigated the detected variations in terms of biotin synthesis, the subtilisin NAT, the production of γ-PGA, and the motility of *B. subtilis*.

Biotin auxortophy of *B. subtilis* natto. Biotin is synthesized via the *bio* operon (*bioWAFDBI*) in *B. subtilis*. In the natto-fermenting strains, the *bioB* gene is functionally expressed, but the *bioW* gene and *bioF* gene are defective because of a nonsense mutation and a large deletion, respectively [44]. These defects result in biotin auxotrophy of the natto-fermenting strains. According to the experimental studies, the biotin depletion condition leads to an overproduction of L-glutamic acid that is a component of γPGA [45, 46]. Therefore, biotin auxotrophy might be related to transcriptional regulation that is favorable for γPGA synthesis and natto fermentation [10]. The multiple sequence alignment of *bioF* and *bioW* in the genomes of the eight strains, BEST195, and 168 showed that all eight strains had the same deletion in *bioF* as BEST195. However, strains LaoA1 and MyaA2 did not have the nonsense mutation in *bioW* that was also missing in strain 168 (Fig 4). ThaB had the nonsense mutation in *bioW*, but this strain is biotin prototrophic [11]. Thus, it is assumed that biotin auxotrophy of the natto-fermenting strains depends mainly on the deletion in *bioF*. The non-Japanese strains excluding NepD5 have many nucleotide changes in the *bio* operon. Especially, MyaA2 has a variant causing frame shift in *bioD* that encodes a dethiobiotin synthetase. The positions and details of these variants are shown in Table E in S1 File. Additionally, in the ThaB strain, no read was mapped to *bioI* and a portion of BSNT_09438 (Fig 4). *bioI* encodes a biotin biosynthesis cytochrome P450 protein, and BSNT_09438 encodes a hypothetical protein and corresponds to *ytbQ* in strain 168. Pair-wise alignment of BEST195 and the draft genome of strain ThaB using Murasaki also showed the deletion of these regions (Fig 4). Thus, we inferred that biotin

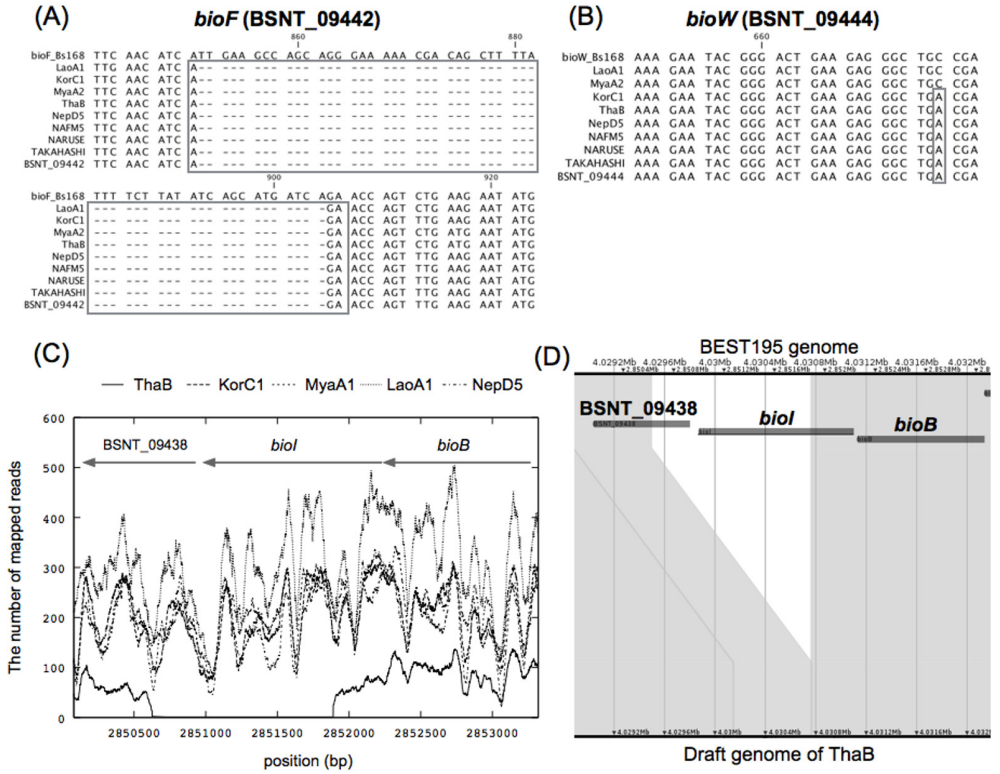

Fig 4. Variant analysis of the *bio* operon. The genome sequences of eight *B. subtilis* strains were obtained via the variant data, and the multiple sequence alignments of them with BEST195 and *B. subtilis* 168 were performed focusing on BSNT_09442 (A) and BSNT_09444 (B). A large nucleotide deletion in *bioF* and a nonsense mutation in *bioW* are found in BEST195. All eight *B. subtilis* strains have the same deletions in *bioF* as BEST195, while the nonsense mutation in *bioW* was not found in strains LaoA1 and MyaA2. (C) The number of mapped reads of the non-Japanese strains against BSNT_09438, *bioI* and *bioB* of BEST195. For only strain ThaB, no read was mapped to the end of BSNT_09438 and most of *bioI*. (D) The pairwise sequence alignment of the draft genome of strain ThaB and BEST195. An unmapped region in (C) was deleted in the genome sequence of strain ThaB.

doi:10.1371/journal.pone.0141369.g004

synthetic organization of strains LaoA1, MyaA2, and ThaB are different from those of other strains, and these changes might result in different biotin metabolism and growth conditions.

Production of subtilisin NAT. Subtilisin NAT (formerly designated as nattokinase) is an extracellular enzyme secreted by *B. subtilis* (natto) [47], and belongs to the alkaline serine protease family. Subtilisin NAT is considered to be the most important enzyme for the characteristic taste and flavor of natto, and a gene encoding subtilisin NAT was determined to be *aprN* [48]. Focusing on *aprN* and neighboring genes, three nucleotide changes were found to be common in *aprN* in strains LaoA1, KorC1, and MyaA2. However, a thymine-to-cytosine nucleotide change in *aprN* that leads to a non-synonymous mutation (asparagine to serine) was found only in strain ThaB. Some nucleotide changes were also found in genes located upstream and downstream of *aprN* in strains KorC1, LaoA1, MyaA2, ThaB, and NepD5. The details of these variations are shown in Figure A and Table F in S1 File. It has been reported that some *B. subtilis* strains isolated from "Tua nao" exhibited higher value of production of subtilisin NAT and PGA than the Japanese commercial strain used in natto production [49]. Therefore, there is a possibility that the nucleotide changes uniquely found in strain ThaB are related to the high productivity of subtilisin NAT, and the experimental verifications of these variants may provide an insight into the difference on the subtilisin NAT production.

Production of γPGA. The production of γPGA has been extensively studied because γPGA plays an important role in industrial production and medical treatment. In strain 168, which is incapable of producing γPGA, a single nucleotide is substituted from cytosine to thymine in the promoter region of *degQ* and a single adenine is inserted into the coding region of *swrAA*. These two nucleotide substitutions are specifically present in the BEST195 genome, and *swrAA* has been reannotated as *yvzD*.

First, we confirmed whether or not the eight strains have the same substitutions as BEST195. No different substitution was found in *degQ*; that is, all eight strains have the same substitution in *degQ*. However, two nucleotides deletion was found in *swrAA* only of LaoA1, while the other strains have the same sequence as BEST195. This two nucleotides deletion also causes the pseudogenization of *swrAA* in common with BEST195 and other strains. The multiple sequence alignment of *swrAA* (*yvzD*) from the eight strains, BEST195, and 168 is shown in Figure B in S1 File.

Cell density-dependent phenotypes of *B. subtilis* are regulated by the ComQXPA quorum-sensing system, which involves the ComP-ComA two-component signal transduction system [50, 51]. It is known that the synthesis of γPGA in *B. subtilis* (natto) is also controlled by this system [52]. *B. subtilis* (natto) uses γPGA as an extracellular nutrient reservoir. A high cell density, which is a sign of overhanging starvation, triggers the synthesis of γPGA [53, 54]. ComQXPA quorum-sensing machinery of *B. subtilis* is known to have divergent structure as for the extracellular signaling peptide ComX and the N-terminal ComX binding domain of the membrane receptor kinase ComP [43, 52].

All strains have the same variations as BEST195 in the DNA region of ComQXPA, and three nucleotide changes that lead to non-synonymous mutations were found in strains KorC1, LaoA1, and MyaA2 at the end of the coding region of *comP* (Figure C and Table G in S1 File). One of these nucleotide changes found in KorC1, LaoA1, and MyaA2 corresponded to the same variant in 168 strain. Of the other two nucleotide changes one was unique for strain KorC1 and one was commonly detected in KorC1 and MyaA2. The N-terminal part of the coding region of *comP* is a transmenbrane sensor domain that binds *comX*, while the end part is a kinase domain that is localized in the cytoplasm [52]. Therefore, these nucleotide changes have a potential impact on the phosphorylation of *comA*. Details of variations and the multiple sequence alignment of sequences corresponding to the nucleotide changes in *comP* are shown in Figure C and Table G in S1 File, respectively. The cell density-dependent phenotype and the motility of bacteria are related to each other. In the following section, we focused on the relationship between the motility of *B. subtilis* and the γPGA production.

Motility of *B. subtilis*. Recent studies reported a relationship between the motility and γPGA production of *B. subtilis* [55, 56]. These studies suggested that flagellar rotation negatively affects γPGA synthesis, and a lack of motility might enhance γPGA synthesis. The bacterial flagellum is a complex molecular machine composed of about 30 different proteins. It is organized into three main parts: basal body (motor), hook, and filament [57]. A diagram of the bacterial flagellum (for *B. subtilis*) with 31 proteins is shown in Fig 5-A. We investigated genes that encode 24 proteins given in Fig 5, excluding proteins that encode genes that are unannotated in BEST195 and 168.

First, we focused on *motA* and *motB*, which encode the MotA and MotB flagellar stator proteins in *B. subtilis*. No nucleotide change was found in *motB* in the eight strains, but two nucleotide changes were found in *motA* in strain LaoA1. These changes cause a codon insertion (glutamate) and a non-synonymous mutation (alanine to threonine); strain 168 has the same sequence as strain LaoA1 (Table H in S1 File). Among the 24 genes given in Fig 5-A, some nucleotide changes were found in 11 genes only in LaoA1 including the above *motA*, and no change was found in 8 genes of the eight strains. Some common changes in more than one

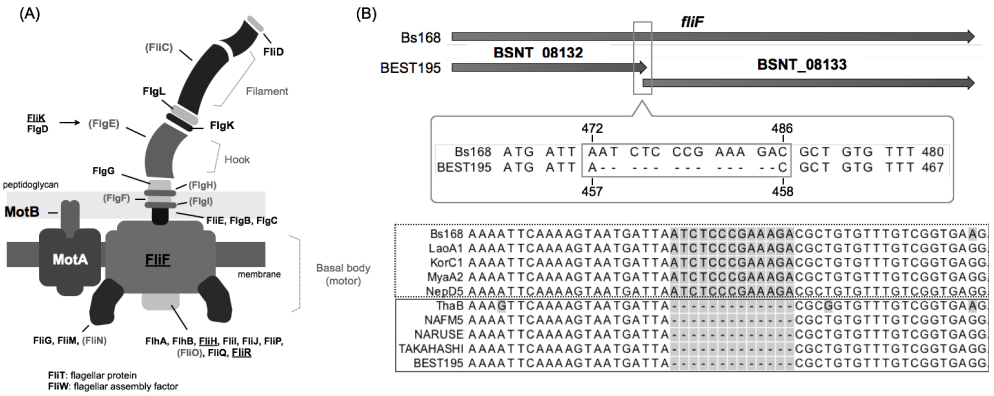

Fig 5. Variant analysis focused on flagellum. (A) Diagram of the bacterial flagellum. Underlined proteins are encoded by a gene that was reannotated in the BEST195 genome. Proteins written in gray and in brackets are not annotated in BEST195 and *B. subtilis* 168; these proteins were removed from the analysis. (B) The nucleotide deletions in *fliF* that encodes FliF and the sequence alignment of *fliF* and corresponding genes BSNT_08132 and BSNT_01833. The same deletion in *fliF* that is in BEST195 was found in strains ThaB, NAFM5, NARUSE and TAKAHASHI, while it was not found in other non-producing γPGA strains.

doi:10.1371/journal.pone.0141369.g005

strain were also found: an adenine-to-thymine nucleotide change in *flgK* encoding FlgK was identified in strains KorC1 and LaoA1, a thymine-to-adenine nucleotide change in *flgC* encoding FlgC was identified in strains LaoA1 and ThaB, and a thymine-to-adenine nucleotide change in *flhA* encoding FlhA was identified in strains LaoA1, KorC1, MyaA1, and ThaB. We also observed a cytosine-to-guanine nucleotide change in *flhA* and a guanine-to-adenine nucleotide change in *flgB* encoding FlgB in strain ThaB. The details of variation in 24 genes are shown in Table H in S1 File.

Focusing on *fliF*, which encodes the flagellar MS-ring protein FliF. A 13-bp fragment encoding 5 amino acids in *fliF* was largely deleted, and *fliF* is separated into two genes, BSNT_08132 and BSNT_08133 in the BEST195 genome. This deletion is estimated to deactivate *fliF* in BEST195. It should be noted that the same deletions were found in strain ThaB and the three Japanese strains, but the other strains had the same sequence as strain 168 (Fig 5-B). According to previous experimental data [11], strains LaoA3, which was isolated from the same source as LaoA1, MyaA2, and NepD5 did not produce γPGA, but strain ThaB did. In other words, a common difference in the non-Japanese strains excluding strain ThaB might have an influence on the production of γPGA. Additionally, in the experiment in [56], γPGA overproduction was observed in mutants defective in flagellar basal body assembly, including a mutation in *fliF*. Although the production of γPGA by KorC1 was not indicated in [11], it is highly likely that the 13-bp deletion of *fliF* causes the strain to be nonmotile and improves γPGA yields. Moreover, this result strongly supports the suggestions from the experimental studies at the nucleotide sequence level. The common deletion in *fliF* among *B. subtilis* strains producing γPGA is a previously unreported variation.

Additionally, as described previously, the production of γPGA in *B. subtilis* isolated from "Tua nao" (the same source foods as ThaB) were higher than that of the Japanese natto-fermenting strains. Thus, it is possible that the nucleotide changes in *flhA* and *flgB* that are uniquely found in strain ThaB trigger the overproduction of γPGA.

MLST analysis

There are many methods to investigate phylogenetic relationship among bacteria strains. Since whole genome of eight strains were sequenced in this study, it is possible to employ

phylogenetic analysis with whole-genome, such as the assembled draft genome sequences and the sequences constructed from detected variants. However, as shown in Figs 1 and 4, incomplete regions were remained on these genomes. Therefore, we employed MLST analysis that uses only seven housekeeping genes selected with consideration for evolutionary rate as phylogenetic analysis in this study.

The housekeeping genes *glpF*, *ilvD*, *pta*, *purH*, *pycA*, *rpoD*, and *tpiA* were chosen for MLST analysis for *B. subtilis* [38]. The internal fragment sequences of these genes were identified using the database and concatenated into a sequence for each strain. Using MEGA 6.0, we analyzed the phylogeny of the eight strains, BEST195, 168, W23, and *B. amyloliquefaciens* LL3. Fig 6 shows an unrooted phylogenetic tree of the eight *B. subtilis* strains of this study, BEST195, *B. subtilis* 168, W23, and *B. amyloliquefaciens* LL3. BEST195, NepD5, and three Japanese strains (NAFM5, NARUSE and TAKAHASHI) formed a tight cluster. These strains are more evolutionary close to each other than the other non-Japanese strains, and these five strains have the same allelic profile for all seven genes and sequence type by MLST analysis. (The allelic profiles for all strains are shown in Table I in S1 File.) The same sequence type was also shown in 12 *B. subtilis* strains registered in PubMLST of *B. subtilis*, and most of the annotated strains were

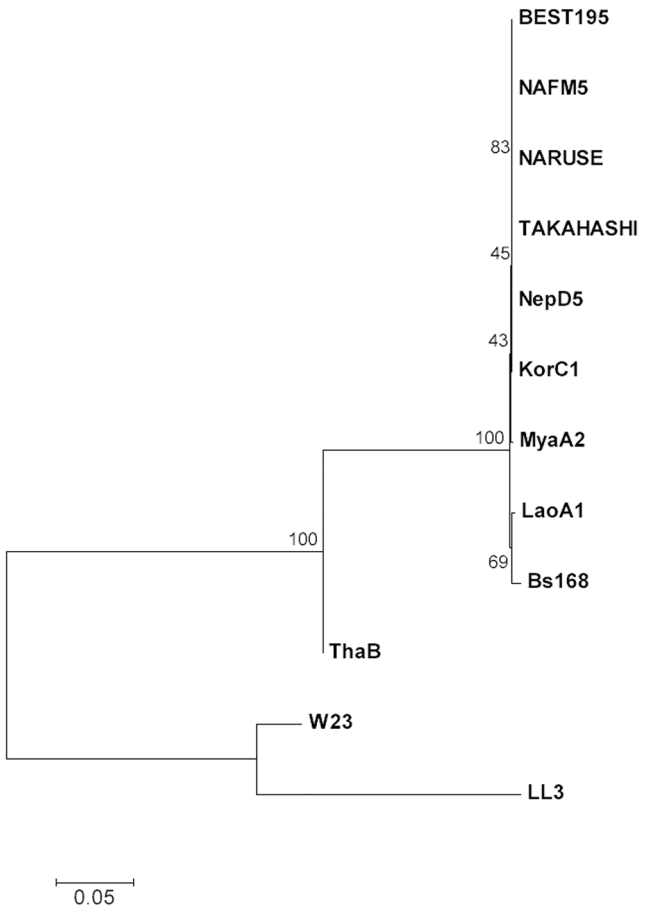

Fig 6. Phylogenetic relationship between the eight *B. subtilis* strains isolated from non-salted soybean foods, *B. subtilis* BEST195, 168, W23, and *B. amyloliquefaciens* LL3. An unrooted phylogenetic tree was generated using the neighbor-joining method based on the seven genes of multilocus sequence typing analysis. The numbers along branches indicate bootstrap percentages.

doi:10.1371/journal.pone.0141369.g006

isolated from soybeans in Japan. The strain LaoA1 formed a cluster with 168. This is a reasonable result because they both are incapable of producing γPGA and have homology to each other as shown in the ortholog analyses. The ThaB strain was thought to be apart from other *B. subtilis* strains earlier than other strains, and it is inferred that a strain with a survival advantage in terms of biotin synthesis and production of γPGA was survived in Thailand. *B. amyloliquefaciens* LL3 was isolated from a fermented food (Korean bibimpa) and synthesizes γPGA [18], but it was on the same branch as W23. From these results, the evolutionary process of the soybean-fermenting strains was thought to be independent from the ability of fermentation and the production of γPGA.

de novo Assembly with unmapped reads

To investigate extrachromosomal DNAs such as a plasmid and the differences from the BEST195 genome, we assembled unmapped reads, which were not used in the variant analysis, into longer scaffolds using SPAdes. A small number of scaffolds with more than 1kb of sequence from the Japanese strains and between 24 and 935 scaffolds for the non-Japanese strains were obtained. Using BLASTn against the BEST195 genes, all scaffolds of strain NAFM5 and most of the scaffolds of the other seven strains were identified as sequences in the BEST195 genome. Between 4 and 15 scaffolds of the seven strains remained without similarity to the BEST195 genes; they were used in a BLASTn search of the nr database that only targeted the *Bacillus subtilis* group. The statistics of *de novo* assembly and BLASTn search are shown in Table 3.

All remaining scaffolds of the NARUSE strain were identified as a plasmid with the same sequence as *B. subtilis* (natto) plasmid pL20 (65,774 bp; GenBank accession number AB615352), and the scaffolds of TAKAHASHI were identified as plasmids with the same sequence as pLS20 and B. subtilis natto plasmid pBEST195S (5838 bp; GenBank accession number AP011542). BEST195 contains two plasmids, pBEST195L and pBEST195S; pBEST195L is similar to pLS20 [58]. Therefore, it is thought that stain TAKAHASHI contains the same plasmids as BEST195 and that strain NARUSE contains a plasmid with the same sequence as pLS20 (pBEST195L).

Most of the remaining scaffolds of the non-Japanese strains were also identified as plasmids. Although the other scaffolds, which were not identified as plasmids, had similarities with some sequences of other *B. subtilis* and *B. amyloliquefaciens* at the nucleotide level, many of them have only partial match with fractions of scaffolds and some of them had similarities with transposases, which have repetitive sequences. Thus, we could not eliminate the possibility of misassembly and did not discuss them into detail. The BLASTn results suggested that stain KorC1 contains a plasmid with the same sequence as pBEST195S and a plasmid similar to *B. subtilis* plasmid p1414 (7949 bp; GenBank accession number AF091592), stain LaoA1 contains a plasmid with the same sequence as p1414 and a plasmid similar to *B. amyloliquefaciens* LL3 plasmid pMC1(6758 bp; GenBank accession number CP002635), and strain MyaA2 contains plasmids

Table 3. The results of *de novo* assembly and BLASTn search against the BEST195 genes.

	KorC1	LaoA1	MyaA2	ThaB	NepD5	NAFM5	NARUSE	TAKAHASHI
No. unmapped reads	1,934,400	5,654,426	1,787,502	3,370,384	742,132	77,136	686,796	469,936
No. SCF (≥ 1kbp)	76	132	67	935	24	5	11	8
No. SCF w/o similarity to BEST195	4	10	15	11	5	0	7	5

SCF: Assembled genomic scaffold

doi:10.1371/journal.pone.0141369.t003

Table 4. The plasmids of BEST195 and the identified plasmids in the eight *B. subtilis* strains.

	BEST195	KorC1	LaoA1	MyaA2	ThaB	NepD5	NAFM5	NARUSE	TAKAHASHI
pLS20 [pBEST105L]	✓							✓	✓
pBEST195S	✓	✓		✓		✓			✓
B. subtilis p1414		✓	✓	✓	✓				
B. amyloliquefaciens pMC1			✓		✓				
B. subtilis pPL1				✓					

doi:10.1371/journal.pone.0141369.t004

similar to pBEST195S, p1414, and *B. subtilis* ATCC 15841 plasmid pPL1 (6704 bp; GenBank accession number DQ140187). Strain ThaB seems to contain plasmids with the same sequences as pMC1 and p1414, and strain NepD5 is thought to contain a plasmid with the same sequence as pBEST195S. The identified plasmids for each strain are summrized in Table 4.

Conclusion

This study performed whole-genome shotgun sequencing of eight *B. subtilis* strains isolated from non-salted fermented soybean foods in Southeast Asia, and it investigated genetic differences among them using comparative genomics approaches. Using comparative variant analysis, we showed the differences in biotin auxotrophism for the strains, and examined potential nucleotide changes that improve the production of subtilisin NAT (nattokinase) and γPGA. Furthermore, our results suggested that the deletion in *fliF* encoding FliF, which constitutes the flagellar basal body, is related to the production of γPGA. We hope that the genomic differences detected in this work promise new insights into phenotypic characteristic of *B. subtilis*.

Although the natto-fermenting strains are classified as *B. subtilis* species in the National Center for Biotechnology Information Taxonomy, there is no sharp taxonomic distinction between the natto-fermenting *B. subtilis* strains and other *B. subtilis* strains. Moreover, it is known that the production of γPGA alone is not a predictor of the ability to ferment natto because many γPGA-positive strains cannot be used for natto production.

Phylogenetic analysis revealed that the natto-fermenting Japanese strains fell into a tight cluster in the phylogenetic tree as previously described [10]. However, when strains that were isolated outside Japan were included in the analysis, the strains with the ability to ferment soybeans did not fall into a single cluster in the phylogenetic tree. Thus, this study showed that *B. subtilis* strains that could be used in fermented soybean production could not be classified into a single taxonomic group based on lineage analysis.

There are some discussions about the origin of non-salted fermented soybean foods [59]. Our results did not show correlations between geographical location and phylogenetic history for *B. subtilis* strains isolated from non-salted fermented soybean foods, and also suggested that strain ThaB, which was isolated from fermented food in Thailand, followed a different diffusion process than other strains. An comparison of genome sequences constructed from short sequence reads revealed that strain LaoA1 was different from the other strains but similar to the standard laboratory strain *B. subtilis* 168. The detailed analysis will be needed to understand this result, but there is a possible involvement of the cultural history of the non-salted fermented soybean foods. The genetic differences revealed in this study provide a clue to understand the origin and routes of non-salted fermented soybean foods in Southeast Asia.

The expected merits of this study are not merely confined to industrial applications. Natto draws attention as a health food and some efforts have been made to develop natto to suit each person's taste through trial and error; for instance, natto with a decreased aroma or the softening of beans for people who experience difficulty in mastication or deglutition. There are

further possibilities for the flexible production of natto using correlation analysis of the genotype and phenotype of fermented soybean foods based on the comparative genome analysis of this study.

Supporting Information

S1 File. Additional information of the results. The details of sequencing output (**Table A**). The results of BLASTn hits with the transposase of the insertion sequence (**Table B**). The details of mapping and variant calls (**Table C**). The details of GO term counts for genes with variations (**Table D**). The details of variations in the *bio* operon (**Table E**). The details of variations in *aprN* and neighboring genes (**Figure A** and **Table F**). The result details for the analysis of the productivity of γPGA (**Figure B**, **Figure C**, and **Table G**). The result details for the analysis of the motility of *B. subtilis* (**Table H**). The details of the MLST analysis (**Table I**).
(PDF)

Acknowledgments

This work was supported by a Grant-in-Aid for KAKENHI (Grant-in-Aid for Scientific Research) on Innovative Areas No. 221S0002 and Scientific Research (A) No. 23241066 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author Contributions

Conceived and designed the experiments: KK YS. Performed the experiments: SH KF MM. Analyzed the data: MK. Wrote the paper: MK KS KK YS.

References

1. Steinkraus KH. Fermentations in World Food Processing. Comprehensive Reviews in Food Science and Food Safety. 2002; 1(1):23–32. Available from: http://dx.doi.org/10.1111/j.1541-4337.2002.tb00004.x.
2. Fukutake M, Takahashi M, Ishida K, Kawamura H, Sugimura T, Wakabayashi K. Quantification of genistein and genistin in soybeans and soybean products. Food and Chemical Toxicology. 1996; 34 (5):457–461. Available from: http://www.sciencedirect.com/science/article/pii/0278691596873558. doi: 10.1016/0278-6915(96)87355-8 PMID: 8655094
3. Messina MJ, Persky V, Setchell KDR, Barnes S. Soy intake and cancer risk: A review of the *in vitro* and *in vivo* data. Nutrition and Cancer. 1994; 21(2):113–131. doi: 10.1080/01635589409514310 PMID: 8058523
4. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, et al. The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature. 1997; 390:249–256. doi: 10.1038/36786 PMID: 9384377
5. Kamada M, Hase S, Sato K, Toyoda A, Fujiyama A, Sakakibara Y. Whole Genome Complete Resequencing of *Bacillus subtilis* Natto by Combining Long Reads with High-Quality Short Reads. PLoS ONE. 2014 10; 9(10):e109999. doi: 10.1371/journal.pone.0109999 PMID: 25329997
6. Meerak J, Yukphan P, Miyashita M, Sato H, Nakagawa Y, Tahara Y. Phylogeny of γ-polyglutamic acid-producing *Bacillus* strains isolated from a fermented locust bean product manufactured in West Africa. The Journal of General and Applied Microbiology. 2008; 54(3):159–166. doi: 10.2323/jgam.54.159 PMID: 18654037
7. Buescher JM, Margaritis A. Microbial Biosynthesis of Polyglutamic Acid Biopolymer and Applications in the Biopharmaceutical, Biomedical and Food Industries. Critical Reviews in Biotechnology. 2007; 27 (1):1–19. doi: 10.1080/07388550601166458 PMID: 17364686
8. Government of Japan, Highlighting JAPAN—"Nippon Poly-Glu";. Accessed 2014 September 24th. http://www.gov-online.go.jp/eng/publicity/book/hlj/html/201402/201402_03_en.html.
9. Lee TY, Kim DJ, Won JN, Lee IH, Sung MH, Poo H. Oral administration of poly-γ-glutamate ameliorates atopic dermatitis in Nc/Nga mice by suppressing Th2-biased immune response and production of IL-

17A. Journal of Investigative Dermatology. 2014; 134(3):704–711. doi: 10.1038/jid.2013.389 PMID: 24025551

10. Kubo Y, Rooney AP, Tsukakoshi Y, Nakagawa R, Hasegawa H, Kimura K. Phylogenetic Analysis of *Bacillus subtilis* Strains Applicable to Natto (Fermented Soybean) Production. Applied and Environmental Microbiology. 2011; 77(18):6463–6469. doi: 10.1128/AEM.00448-11 PMID: 21764950
11. Inatsu Y, Kimura K, Itoh Y. Characterization of *Bacillus subtilis* strains isolated from fermented soybean foods in southeast Asia: comparison with *B. subtilis* (*natto*) starter strains. JARQ Japan agricultural research quarterly. 2002; Q36:169–175. doi: 10.6090/jarq.36.169
12. Takahashi K, Sekine Y, Chibazakura T, Yoshikawa H. Development of an intermolecular transposition assay system in *Bacillus subtilis* 168 using IS*4Bsu1* from *Bacillus subtilis* (*natto*). Microbiology. 2007; 153(8):2553–2559. doi: 10.1099/mic.0.2007/007104-0 PMID: 17660419
13. Nagai T, Tran LS, Inatsu Y, Itoh Y. A New IS4 Family Insertion Sequence, IS*4Bsu1*, Responsible for Genetic Instability of Poly-γ-Glutamic Acid Production in *Bacillus subtilis*. Journal of Bacteriology. 2000; 182(9):2387–2392. doi: 10.1128/JB.182.9.2387-2392.2000 PMID: 10762236
14. Kimura K, Itoh Y. Characterization of Poly-γ-Glutamate Hydrolase Encoded by a Bacteriophage Genome: Possible Role in Phage Infection of *Bacillus Subtilis* Encapsulated with Poly-γ-Glutamate. Applied and Environmental Microbiology. 2003; 69:2491–2497. doi: 10.1128/AEM.69.5.2491-2497.2003 PMID: 12732513
15. Saito H, Miura K. Preparation of transforming deoxyribonucleic acid by phenol treatment. Biochimica et Biophysica Acta (BBA)—Specialized Section on Nucleic Acids and Related Subjects. 1963; 72:619–629. doi: 10.1016/0926-6550(63)90386-4
16. Barbe V, Cruveiller S, Kunst F, Lenoble P, Meurice G, Sekowska A, et al. From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. Microbiology. 2009; 155:1758–1775. doi: 10.1099/mic.0.027839-0 PMID: 19383706
17. Zeigler DR. The genome sequence of *Bacillus subtilis* subsp. *spizizenii* W23: insights into speciation within the *B. subtilis* complex and into the history of *B. subtilis* genetics. Microbiology. 2011; 157 (7):2033–2041. doi: 10.1099/mic.0.048520-0 PMID: 21527469
18. Geng W, Cao M, Song C, Xie H, Liu L, Yang C, et al. Complete Genome Sequence of *Bacillus amyloliquefaciens* LL3, Which Exhibits Glutamic Acid-Independent Production of Poly-γ-Glutamic Acid. Journal of Bacteriology. 2011; 193(13):3393–3394. doi: 10.1128/JB.05058-11 PMID: 21551302
19. Lim EC, Müller J, Hagmann J, Henz SR, Kim ST, Weigel D. Trowel: a fast and accurate error correction module for Illumina sequencing reads. Bioinformatics. 2014; 30(22):3264–3265. doi: 10.1093/bioinformatics/btu513 PMID: 25075116
20. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012; 19(5):455–477. doi: 10.1089/cmb.2012.0021 PMID: 22506599
21. Pop M, Phillippy A, Delcher AL, Salzberg SL. Comparative Genome Assembly. Briefings in Bioinformatics. 2004; 5(3):237–248. doi: 10.1093/bib/5.3.237 PMID: 15383210
22. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011; 27(4):578–579. doi: 10.1093/bioinformatics/btq683 PMID: 21149342
23. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29(8):1072–1075. doi: 10.1093/bioinformatics/btt086 PMID: 23422339
24. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007; 23(6):673–679. doi: 10.1093/bioinformatics/btm009 PMID: 17237039
25. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25(5):955–964. doi: 10.1093/nar/25.5.0955 PMID: 9023104
26. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research. 2007; 35(9):3100–3108. doi: 10.1093/nar/gkm160 PMID: 17452365
27. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. "BLAST+: architecture and applications. BMC Bioinformatics. 2008; 10:421. doi: 10.1186/1471-2105-10-421
28. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biology. 2004; 5:R12. doi: 10.1186/gb-2004-5-2-r12 PMID: 14759262
29. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics. 2009; 25(1):119–20. doi: 10.1093/bioinformatics/btn578 PMID: 18990721

30. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009; 25(15):1968–1969. doi: 10.1093/bioinformatics/btp347 PMID: 19497936

31. Popendorf K, Hachiya T, Osana Y, Sakakibara Y. Murasaki: A Fast, Parallelizable Algorithm to Find Anchors from Multiple Genomes. PLoS ONE. 2010; 5(9):e12651. doi: 10.1371/journal.pone.0012651 PMID: 20885980

32. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009; 25:1754–1760. doi: 10.1093/bioinformatics/btp324 PMID: 19451168

33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297–1303. doi: 10.1101/gr.107524.110 PMID: 20644199

35. Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land SJ, 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. 2012; 6(2):80–92. doi: 10.4161/fly.19695 PMID: 22728672

36. Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21:3674–3676. doi: 10.1093/bioinformatics/bti610 PMID: 16081474

37. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: Open Software Development for Computational Biology and Bioinformatics. Genome Biology. 2004; 5(10):R80. doi: 10.1186/gb-2004-5-10-r80 PMID: 15461798

38. Jolley KA, Maiden M. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010; 11(1):595. doi: 10.1186/1471-2105-11-595 PMID: 21143983

39. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution. 2013; 30:2725–2729. doi: 10.1093/molbev/mst197 PMID: 24132122

40. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 1980; 16(2):111–120. doi: 10.1007/BF01731581 PMID: 7463489

41. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 2014; 15:121–132. doi: 10.1038/nrg3642 PMID: 24434847

42. Kisand V, Lettieri T. Genome sequencing of bacteria: sequencing, *de novo* assembly and rapid analysis using open source tools. BMC Genomics. 2013; 14(1):211. doi: 10.1186/1471-2164-14-211 PMID: 23547799

43. Nishito Y, Osana Y, Hachiya T, Popendorf K, Toyoda A, Fujiyama A, et al. Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data. BMC Genomics. 2010; 11:243. doi: 10.1186/1471-2164-11-243 PMID: 20398357

44. Sasaki M, Kawamura F, Kurusu Y. Genetic Analysis of an Incomplete bio Operon in a Biotin Auxotrophic Strain of *Bacillus subtilis* Natto OK2. Bioscience, Biotechnology, and Biochemistry. 2004; 68 (3):739–742. doi: 10.1271/bbb.68.739 PMID: 15056910

45. Ogunleye A, Bhat A, Irorere VU, Hill D, Williams C, Radecka I. Poly-γ-glutamic acid: production, properties and applications. Microbiology. 2015; 161:1–17. doi: 10.1099/mic.0.081448-0 PMID: 25288645

46. Kimura E. Triggering mechanism of L-glutamate overproduction by DtsR1 in coryneform bacteria. Journal of Bioscience and Bioengineering. 2002; 94:545–551. doi: 10.1016/S1389-1723(02)80193-1 PMID: 16233348

47. Sumi H, Hamada H, Tsushima H, Mihara H, Muraki H. A novel fibrinolytic enzyme (nattokinase) in the vegetable cheese Natto; a typical and popular soybean food in the Japanese diet. Experientia. 1987; 43(10):1110–1111. doi: 10.1007/BF01956052 PMID: 3478223

48. Nakamura T, Yamagata Y, Ichishima E. Nucleotide sequence of the subtilisin NAT gene, *aprN*, of *Bacillus subtilis* (*natto*). Bioscience, Biotechnology, and Biochemistry. 1992; 56(11):1869–1871. doi: 10.1271/bbb.56.1869 PMID: 1369081

49. Inatsu Y, Nakamura N, Yuriko Y, Fushimi T, Watanasiritum L, Kawamoto S. Characterization of *Bacillus subtilis* strains in Thua nao, a traditional fermented soybean food in northern Thailand. Letters in Applied Microbiology. 2006; 43(3):237–242. doi: 10.1111/j.1472-765X.2006.01966.x PMID: 16910925

50. Weinrauch Y, Penchev R, Dubnau E, Smith I, Dubnau D. A *Bacillus subtilis* regulatory gene product for genetic competence and sporulation resembles sensor protein members of the bacterial two-

component signal-transduction systems. Genes & Development. 1990; 4(5):860–872. doi: 10.1101/gad.4.5.860

51. Solomon JM, Magnuson R, Srivastava A, Grossman AD. Convergent sensing pathways mediate response to two extracellular competence factors in *Bacillus subtilis*. Genes & Development. 1995; 9 (5):547–558. doi: 10.1101/gad.9.5.547

52. Tran LS, Nagai T, Itoh Y. Divergent structure of the ComQXPA quorum-sensing components: molecular basis of strain-specific communication mechanism in *Bacillus subtilis*. Molecular Microbiology. 2000; 37(5):1159–1171. doi: 10.1046/j.1365-2958.2000.02069.x PMID: 10972833

53. Kimura K, Tran LS, Itoh Y. Roles and regulation of the glutamate racemase isogenes, *racE* and *yrpC*, in *Bacillus subtilis*. Microbiology. 2004; 150:2911–2920. doi: 10.1099/mic.0.27045-0 PMID: 15347750

54. Kimura K, Tran LS, Uchida I, Itoh Y. Characterization of *Bacillus subtilis* γ-glutamyltransferase and its involvement in the degradation of capsule poly-γ-glutamate. Microbiology. 2004; 150:4115–4123. doi: 10.1099/mic.0.27467-0 PMID: 15583164

55. Cairns LS, Marlow VL, Bissett E, Ostrowski A, Stanley-Wall NR. A mechanical signal transmitted by the flagellum controls signaling in *Bacillus subtilis*. Molecular Microbiology. 2013; 90(1):6–21. doi: 10.1111/mmi.12342 PMID: 23888912

56. Chan JM, Guttenplan SB, Kearns DB. Defects in the Flagellar Motor Increase Synthesis of Poly-γ-Glutamate in *Bacillus subtilis*. Journal of Bacteriology. 2014; 196(4):740–753. doi: 10.1128/JB.01217-13 PMID: 24296669

57. Chevance FF, Hughes KT. Coordinating assembly of a bacterial macromolecular machine. Nature Reviews Microbiology. June 2008; 6:455–465. doi: 10.1038/nrmicro1887 PMID: 18483484

58. Qiu D, Fujita K, Sakuma Y, Tanaka T, Ohashi Y, Ohshima H, et al. Comparative Analysis of Physical Maps of Four Bacillus subtilis (natto) Genomes. Applied and Environmental Microbiology. 2004; 70 (10):6247–6256. doi: 10.1128/AEM.70.10.6247-6256.2004 PMID: 15466572

59. Tamang JP, Kailasapathy K. Fermented Foods and Beverages of the World. Food science and technology. CRC Press; 2010. Available from: https://books.google.co.jp/books?id = MJTLBQAAQBAJ.