Comparative Genomics of Bacillus Subtilis MZK05 and Its Mutant Strain Revealed Genetic Factors Responsible for Enhanced Serine Protease Expression

Md. Mozammel Hoq
Fermentation and Enzyme and Biotechnology laboratory, Department of Microbiology, University of Dhaka, Dhaka-1000

Md. Mahmud Hasan
National Institute of Biotechnology, Ganakbari, Savar, Dhaka-1349

Muhammad Manjurul Karim
Fermentation and Enzyme and Biotechnology laboratory, Department of Microbiology, University of Dhaka, Dhaka-1000

Md. Arafat Al Mamun
Pilot Plant Research Laboratory, Centre for Advanced Research in Sciences, University of Dhaka, Dhaka-1000

Shakila Nargis Khan (shakila@du.ac.bd)
Fermentation and Enzyme and Biotechnology laboratory, Department of Microbiology, University of Dhaka, Dhaka-1000

Research Article

Keywords: Complete genome sequence, Bacillus subtilis MZK05, Mutant M9, Serine protease, Signal peptidase

Posted Date: January 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1242999/v1

Comparative genomics of *Bacillus subtilis* MZK05 and its mutant strain revealed genetic factors responsible for enhanced serine protease expression

Md. Mozammel Hoq[1], Md. Mahmud Hasan[1,2], Muhammad Manjurul Karim[1], Md. Arafat Al Mamun[1,3] and Shakila Nargis Khan[1*]

[1]Fermentation and Enzyme and Biotechnology laboratory, Department of Microbiology, University of Dhaka, Dhaka-1000, Bangladesh

[2]National Institute of Biotechnology, Ganakbari, Savar, Dhaka-1349, Bangladesh

[3]Pilot Plant Research Laboratory, Centre for Advanced Research in Sciences, University of Dhaka, Dhaka-1000, Bangladesh

Address correspondence: Shakila Nargis Khan, shakila@du.ac.bd.

Abstract

Genome sequence study of an industrially-important strain, *Bacillus subtilis* M9, a mutant version of wild strain *Bacillus subtilis* MZK05 was conducted to uncover genetic factors responsible for enhanced serine proteases expression in addition to its other industrial enzymes, metabolites and bacteriocins producing efficacy. The wild type and the mutant genome contained a size of 4,145,727 and 4,045,950 bp, with 4,352 and 4,383 genes; and 477 and 478 subsystems respectively. Genomic comparison with 31 *B. subtilis* sourced from different countries showed both wild and mutant shared same type of genome structure with 20 others. Moreover, 6,000 kb pangenome showed that they share 3082, 1449, and 25757 core, unique and accessory genes respectively. A sum of 32,559 mutations were found with three major genomic structural changes in the upstream and downstream of an extracellular alkaline serine protease, AprX and a periplasmic serine protease, HtrC in M9 genome when compared to the wild type. Furthermore, 11 different serine protease genes and 4 different signal peptidases were found with several mutational changes in M9. In addition, mutations found in core genome of BsM9 in phage-like element, major capsid protein, and phage portal protein are the likely reasons of high-level serine protease activity.

Keywords: Complete genome sequence, *Bacillus subtilis* MZK05, Mutant M9, Serine protease, Signal peptidase

Introduction

Bacillus subtilis is a rod-shaped, Gram-positive, spore forming, nonpathogenic soil bacterium that secretes numerous enzymes to degrade a variety of substrates, enabling the bacterium to survive in a continuously changing environment. Proteases from *Bacillus* spp are one of the most commercially valuable enzymes[1] having enormous applications in food, feed, leather processing, detergent, digestive aid and therapeutic purposes[2]. The genes coding for eight extracellular serine proteases were identified in *B. subtilis*: *aprE*[3,4], *bpr* [5,6], *epr*[7,8], *mpr*[9,10], *nprB*[11], *nprE*[12], *vpr*[13], and *wprA*[14]. While two of them were considered as major protease-coding genes, viz. *aprE* (alkaline serine protease subtilisin) and *nprE* (neutral metalloprotease) accounting for 95% of the total extracellular protease[15], the rest were minor extracellular proteases. These enzymes help cells supplying amino acids for growth by degradation of extracellular proteins[16].

Earlier, we isolated eight *Bacillus* species from the effluents of tannery and poultry farms of Bangladesh[17], of which *B. subtilis* MZK05 (Bs MZK05) demonstrated potential extracellular enzymes, identified as serine protease[17]. With a view to increasing the performance, the bacterium was subjected to random mutagenesis using both chemical (ethyl methane sulfonate) and UV radiation[18] that eventually generated a mutant strain, *B. subtilis* M9 from a selection of potential strains, able to produce more efficient extracellular protease than that of the wild type. This was demonstrated in unhairing and bating of skin and hides, a performance comparable to that of the commercial enzymes, thereafter was applied successfully in leather industries[19],[20],[21]. Therefore, we attempted here to compare the genome sequencing of the two strains to pinpoint the probable cause of high-level expression of serine proteases in *B. subtilis* M9 mutant, the number of proteases, and other industrially useful enzymes and metabolites, genes responsible for extracellular protease secretion, genome variation in between the strains studied here with that of strains from different countries of origin, and evolutionary relationships present with other bacterial species of *Bacillus subtilis* group.

Results and Discussion

Genome sequence information

Classification, general features, and genome sequencing information of *B. subtilis* MZK05 are tabulated (Table 1) according to the MIxS recommendations. A total of 297,674 read pairs (average read length, 35-301) and 372,189 read pairs (average read length, 35-301) were found with FastQC from *B. subtilis* MZK05 and *B. subtilis* M9 respectively, and 37 contigs were obtained from each strain by SPAdes v. 3.11.1 *de novo* genome assembly. N50 genome sequence of *B. subtilis* MZK05 comprises a circular chromosome with a total size of 4,145,727 bp, and draft whole genome of *B. subtilis* M9 consists of 4,045,950 bp, and both the strains comprised of 43.5% of G+C contents.

Item	Description
Organism	Cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Bacillaceae; *Bacillus*; *Bacillus subtilis* MZK05
Isolate type	Monoisolate
Sample type	Chromosomal DNA
Investigation type	Bacterial genome sequencing and annotation
Project name	Complete genome sequencing of *Bacillus subtilis* MZK05 and *Bacillus subtilis* M9
Geographic location of the sampling site	Hazaribagh, Dhaka City Latitude 23.777176, Longitude 90.39942
Geographic location (country and/or sea, region)	Dhaka, Bangladesh
Collection date	February, 2000
Environment (biome)	Effluents discharge area of leather manufacturing industries
Environment (feature)	Effluent's liquids, mud and soil
Environment (material)	Hides & skins processing waste liquid, muds, and soils
Environmental package	Liquids, muds and soils
Sequencing method	Illumina Miseq
Assembly method	SPAdes 3.11.1
Finishing quality	Finished (complete)

Table 1. Classification, general features, and genome sequencing information of *B. subtilis* MZK05 and *B. subtilis* M9 strains according to the MIxS

Gene function annotation, genomic features, variants and genome overview

The genetic features of *B. subtilis* MZK05 and *B. subtilis* M9 genome sequences, assembly, and annotations are summarized in Table 2, and major subsystem features predicted from RAST are shown in supplementary information (Figs. S1-S2, and Table S1). We found that both *B. subtilis* MZK05 and *B. subtilis* M9 strains contain genes coding for industrially useful enzymes and metabolites, such as major and minor serine proteases, signal peptidases, metalloproteases, alpha-amylase, phytase, xylanases, pectate lyase, insulinase, bacteriocin-encoding genes (nisin, sublancin, lanthipeptide and subtilin) and serine protease transcriptional regulators. Summary of the related results are shown in supplementary information (Table S2). Circular graphical map of the genomes with COGs annotations are shown in Figs. 1a-b. Variant calling is an essential genome comparison method that provides insights into organismal differences at nucleotide level. It comprises structural variants, single and multiple nucleotide polymorphisms (SNPs, MNPs), insertions and deletions (indels). It coordinates genomes with polymorphisms relative to a reference[22] and estimates the impact of variants that occur in genes. Using snippy 3.2-dev variant calling[23], we have found 32,559 total variants, of them 3,814 Variant-COMPLEX, 323 Variant-DEL, 304 Variant-INS, 5 Variant-MNP, and 28,113 Variant-SNP. Major genomic structural differences between the wild and mutant Bs M9 genomes and variants are shown (Fig. 1).

Features	*Bacillus subtilis* MZK05	*Bacillus subtilis* M9
Base pairs (bp)	4,145,727	4,045,950
GC content	43.5%	43.5%
No of contigs	37	37
N50 value of contigs	1,045,489	325,303
Longest contig length (bp)	1,057,670	612,811
Genes	4,352	4,383
CDS (coding)	4,140	4,130
Subsystems	477	478
rRNAs	12	9
tRNAs	85	83
ncRNAs	5	5
Pseudogenes (total)	110	134
GenBank accession	CP032315	RFFO00000000
SRA accession	SRR8199929	SRR8200505

Table 2. Summarized features of genome sequence, assembly and annotation of *B. subtilis* MZK05 and *B. subtilis* M9

Mutational changes in genome sequence

The analysis of next generation genome sequence data yields major insights into the genetic features of the organism[24],[25]. Here, we found that the mutations in the *B. subtilis* MZK05 strain produced nearly about 0.1 Mb reduced genome sequence than that of the mutant strain, *B. subtilis* M9. The related mutational variants are shown in Fig. 1c and the major genetic changes are described in pangenome section. From mauve genome alignment, it was clearly revealed that there were three major changes occurred in very close region of two serine protease genes in *B. subtilis*M9. A frameshifted region was identified in the downstream of a periplasmic serine protease *htrC* (Fig. 2a). Besides, deletions of about 75 kb and 125 kb nucleotides were detected from upstream and downstream of an extracellular alkaline serine protease *aprX* (Fig. 2b). These deletions occurred just before 17 genes and after 2 genes of *aprX*, changes of which are thought to be linked with the increased production of serine protease in *B. subtilis* M9. Similar to this finding, about 874 kb (20.7%) of the genomic sequence was deleted from *B. subtilis* 168, and the resulting mutant strain, *B. subtilis* strain MBG874 exhibited high level productivity of extracellular cellulase and alkaline protease[26]. In another study, a deletion of 15% of the genome sequence (0.71 Mb) truncating all mobile DNA and cryptic virulence genes from an *Escherichia coli* strain yielded some unexpected beneficial properties, yet ensuring normal cellular growth and protein expression[27].

Mutational changes in serine proteases

There are 11 different serine protease genes present in both *B. subtilis* MZK05 and *B. subtilis* M9 (Figure S3). Amongst them, there are 2 major alkaline serine protease (*aprX*) genes with peptidase S8 family domain, 1 minor serine protease (fibrinolytic protease Vpr) with inhibitor I9 and peptidase S8 subtilisin Vpr-like family domain, and 1 trypsin-like extracellular serine metalloprotease (Mpr) (V8-like Glu-specific endopeptidase). In addition, both wild and mutant strains had 1 intracellular serine protease (Isp), 2 rhomboid family intramembrane serine protease (one of them is YqgP), 1 Bacillopeptidase F (Bpr) with Peptidase S8 family domain with multiple catalytic active sites, and 3 membrane-bound and periplasmic serine protease do-like genes *HtrA, HtrB,* and *HtrC*. Using MEGA-X protein alignment, we identified

several mutations in different types of serine protease enzymes found in *B. subtilis* MZK05 and *B. subtilis* M9 (Suppl. Mat. Table S3). Major structural changes in rhomboid family intramembrane serine protease owing to mutation were detected in the serine protease of *B. subtilis* M9, particularly 3 changes in active sites - S398N, E436K and N463Y; and 2 mutations at G359R and V376A that produced major and noticeable 3D structural changes (Fig. 3a). There were four changes observed in extracellular serine metalloprotease, one of which was Q88L that caused 3D structural change owing to loss of an alpha helix (Fig. 3b). Two different sized AprX proteins were evident in both *B. subtilis* MZK05 and *B. subtilis* M9. The larger AprX harbors 52 extra amino acids upstream of the coding sequence, the rest of the amino acid sequences were similar to both AprX. Two mutations were found in the small AprX in two different positions: N230S and T378A (Fig. 3c). The large AprX, on the other hand, had two substitutions: N282S and T430A (Fig. 3d). Remarkably, both of these large and small AprX had same type of two non-synonymous changes: asparagine to serine at positions 282 and 230 (Fig. 3e); and threonine to alanine at positions 430 and 378 in large and small AprX respectively (Fig. 3f). Interestingly, the positions of the changes between the two varied sizes of AprX become alike if the stretch of upstream 52-amino-acid residues present in the large AprX is deducted. Further, 6 mutational changes were evident in Vpr, one of these were T719S that brought a 3D structural change in the enzyme (Fig. 3g). Besides, 1 mutational change was observed in Isp (Fig. 3h), and each in HtrA, HtrB, and HtrC, however, no significant structural change was resulted due to these mutations. In addition, 7 mutational changes were found in immune inhibitor A peptidase M6 region of bacillopeptidase F (Suppl. Mat. Table S3). Overall, the deletion of upstream and downstream of an *aprX* gene (Fig. 2b), frameshift mutation in the downstream of a periplasmic serine protease *htrC* (Fig. 2a), and multiple mutations in different types of serine proteases genes, specially *aprX, vpr, mpr* and *yqgP* in the *B. subtilis* M9 genome could be the likely reasons for high-level enzymatic activity of alkaline serine protease in *B. subtilis* M9. On a relevant note, as mutation in YqgP (GluP) region of rhomboid proteases altered glucose uptake and cell division in *B. subtilis*[28,29], the presence of 15 mutations altering major structural changes in rhomboid protease YqgP of

B. subtilis M9 could also be an important factor for high-level cell division and consequent serine protease expression in *B. subtilis* M9.

Observation in transcriptional regulators

There were 6 different types of serine protease regulatory genes in both *B. subtilis* MZK05 and *B. subtilis* M9 identified as *degQ, codY, lexA, abrB, ylbF*, and *marR* which are responsible for the expression of *aprX, vpr, isp, mpr, yqgP, bpr, htrA, htrB*, and *htrC* genes. No changes at the nucleotide level were observed in these regulatory genes.

Mutational changes in signal peptidases

We found that both *B. subtilis* MZK05 and *B. subtilis* M9 had 6 different serine types of signal peptidases (SPases) protein coding genes, and among them *sipI, sipT, sipV* and 2 different *sipS* were from type I signal peptidase (SPase I) and a lipoprotein signal peptidase (*lspA*) from type II signal peptidase (SPase II). It is ascertained that LspA, and SipT didn't have any mutational change in their amino acid sequences however, SipI, SipV and each type of SipS enzyme had mutational and structural changes that distinguished *B. subtilis* M9 from the wild type. These are V33I and T128I, the later change affected three active catalytic sites: 100 (alanine), 101 (glutamine), and 102 (proline) resulting in alteration of the local (amino acids 95 to 127) structure of SipV protein (Figs. 4a-b). The SipS, a signal peptidase of the S26 family, encountered 2 changes: L15T and S141T. While the later didn't have any impact, the former affected the upstream protein structure until amino acid position 35 (Figs. 4c-d). SiPI produced 2 mutations: both of them influenced alteration of protein's 3D structure. These are T11S affecting amino acid position 32 to 11 (Figs. 4e-f) and I179M affecting downstream terminal structure of the enzyme (Fig. 4e). Further, mutation at T64A of a second S26 family signal peptidase (Fig. 4g) altered most of the 3D structure of the enzyme. This change affected catalytic sites: S47 and H87 (Fig. 4h). Overall, multiple mutations in different serine type signal peptidases are thought to produce a high-level enzyme activity of serine proteases in *B. subtilis* M9.

Identification of bacteriocin gene clusters

From AntiSMASH 3.0[30] and BAGEL4[31], four different classes of bacteriocin (nisin, sublancin, lanthipeptide and subtilosin) encoding genes were found both in *B. subtilis* MZK05 and Bs M9. The structure of bacteriocin genes with operon in MZK05 genome is illustrated in supplementary information, Figure S4.

Comparative genomics and phylogeny

Comparative genomics of 31 complete genome sequences of *B. subtilis* strains from different countries, deposited in NCBI GenBank was addressed using Progressive Mauve genome alignment. Seven different types of structural and architectural patterns (Suppl. Mat. Fig. S5) were revealed; of them, 22 bacterial strains hailed from South Korea, Denmark, Germany, Greece, Belgium, China, India and USA including the present *B. subtilis* MZK05 and *B. subtilis* M9 strains grouped together to share common type of genome structure. Further analyses of these 31 *B. subtilis* strains by the BRIG-BLAST method presented a snapshot of the genetic diversity and close relatedness. The different color gradients of rings 4-34 (Suppl. Mat. Fig. S6), indicated a BLAST match of a percent identity in the key, located in the right side of the circular comparative genome map. From BRIG image, it was clearly evidenced that *B. subtilis* MZK05 had three major specific regions: 16.5 Mbp to 17.5 Mbp, 17.8 Mbp to 19.2 Mbp, and 24.2 to 25.5; and twenty-five small, distinctive segments scattered throughout *B. subtilis* MZK05 genome as compared to other strains. The symmetric identity of *B. subtilis* MZK05 with 30 other *B. subtilis* strains ranged from 70% to 90%, and the greatest similarities were observed with strains reported from Germany (ring 4, strain 168, CP019662), Belgium (ring 6, strain 2014-3557, CP045672), China (ring 8, strain 7PJ-16, CP023409), India (ring 10, strain FB6-3, CP032089), South Korea (ring 13, strain HJ0-6, CP016894), Japan (ring 16, strain NBRC 13719, AP019714), USA (ring 18, str. SMY, CP050532), South Korea (ring 23, strain SRCM102756, CP028218), China (ring 24, strain H1, CP026662), and South Korea (ring 28, strain SSJ-1, CP032860). Phylogenetic tree analysis revealed that bacterial strain *B. subtilis* MZK05 was grouped with *B. subtilis* strains (Suppl. Mat. Fig. S7) hence, the strain was identified as *B. subtilis*. The

16S rRNA gene sequence of strain *B. subtilis* MZK05 is closely related to strains reported from South Korea, Denmark, Germany, Greece, Belgium, Egypt, Poland, China, Japan, South Africa, India and USA, generating 100% sequence identity. Besides, this strain is also very close to *B. velezensis* and *B. amyloliquefaciens* than *B. licheniformis* and *B. paralicheniformis* of *Bacillus subtilis* group bacteria.

Pan-genome and COGs protein function analysis

The NCBI-retrieved 30 *B. subtilis* strains and the strains of interest of this study produced a 6,000 kb long pangenome (Fig. 6a), where each of the strains shared 3,082 core genes. Besides, 1,449 unique genes and 25,757 accessory genes were distributed in scattered manner among the strains. In pangenome, the position of the new genes in the different genomes are shown between 4,000 kb to 6,000 kb (just before the start position 0). The pan-versus-core gene plot clearly indicates that the pangenome window is still open but may be closed soon in the case of *B. subtilis* (Fig. 5a) and with the addition of each new genome to the *B. subtilis* multipart, a change in the pan-genome shape can result. Distribution of new genes within the pangenome of the *B. subtilis* genome are shown (Fig. 5b). Representative protein sequences of each strain were identified and their COGs and KEGG identities were determined. The details of core, accessory and unique genes present in each of the 30 genomes are shown (Suppl. Mat. Table S4). From the resulted table, it was found that 45 accessory genes were deleted or broken in *B. subtilis* M9 from *B. subtilis* MZK05. While the number of exclusively absent gene was 3 for *B. subtilis* MZK 05, the mutant lost additional 15 more genes making the figure 18 for *B. subtilis* M9. Besides, *B. subtilis* M9 had 43 extra unique genes that *B. subtilis* MZK05 and these genes were originated due to mutational breakdown of accessory genes. From *B. subtilis* MZK05 and *B. subtilis* M9 core genome blast result, it was found that major mutational changes were occurred in phage-like element PBSX protein XkdG, XkdM, XkdQ, phage major capsid protein, phage portal protein, lytic transglycosylase, sulfate adenylyltransferase, baseplate J/gp47 family protein, LysM peptidoglycan-binding domain-containing protein, ABC transporter ATP-binding protein and numerous hypothetical proteins (Suppl. Mat. Table S5). Previous studies[26, 27] found that reduction of mobile element and genomic DNA increased high level productivity of

extracellular cellulase, alkaline protease and beneficial properties. Likewise, reduction of some accessory genes and mutation in core genes may increase the production and activity of alkaline serine proteases in this study. The KEGG and COGs distribution of the representative proteins present in the core, accessory, and unique genome are shown in Fig. 5c and Fig. 5d respectively, and the COGs frequency heatmap of the representative proteins present in all of 30 genomes are shown (Fig. 5e). Relative evolutionary divergence of Neighbor Joining (NJ) core and pan phylogenomic tree of 30 strains, constructed on the basis of core genome and pan gene alignments are shown in (Figs. 6b-c). It was clearly revealed that the position of *B. subtilis* MZK05 and *B. subtilis* M9 in core phylogenomic tree and pan phylogenomic tree were different. In core phylogenomic tree *B. subtilis* MZK05 and *B. subtilis* M9 belonged to same node with strains of China and Poland, whereas in pan phylogenomic tree mutant M9 strain moved to a new position of another node with two South Korean strains. From core and pan phylogenomic tree analysis of wild and mutant genomes, it was clearly revealed that mutational changes affected the accessory genes of *B. subtilis* M9.

Conclusions

The comparative genomics of wild type *B. subtilis* MZK05 and its mutant strain M9 revealed underlying genetic factors that prompted enhanced activity of serine protease in the mutant. Further, the analyses provided insight for some very useful genes, coding for enzymes and metabolites, such as, major and minor serine proteases, signal peptidases, metalloproteases, alpha-amylase, phytase, xylanases, pectate lyase, insulinase, bacteriocin-encoding genes (nisin, sublancin, lanthipeptide and subtilin) and serine protease transcriptional regulators, thereby making the strains industrially important for potential bioeconomy.

Data availability

The complete genome sequence of the *B. subtilis* MZK05 and the draft whole genome sequence of *B. subtilis* M9 are available at GenBank under the accession CP032315 and RFFO00000000 respectively.

The raw data (Illumina) are available in the SRA under the accession number SRR8199929 and SRR8200505, respectively.

Methods

Bacterial strains and genomic DNA preparation

The bacterial cultures of *B. subtilis* MZK05, previously isolated and identified (Hoq *et al*., 2005) and its mutant *B. subtilis* M9 developed[18] were prepared for sequencing by growing the strains separately overnight at 37°C in 100 ml LB broth. Genomic DNA from *B. subtilis* MZK05 and *B. subtilis* M9 strains were extracted with Thermo Scientific GeneJET Genomic DNA Purification Kit according to manufacturer's instructions. Genomic DNA was quantified using NanoDrop One spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and the integrity of the genomic DNA was analyzed by 1% agarose gel electrophoresis. Then, a 300-bp paired-end library was constructed using the Nextera XT DNA library preparation kit (Illumina) and bead-based normalization following the standard manufacturer's protocol.

Genome sequencing, and *De novo* assembly

DNA of the strains (*B. subtilis* MZK05 and mutant *B. subtilis* M9) were sequenced with the Illumina MiSeq technology (Illumina, San Diego, CA, USA) by MiSeq Reagent Kit v3 (600-cycle) at the Genome Research Institute of North South University (NSU), Bangladesh. The quality of the raw reads was checked with FastQC v. 0.11.7[32]. Illumina adapter removal, read trimming and filtering were performed using BaseSpace v3. Reads of both genomes were then assembled with SPAdes v. 3.11.1[33] in 'careful' mode using the default k values of 21, 33, 55, 77, 99 and 127. Resulted contigs of both strains were scaffolded using the move contigs command with default parameters of Mauve v. 2.4.0[34],[35], where the genome of *B. subtilis* (CP021507) was used as reference.

Genome annotation, gene prediction and variant calling

Automated functional annotation of scaffolded contigs of *B. subtilis* MZK05 and *B. subtilis* M9 was performed by the RAST server[36]. Classic RAST v. 2.0 along with FIGfam (release 70) was used to annotate both strains. Errors were fixed automatically. Gaps were backfilled and metabolic models were created. Debug was not turned on and replication was disabled. Besides, final annotations were performed using NCBI Prokaryotic Genome Annotation Pipeline (PGAP)[37]. Cluster of orthologous groups (COGs) protein function analysis and KEGG pathway were predicted by eggNOG-mapper v2[38] and WebMGA server[39]. Circular graphical maps with COGs annotations were created by GView[40]. Wild and mutant bacterial genomic variants were detected by snippy 3.2-dev[23] variant calling from NGS reads and overall genomic differences were observed by Progressive Mauve genome alignment[34].

Identification of differences between serine protease genes in the genomes of the two strains.

Identification of genetic, structural and positional differences for high-level expression of serine protease genes a genome-wide alignment with the Mauve alignment tool[34] was performed where GenBank (.gb) file of the genome and mauve search tool were used to find out the position of serine protease genes in the aligned genomes. Structural changes of both nucleotide and amino acid sequences of serine proteases were also studied. Genetic level differences were studied by NCBI blastp and MEGA-X. 3D Pbd file of protein sequences were predicted by phyre2[41] webserver and structural and mutational changes were studied by Pymol[42].

Identification of bacteriocin gene cluster

The bacteriocin synthesizing gene clusters were revealed by AntiSMASH 3.0[30] and BAGEL4[31].

Comparative genome and phylogeny analysis

Comparative genome and phylogenic analysis of closely related species of *B. subtilis* strain MZK05 and M9 was performed by NCBI microbial genome BLAST. Best matched genome sequence of *B. subtilis*

306 strains was analyzed and studied by two techniques, Mauve and Blast Ring Image Generator (BRIG).
307 Progressive Mauve command was used with default parameters to perform whole genome alignment with
308 29 best BLAST matched complete genome sequence of *B. subtilis* strains to reveal the genomic structural
309 and architectural variations. Besides, best matched genome sequence of *B. subtilis* strains was also
310 analyzed and studied by BRIG version 0.95 to determine overall genome comparison[43]. The circular
311 comparative genomic map was constructed by BRIG with NCBI local blast-2.8.1+, with standard default
312 parameters. To determine the phylogenetic relationship, complete 16S rRNA gene sequences were
313 extracted from the genome sequences of 40 *Bacillus* strains (24 *B. subtilis,* 4 *B. licheniformis,* 4 *B.*
314 *paralicheniformis,* 4 *B. amyloliquifaciense,* 4 *B. velezensis*) and were aligned using MUSCLE with
315 default parameters. Phylogenetic tree was estimated by Maximum Likelihood Phylogeny statistical
316 method with 1000 Bootstrap replications in MEGA-X with default parameters where substitution model
317 and tree interference were developed by Tamura-Nei method and Nearest-Neighbor-Interchange (NNI)
318 ML Heuristic Method respectively.
319
320 **Pangenome and COGs protein function analysis**
321 Pan- and core-genome analyses of 30 *B. subtilis* genomes were performed using GView[40] and BPGA
322 v1.3[44]. GView was used to generate pangenome atlas as well as to study both pan and core genomes.
323 Besides, BPGA v1.3 was used with default clustering algorithm for orthologous gene identification and
324 clustering with a 50% sequence identity cut-off. BPGA software uses USEARCH, CD-HIT, OrthoMCL
325 and MUSCLE software for the orthology analyses and powerlaw regression and exponential curve fit for
326 the pangenome, core genome developments, phylogeny, subset analysis, and KEGG & COGs mapping.
327 Pangenome plot was extracted by plotting the total number of distinct gene families identified with each
328 subsequent addition of a genome against the number of genomes. Core-genome plot was extracted by
329 plotting the total number of shared genes with the addition of each genome against the number of
330 genomes. To avoid sampling biases, 30 iterations (random samples) were used for extracting these plots.
331 Representative sequences of both pan- and core-genomes were used for comprehensive functional

analyses. Pan and core phylogeny were created from default parameter. COGs protein function of 30 *B. subtilis* strains were predicted by WebMGA server[39] and their comparative study were observed by R heatmap.

References

1. Rao, M.B., Tanksale, A.M., Ghatge, M.S. & Deshpande, V.V. Molecular and biotechnological aspects of microbial proteases. *Microbiol Mol Biol Rev*. **62**(3), 597-635 (1998).
2. Buescher, J.M. & Margaritis, A. Microbial biosynthesis of polyglutamic acid biopolymer and applications in the biopharmaceutical, biomedical and food industries. *Crit. Rev. Biotechnol.* **27**, 1–19 (2007).
3. Stahl, M.L. & Ferrari, E. Replacement of the *Bacillus subtilis* subtilisin structural gene with an in vitro-derived deletion mutation. *Journal of Bacteriology.* **158**(2), 411-418. ISSN 0021-9193 (1984).
4. Wong, S.L., Price, C.W., Goldfarb, D.S. & DoI, R.H. The subtilisin E gene of *Bacillus subtilis* is transcribed from a sigma 37 promoter in vivo. *Proceedings of the National Academy of Sciences.* **81**(4), 1184-1188. ISSN 0027-8424 (1984).
5. Sloma, A.L. *et al.* Bacillopeptidase F of *Bacillus subtilis*: purification of the protein and cloning of the gene. *Journal of bacteriology.***172**(3), 1470-1477. ISSN 0021-9193 (1990).
6. Wu, X.C, Nathoo, S., Pang, A.S., Carne, T. & Wong, S.L. Cloning, genetic organization, and characterization of a structural gene encoding bacillopeptidase F from *Bacillus subtilis*. *Journal of Biological Chemistry.* **265**(12), 6845-50. ISSN 0021-9258 (1990).
7. Bruckner, R., Shoseyov, O. & Doi, R.H. Multiple active forms of a novel serine protease from *Bacillus subtilis*. *Molecular and General Genetics MGG*. **221**(3), 486-490. ISSN 0026-8925 (1990).
8. Sloma, A., Ally, A., Ally, D. & Pero, J. Gene encoding a minor extracellular protease in *Bacillus subtilis*. *Journal of Bacteriology.* **170**(12), 5557-5563. ISSN 0021-9193 (1988).
9. Rufo, G.A., Sullivan, B.J., Sloma, A.L. & Pero, J.A. Isolation and characterization of a novel extracellular metalloprotease from *Bacillus subtilis*. *Journal of Bacteriology.* **172**(2), 1019-1023, ISSN 0021-9193 (1990).
10. Sloma, A. *et al.* Gene encoding a novel extracellular metalloprotease in *Bacillus subtilis*. *Journal of Bacteriology.* **172**(2), 1024-1029. ISSN 0021-9193 (1990).

11. Tran, L.O., Wu, X.C. & Wong, S.L. Cloning and expression of a novel protease gene encoding an extracellular neutral protease from *Bacillus subtilis*. *Journal of bacteriology*. **173**(20), 6364-6372. ISSN 0021-9193 (1991).
12. Yang, M.Y., Ferrari, E.U. & Henner, D.J. Cloning of the neutral protease gene of *Bacillus subtilis* and the use of the cloned gene to create an in vitro-derived deletion mutation. *Journal of Bacteriology.* **160**(1), 15-21. ISSN 0021-9193 (1984).
13. Sloma, A.L. *et al.* Cloning and characterization of the gene for an additional extracellular serine protease of *Bacillus subtilis*. *Journal of bacteriology.***173**(21), 6889-6895. ISSN 0021-9193 (1991).
14. Margot, P. & Karamata, D. The wprA gene of Bacillus subtilis 168, expressed during exponential growth, encodes a cell-wall-associated protease. *Microbiology.* **142**(12), 3437-3444, ISSN 1350-0872 (1996).
15. Kawamura, F. & Doi, R.H. Construction of a *Bacillus subtilis* double mutant deficient in extracellular alkaline and neutral proteases. *J Bacteriol.* **160**, 442-444 (1984).
16. Strauch, M.A. & Hoch, J.A. Transition-state regulators: sentinels of *Bacillus subtilis* post-exponential gene expression. *Mol. Microbiol.* **7**, 337–342 (1993).
17. Hoq, M.M., Siddiquee, K.A.L., Kawasaki, H. & Seki, T. Keratinolytic activity of some newly isolated *Bacillus* species. *J. Biol. Sciences*. **5**(2), 193-200 (2005).
18. Salaheen, S., Mamun, M.A.A., Khan, S.N & Hoq, M.M. Improvement of *Bacillus licheniformis* MZK05 by mutation for increased production of keratinase. *Dhaka Univ. J. Biol. Sci.* **24**(1), 17-23 (2015).
19. Mamun, M.A.A. *et al.* Development of an Alternative Enzyme-assisted Dehairing Method of Animal Skins using Proteases from *Bacillus licheniformis* MZK05M9. *Bangladesh J. Microbiol.* **2**, 33-7 (2015).
20. Mamun, M.A.A., Khan, M.M., Akand, M.N., Khan, S.N. & Hoq, M.M. Characterization of an alkaline protease with high quality bating potential in leather processing from *Bacillus licheniformis* MZK05M9 mutant. *Intl. J. Biol. Res.* **3**(1):36-41 (2015).
21. Mamun, M.A.A., Mian, M.M., Saifuddin, M., Khan, S.N. & Hoq, M.M. Optimization of fermenting medium by statistical method for production of alkaline protease by Bacillus licheniformis MZK05M9. *J. Applied Biol. Biotechnol.* **5**(6), 24-28 (2017).
22. DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491 (2011).
23. Seemann, T. Snippy: fast bacterial variant calling from NGS reads. *Snippy: fast bacterial variant calling from NGS reads*. Available online at: https://github.com/tseemann/snippy (2015).

24. Kunst, F. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature.* **390**(6657), 249-256 (1997).
25. Hasan, M.M., Khan, S.N., Karim, M.M., Begum, A. & Hoq MM. Complete genome and plasmid sequence of a novel *Bacillus sp*. BD59S, a parasporal protein synthesizing bacterium. *3 Biotech.* **9**(9), 318 (2019).
26. Morimoto, T. *et al.* Enhanced recombinant protein productivity by genome reduction in *Bacillus subtilis*. *DNA Res.* **15**, 73–81 (2008).
27. Posfai, G. *et al.* Emergent properties of reduced-genome *Escherichia coli*. *Science.* **312**, 1044–1046 (2006).
28. Mesak, L.R., Mesak, F.M. & Dahl, M.K. Expression of a novel gene, gluP, is essential for normal *Bacillus subtilis* cell division and contributes to glucose export. *BMC microbiology.* **4**(1), 13 (2004).
29. Rather, P. Role of rhomboid proteases in bacteria. *Biochimica et Biophysica Acta (BBA)-Biomembranes.* **1828**(12), 2849-2854 (2013).
30. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, 237–243 (2015).
31. Heel, A.J.V. *et al.* BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.* **46**, 278-281 (2018).
32. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
33. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. **19**, 455–477 (2012).
34. Darling, A.E., Mau, B. & Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one*. **5**(6), e11147 (2010).
35. Darling, A.C., Mau, B., Blattner, F.R. & Perna, N.T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*. **14**(7), 1394-1403 (2004).
36. Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC genomics*. **9**(1), 75 (2008).
37. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic acids res.* **44**(14), 6614-24 (2016).

38. Cantalapiedra, C.P. *et al.* eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *bioRxiv* (2021).
39. Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC genomics.* **12**(1), 1-9 (2011).
40. Petkau, A., Stuart-Edwards, M., Stothard, P. & Van Domselaar, G. Interactive microbial genome visualization with GView. *Bioinformatics.* **26**(24), 3125-3126 (2010).
41. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*. **10**(6), 845-858 (2015).
42. Lilkova, E. *et al.* The PyMOL Molecular Graphics System, *Version 2.0* Schrodinger, LLC (2015).
43. Alikhan, N. F., Petty, N. K., Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics*, **12**, 402 (2011).
44. Chaudhari, N. M., Gupta, V. K. & Dutta, C. (2016). BPGA-an ultra-fast pan-genome analysis pipeline. *Scientific reports.* **6**(1), 1-10.

Acknowledgements

This work was partially financed by the Ministry of Education (Ref: MoE/S-17/10 M-15/2007(Part-1)/146) and Ministry of Science and Technology (Ref: MOSICT/S-9/B.Anu.Pr/86/2003-2004/386), Government of the People's Republic of Bangladesh.

Author contributions

MMH isolated the organisms and supervised the experimental work. MMHasan analyzed the genome sequencing data, and drafted the manuscript. MMK did meticulous review of the manuscript. MAAM performed the supportive laboratory works for optimization of fermentation of alkaline serine proteases. SNK supervised the molecular biology part of the work, reviewed and finalized the manuscript. All authors read, and approved the manuscript.

Competing interests

The authors declare no competing interests.

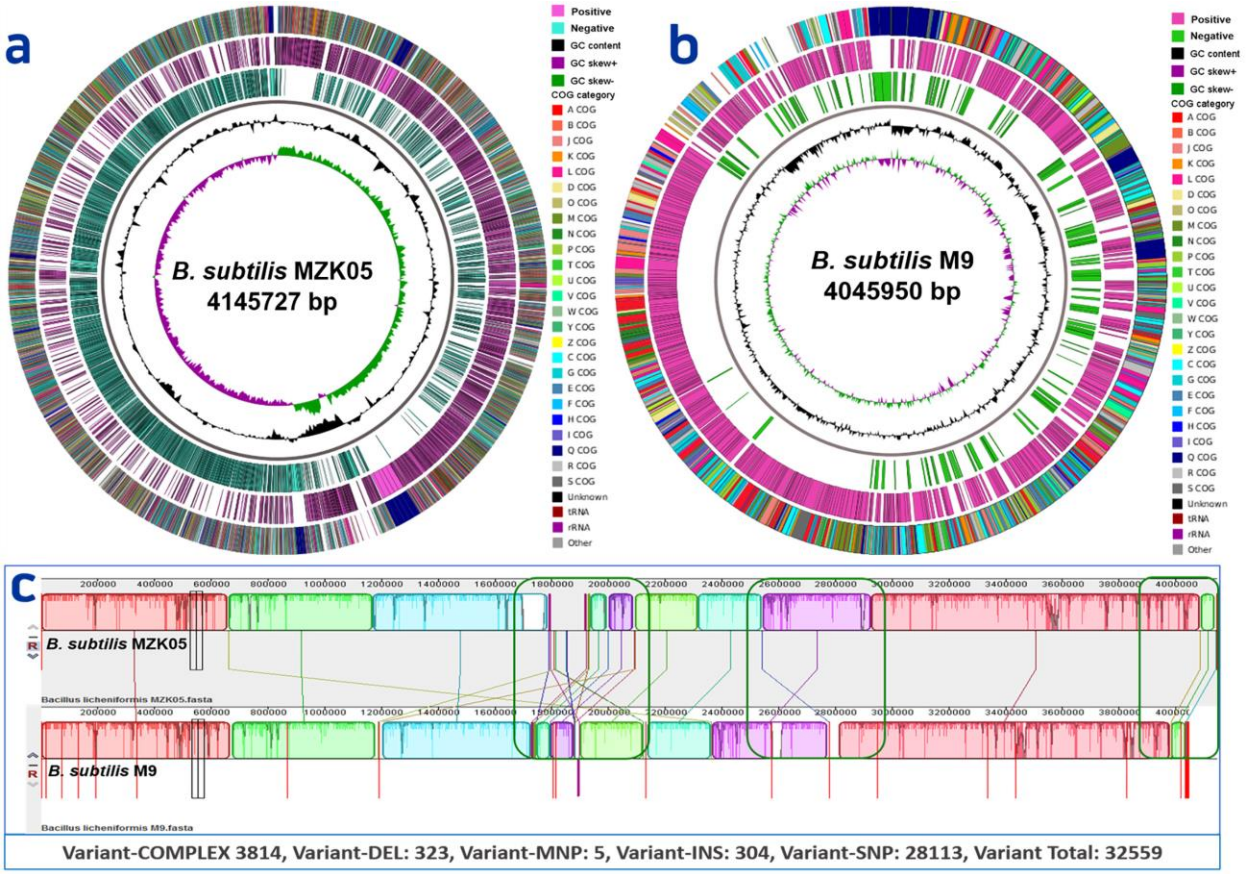


Figure 1. Circular graphical maps of *B. subtilis* MZK05 complete genome (a), and, *B. subtilis* M9 draft genome (b). Circles indicate, from inside to outwards: GC skew; GC content; total genes; rRNAs (green in chromosome); tRNAs (blue in chromosome); pseudogenes (sky blue); CDS on reverse strand (olive); CDS on forward strand (maroon); (c) Genome-wide alignment performed with Mauve progressive alignments compares the differences between *B. subtilis* MZK05 and *B. subtilis* M9 conserved genome regions. Each syntenical placement of the homologous region of the genome is represented as unique colored block, whilst divergent regions are seen as an empty block or line. Indentations within boxes highlight small mutations. Total number of variants after snippy 3.2-dev variant calling analyses are indicated beneath the alignment.

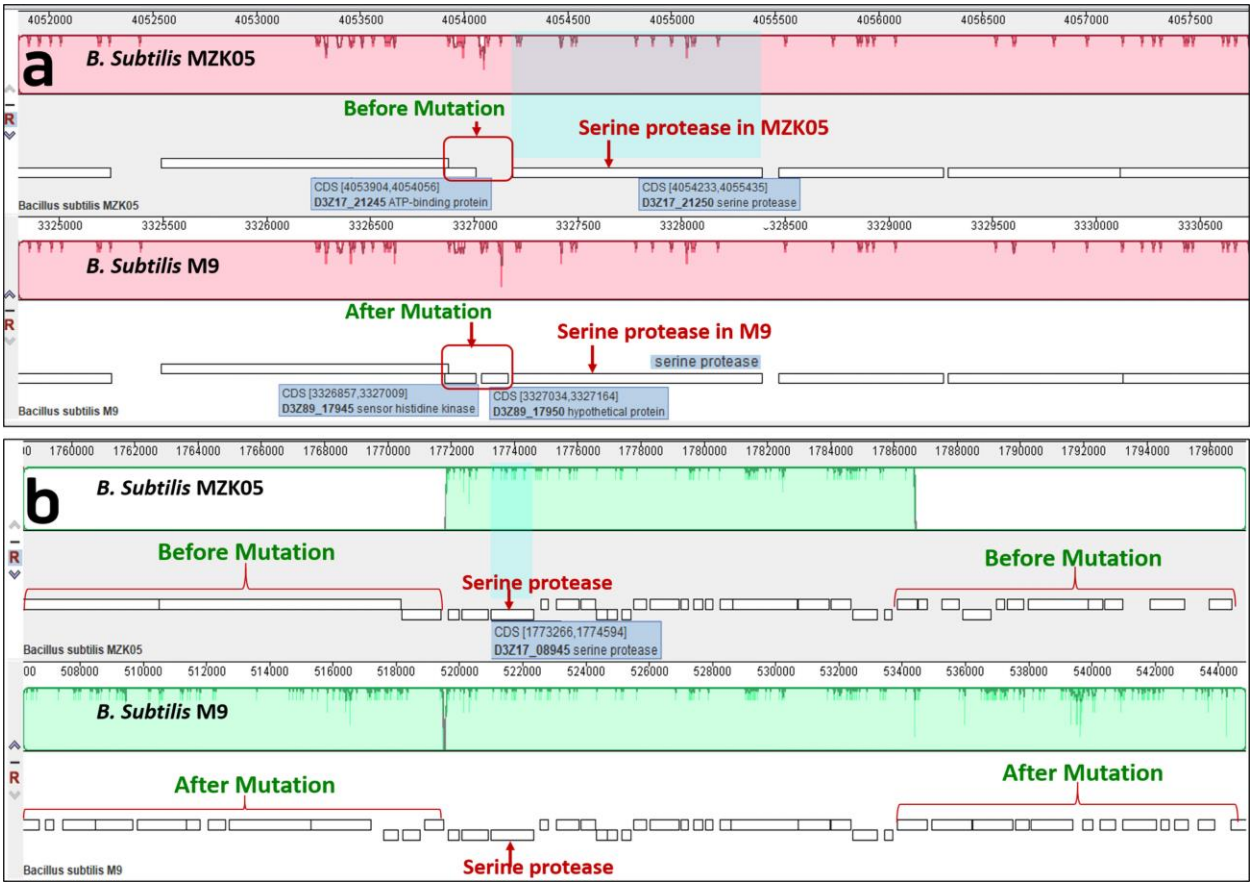


Figure 2. Multiple major changes observed in very close region of two serine protease genes. A single mutated region is noticed just upstream of a serine protease gene (a), and in (b) Two large regions found deleted just before and after another serine protease gene.

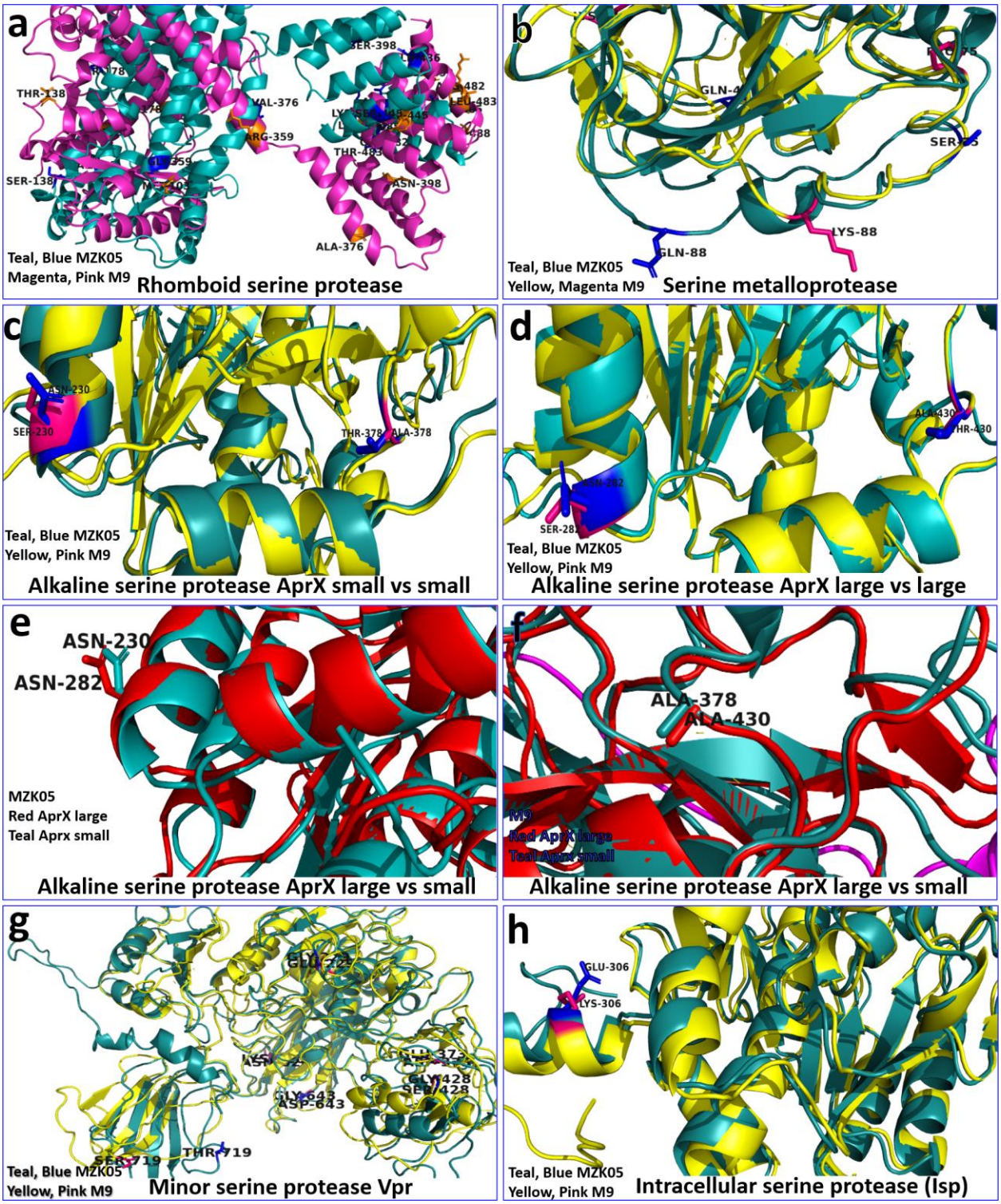
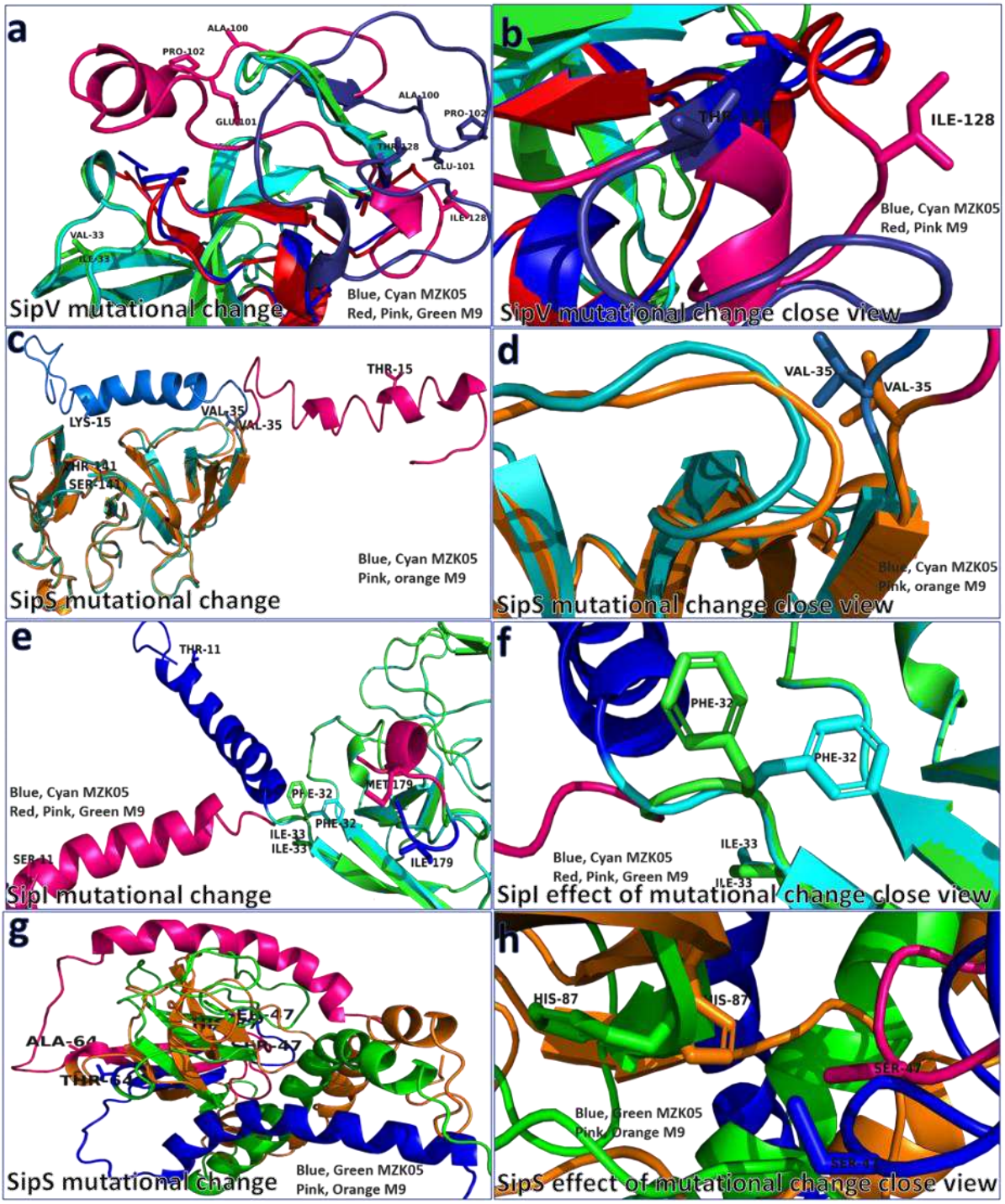


Figure 3. Depiction of amino acid mutations in serine proteases: (a) rhomboid serine protease, (b) extracellular serine metalloprotease, (c-f) AprX, (g) Vpr, and (h) intracellular serine protease.



481

482 Figure 4. Mutational amino acid changes in signal peptidases (a-b) SipV, (c-d) SipS, (e-f) SipI, and (g-h)

483 SipS.

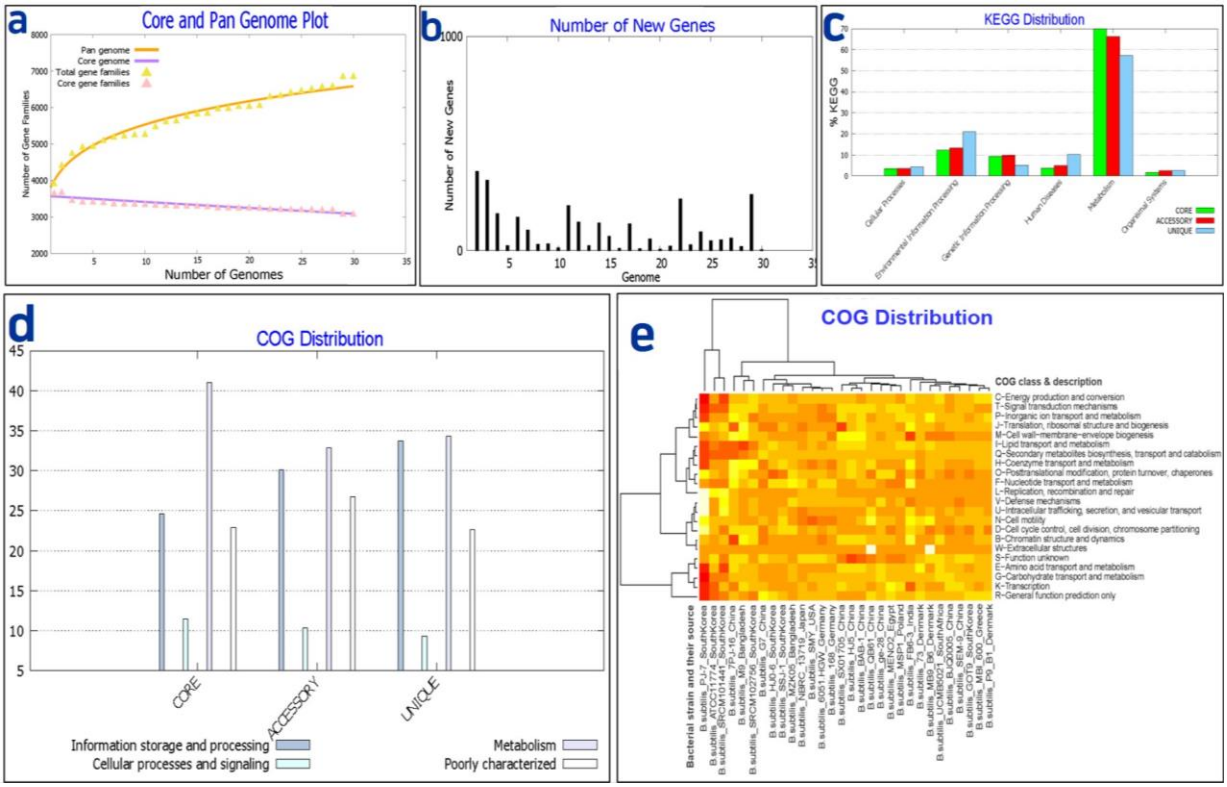


Figure 5. Pangenome plots of *B. subtilis*. (a) the pan and core genome box plot curves shows the evolution of the pangenome (blue) as an accumulated sum of gene families and the core genome (orange), as described in the methods, (b) The columns give the number of new gene families introduced with the addition of each genome into the consideration, (c) the KEGG and (d) COGs distribution of the representative proteins in the core, accessory, and unique genome, and (e) The COGs frequency heatmap of the representative proteins present in all 30 genomes, where, the vertical axis shows the abundance of functional COGs category, and the strains are located on horizontal axis.

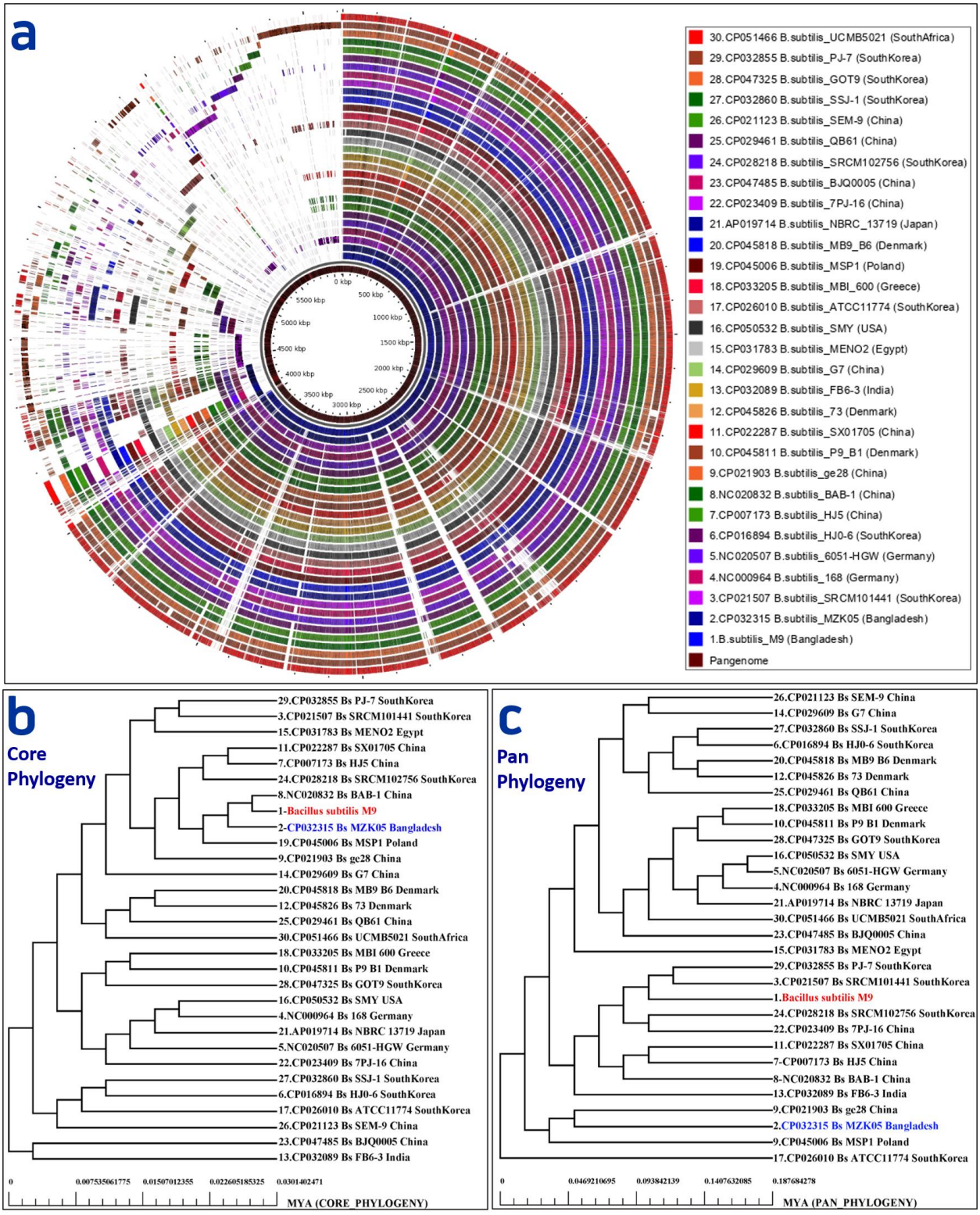


Figure 6. (a) Circular pangenome BLAST atlas of 30 *B. subtilis* genomes, where mutant strain *B. subtilis* M9 was used as the reference strain. The Atlas shows diversities and new genomic properties as well as

497 the genes of *B. subtilis* strains of different countries of origin. The inner most circle shows the pangenome
498 (Brown color). After pangenome, rings 1-2 represents *B. subtilis* M9 and *B. subtilis* MZK05 genome
499 respectively, rings 3-30: *B. subtilis* genome from different countries. The positions of different genomes
500 from inner to outward are mentioned in the legend of right-side box with names of the strain and countries
501 of origin. In pangenome, the position of the new genes in the different genomes are shown between 4000
502 kbp to 6000 kbp. Genomic regions unique to reference mutant strain and not appearing in other strains are
503 lightly colored and blank. (b) relative evolutionary divergence of *B. subtilis*, a Neighbor Joining (NJ) core
504 genome phylogenetic tree, and (c) a pan-genome phylogenetic tree, based on the pangenome of 30 *B.*
505 *subtilis* strains. The bootstrap values are marked below position of each tree. Blue color indicates wild
506 strain *B. subtilis* MZK05, while red color indicates mutant strain *B. subtilis* M9.
507
508

Supplementary Files

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

- SupplementaryinformationBacillussubtilisMZK05M9.pdf