Letter to the Editor

Genomic Diversity and Evolution of *Bacillus subtilis*[*]

YU Gang[1,2,a], WANG Xun Cheng[3,a], TIAN Wang Hong[2], SHI Ji Chun[2], WANG Bin[2], YE Qiang[2], DONG Si Guo[2], ZENG Ming[2,#], and WANG Jun Zhi[1,2,#]

Bacillus subtilis **is the focus of both academic and industrial research. Previous studies have reported a number of sequence variations in different *B. subtilis* strains. To uncover the genetic variation and evolutionary pressure in *B. subtilis* strains, we performed whole genome sequencing of two *B. subtilis* isolates, KM and CGMCC63528. Comparative genomic analyses of these two strains with other *B. subtilis* strains identified high sequence variations including large insertions, deletions and SNPs. Most SNPs in genes were synonymous and the average frequency of synonymous mutations was significantly higher than that of the non-synonymous mutations. Pan-genome analysis of *B. subtilis* strains showed that the core genome had lower dN/dS values than the accessory genome. Whole genome comparisons of these two isolates with other *B. subtilis* strains showed that strains in different subspecies have similar dN/dS values. Nucleotide diversity analysis showed that *spizizenii* subspecies have higher nucleotide diversity than *subtilis* subspecies. Our results indicate that genes in *B. subtilis* strains are under high purifying selection pressure. The evolutionary pressure in different subspecies of *B. subtilis* is complex.**

Bacillus subtilis is an aerobic, endospore-forming, rod-shaped, Gram-positive bacterium. It is a remarkably diverse bacterial species that is capable of growth within many environments. Early analysis in a collection of *B. subtilis* isolates obtained from desert soils in a small sample of loci (*rpoB*, *polC*, and *gyrA*) with restriction fragment length polymorphisms method showed that there was considerable diversity[1]. Despite the fact that all of these strains exhibited 99% sequence identity in their 16S rRNA genes, previous microarray-based comparative genomic M-CGH analyses have found the genetic heterogeneity among members of this species is significant[2]. The microarray hybridization intensities analysis shows that nearly one-third of *B. subtilis* 168-specific genes exhibited variability. Recent high throughput sequencing have been used to investigate whole genome variation in *B. subtilis* to uncover the possible genetic mechanisms involved in physiological or phenotypic variation within the species. Two subspecies, *subtilis.* and *spizizenii*, have been identified in *B. subtilis*[1]. Whole genome re-sequence analyses showed high genome sequence variation of inter- and intra- subspecies in this species. Study the genome sequence of subspecies *spizizenii* W23 have found significant sequence variation with subspecies *subtilis* 168[3]. Further analysis of strains in same subspecies (*subtilis*) found that strain BSn5 has lots of insertions and deletions with length more than 5 kb compared with *B. subtilis* 168[4].

To explore the genome diversity in *B. subtilis* strains in more detail, we conducted whole genome resequencing of two wild isolated *B. subtilis* strains using next-generation sequencing, and compared their genome sequences with other *B. subtilis* strains. *B. subtilis* strain KM was obtained from the National Institutes for Food and Drug Control of China, Beijing; and CGMCC63528 was obtained from the China General Microbiological Culture Collection Center, Beijing. Whole-genome sequencing of the two *B. subtilis* strains was performed on an Illumina Genome Analyzer IIx apparatus, generating 8,200,000 and 9,600,000 high quality 100-bp paired-end reads for strains KM and CGMCC63528, respectively. Reads were mapped to the genome of

doi: 10.3967/bes2015.087

[*]This work was supported by Significant new drugs creation, 12th Five-Year plan special science and Technology Major (2013ZX09304101); the National High Technology Research and Development Program of China (863 Program) (2014AA022210).

1. The State Key Laboratory of Cancer Biology, Department of Biopharmaceutics, School of Pharmacy, Fourth Military Medical University, Xi'an 710032, China; 2. Division of Enteric Bacterial Vaccines, National Institutes for Food and Drug Control, Beijing 100050, China; 3. Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua university, Beijing 100084, China

domesticated *B. subtilis* strain 168 (NC_000964.3) using Bowtie[5]. Overall, 80% of the obtained reads were successfully mapped to a unique position on the bacterial reference genome, with up to two mismatches for each strain. The average sequencing depths of these two strains were 160-fold (KM) and 180-fold (CGMCC63528), and both covered about 96.5% of the reference genome. To analyze the read coverage along the genome in greater detail, we divided the reference genome into windows of 10 kb, and calculated reads per kilobase per million reads (RPKM) for each window. A higher RPKM value indicates a higher read coverage for the corresponding region. As shown in Figure 1A, most of the regions showed a high read density (RPKM>2000) in both of the strains, except for three regions (530,000-550,000 bp, 2,060,000-2,070,000 bp, and 2,160,000-2,280,000 bp) with RPKM values <20. All three regions were absent in both strains.

To determine whether the low density read coverage for these three regions was caused by our read mapping strategy in Bowtie, we changed the mapping parameters to include reads that mapped to multiple positions. Again, very few reads mapped to these three regions. These fragments do not appear to be junk DNA in strain 168, as the gene density is comparable with other regions; more than 200 genes are located in the three fragments (Figure 1A). The first region contains transposon genes, including *ydcQ*, *ydcR*, *yddB*, and *yddH*. Three of these genes, along with *yddC*, *yddD*, *yddE*, and *yddG*, also located in this region, belong to type IV secretion systems[6]. The second region contains Y-family DNA polymerase genes such as *yozL*, *yozK*, and *yobH*, which are correlated with DNA damage repair[7]. The third region spans 120 kb and contains more than 160 genes, including some phage-related genes.

We further assembly the genomes of these two isolates and compare to the *B. subtilis* 168 to find genome variation further. SOAPdenovo were used to assembly the KM and CGMCC63528 genomes. The draft genomes of these two strains have 12 and 89 contigs, ranging in size from 1421 bp to 2,084,809 bp, and 101 bp to 1,052,683 bp, with a total predicted length of 4,030,959 bp and 4,095,208 bp for strains KM and CGMCC63528, respectively. The N50s of the draft genomes are 678,729 bp (KM) and 557,470 bp (CGMCC63528). Glimmer was used to predict protein genes in KM and CGMCC63528. There are 4177 and 4262 ptotein genes found in KM and CGMCC63528 respectively, which is similar in *B.subtilis* 168 (4175). The whole genome alignment of these two isolates with *B. subtilis* 168 showed previous identified 3 regions in *B. subtilis* 168 are clearly deleted in *B. subtilis* 168. We also found lots of positions have insertion and deletions in both strains. The length of insertion or deletion ranged from 1 bp to more than 1000 bp. 4 and 10 DNA fragments (>5 kb) were found inserted and 7 DNA fragments (>5 kb) were found deleted in KM and CGMCC53628 separately.

Using *B. subtilis* 168 as the reference genome, 28,460 and 32,893 SNPs were detected in KM and CGMCC63528, respectively (Figure 1B). Of the identified SNPs, 12,494 were common to both strains. In comparison to *B. subtilis* 168, the average number of SNPs ranged from 1 per 148 bp in KM to 1 per 128 bp in CGMCC63528. We classified identified SNPs into four different categories: transversions, synonymous SNPs (sSNPs), nonsynonymous SNPs (non-sSNPs), and intergenic SNPs (Figure 1C). In both strains the percentage of intergenic SNPs was significant lower (around 9%) than other categories. Further analysis shows that there are about 4% of the genome are intergenic region, indicating a higher level of point mutations in non-coding regions than in coding regions. The majority of coding region SNPs (69%) found in both strains were synonymous (69%). We also found 12 and 10 nonsynonymous SNPs caused premature stop codons both in KM and CGMCC63528, respectively. Among these genes, *yesK* annotated as unknown function, has nonsynonymous SNPs in both strains. The percentage of SNP transition (69%) was similar in both strains, suggesting a substitution bias in favor of nucleotide substitution within the purine or pyrimidine group. Analysis of SNP distribution around the genome showed similar distribution profiles in the two genomes (Figure 1C, R=0.41, P<2.2e-16).

There are about 2800 and 3200 genes harbor SNPs in KM and CGMCC63528 respectively. Most of the genes have their SNPs density no more than 2 per 100 bp. The SNPs number in genes ranges from 1 to 181 in both strains. We found 8 and 7 genes harbor more than 100 SNPs in KM and CGMCC63528 respectively. Among these genes, the putative lytic transglycosylase gene *yqbO*, the antimicrobial peptide (AMP) biosynthetic genes *srfAA* and other four genes (*pskJ*, *pksL*, *pksM*, *pksN*) are common to both strains. However, the SNP density for all these genes are not significantly different to the genome average (three standard deviation), most of these genes are 10 kp, which may account for the high

number of SNPs. We did find 58 and 65 genes with a SNP density higher than 1 per 30 bp, which was three standard deviations higher than the average SNP density. More than 50% of these consisted of genes encoding hypothetical proteins. Twenty genes with high SNPs density were found in both strains, including genes related to translation (*yefA*), cell wall biogenesis (*pbpA*), inorganic ion transport and metabolism (*yclQ*), posttranslational modification (*yrkI*) and amino acid transport and metabolism (*gerBB*).

The high number of SNPs may represent evidence for selection pressure on these genes. We thus calculated values of synonymous differences per synonymous site (dS) and non-synonymous differences per non-synonymous site (dN) across all SNP harboring loci. Genes with 0.01<dS<2 and dN<2 were considered for further analysis. For 2022 genes passed our dS/dN calculation filter criteria, the average dN/dS value was 0.11; most of them are ranged from 0 to 0.5, suggesting that the average frequency of synonymous mutation was significantly higher than that of the non-synonymous mutations; i.e. there was no evidence for diversifying selection. Among the 20 genes found earlier with high SNP density, only one gene, *yycQ*, which was predicted as a membrane protein showed significantly a high dN/dS values (above the mean+3 standard deviations) and might be under diversifying selection. Meanwhile, we found 21 genes with average SNP density in both strains but significantly high dN/dS values, including genes related to transcription (*yxjO*), signal transduction (*natK*), cell motility and vesicular transport (*comGD*), cell wall biogenesis and carbohydrate transport and metabolism (*epsI*), and posttranslational modification and energy production and conversion (*stoA*). We further examined whether certain functional classes of genes were under positive selection by comparing dN/dS values according to the Clusters of Orthologous

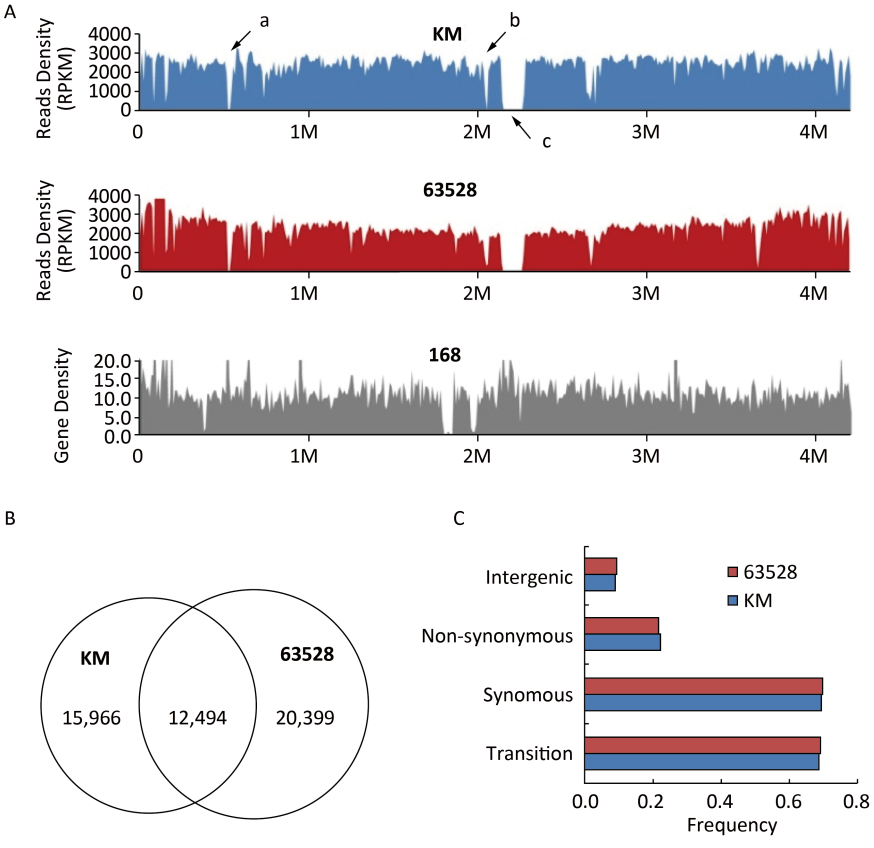

Figure 1. Sequence variation of KM and CGMCC63528 to 168. (A) Distribution of sequence reads from whole-genome resequencing of *B. subtilis* strains KM (top), CGMCC63528 (middle), and gene density for 168 (bottom). 530,000-550,000 bp (a), 2,060,000-2,070,000 bp (b), and 2,160,000-2,280,000 bp (c). (B) Venn diagram showing SNPs from KM and CGMCC63528, as well as those common to the two strains; (C) summary of SNPs in *B. subtilis* strains.

Group (COG) classification. We found no significant overrepresentation of any functional class in the gene group with higher dN/dS values than genome average.

We constructed a phylogenetic tree to better understand the relationship between the two strains, 168, and ten other *B. subtilis* strains, including W23 and TU-B-10, which belong to the *spizizenii* subspecies. The remaining strains all belonged to the *subtilis* subspecies. The genome sequences of these strains were aligned against the 168 genome to produce whole-genome SNP data for each strain. We then constructed a neighbor-joining tree using the SNP data. The phylogenetic tree consisted of three main clusters, one of which contained only the *B. subtilis* subsp. *subtilis* strain BEST195, which is used in the production of natto, a fermented bean product[8] (Figure 2). This suggested that the BEST195 strain may be distinct from other *subtilis* subspecies strains[8]. Three *subtilis* subspecies strains, XF-1, BAB-1, and RO-NN-1, were clustered with the *spizizenii* subspecies strains W23 and TU-B-10, indicating a close evolutionary relationship. The other five subspecies strains and the two newly-sequenced environmental strains were clustered together. In this group, the two isolated strains are not as similar to 168 as BSP1. However, they are both more similar to 168 than to endophytic BSn5 strains[4]. These results indicated that our two wild strains belong to the *subtilis* subspecies. As shown in Figure 2, KM is closer to 168 than CGMCC63528, which agrees with the SNP results in the previous section.

To obtain a deeper understanding of the pan-genome of *B. subtilis*, 53,451 protein coding genes obtained from the 13 *B. subtilis* strains were clustered using the CD-HIT algorithm with an 80% sequence identity cut-off. A total of 6743 clusters were identified, and of these, 3380 orthologs (50%) were identified in all 13 strains as the *B. subtilis* core genome (Figure 3A). The remaining variable 3363 clusters were defined as the *B. subtilis* accessory genome. Species-specific genes were identified among the five tested species (Figure 3A). The relative balance between the core and accessory genomes was completely different to that reported for other bacterial species. In previous studies in *Escherichia coli*[9] and *Sinorhizobium*[10], only 6% and 8% of genes were shared by all the strains (core genome) analyzed.

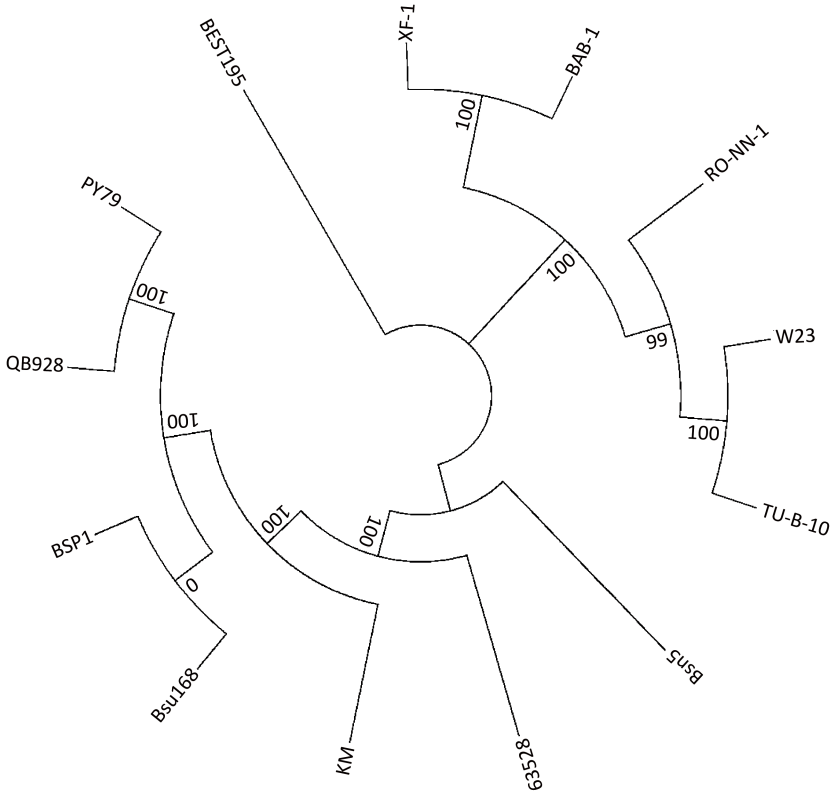

Figure 2. Phylogeny of 13 *B. subtilis* strains. Concatenated whole-genome SNPs were used to construct the phylogenetic tree using the neighbor-joining method.

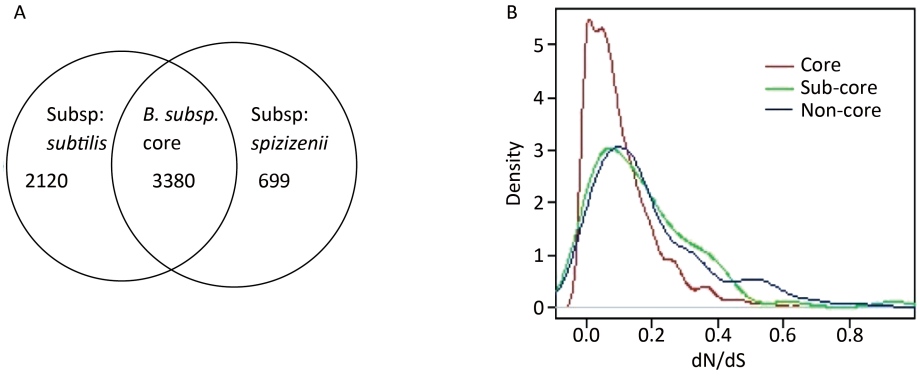

Figure 3. Pan-genome analysis of *B. subtilis* strains. (A) Core orthologous and subspecies-specific unique genes in *B. subtilis*; (B) gene densities of dN/dS values in core genes of *B. subtilis* (core), core genes of *B. subtilis* subsp. *subtilis* (sub-core) and other genes (non-core).

To characterize the evolutionary pressure on genes encoded by the core and/or accessory genome in *B. subtilis*, we divided the genes into three groups: core gene in *B. subtilis*, core orthologous genes in *B. subtilis* subsp. *subtilis*, and other genes (defined as non-core genes). For the 1796, 74, and 152 genes identified in each group, we calculated the gene densities using dN/dS values. As shown in Figure 3B, most dN/dS values for core genes in *B. subtilis* group were significantly lower than other two groups (Wilcox sum test, $P<0.01$). However, there was no significant difference for genes between core genes in *B. subtilis* subsp. *subtilis*, and other genes. These results suggested that genes in the core genome of *B. subtilis* have significant higher purifying selection pressure than those in the accessory genome.

We used the whole genome sequences of the *B. subtilis* isolates for genetic diversity and selection intensity analysis. The nucleotide diversity (*Л* value) for the whole genome sequences of the KM and CGMCC63528 were similar compared to *B. subtilis* 168, ranging from 0.0098 to 0.0122. We also found the similar nucleotide diversity level with *Л* value 0.0137 when compared with the whole genome sequence of eight *subtilis* subspecies from different sources. *B. subtilis* subsp. *subtilis* strain BEST195 is thought to be different with other *subtilis* subspecies strain. However, the *Л* value did not change when this strain was added to the previous 8 *subtilis* subspecies to calculated nucleotide diversity. We then calculated nucleotide diversity level between two *B. subtilis* subsp. spizizenii strains W23 and TU-B-10, and obtained a *Л* value that was more than two-fold (0.032) higher than other *B. subtilis* strains. The dN/dS values for the whole genome sequences were similar between KM and CGMCC63528 when compared with *B. subtilis* 168: about 0.1 for both isolates. We also found the dN/dS values were similar in different group of *B. subtilis* subsp. *subtilis* strains, and were comparable to KM and CGMCC63528 isolates. Unlike the result from nucleotide diversity, the dN/dS value for *B. subtilis* subsp. spizizenii strains were comparable to other *B. subtilis* strains (*Z* test, $P=0.82$), suggesting a complex selection pressure in different *B. subtilis* subspecies.

Our study provides evidence supported the hypothesis that the evolution of *B. subtilis* is clearly a complex and diversified process. Several questions remain further in-depth investigations, such as whether the subspecies functional characters are affected by the deletion of specific genes and disabling of specific metabolic pathways. In addition, further studies using a larger sample of *B. subtilis* isolates from diverse lineages are warranted to better understand the evolution of *B. subtilis* strains in different subspecies.

Nucleotide Sequence Accession Numbers: This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession no. JWHP00000000 (KM) and JWHO00000000 (CGMCC63528)

[a]These authors contributed equally to this work.

[#]Correspondence should be addressed to WANG Jun Zhi, Tel: 86-10-67095782; E-mail: wangjz@nifdc.org.cn; ZENG Ming, Tel: 86-10-67095416; Fax: 86-10-67058402; E-mail: zengming@nifdc.org.cn

Biographical note of the first author: YU Gang, male, born in 1981, Doctor candidate, majoring in microbiology.

Received: October 19, 2014;
Accepted: January 31, 2015

REFERENCES

1. Roberts MS, Cohan FM. Recombination and Migration Rates in Natural Populations of *Bacillus subtilis* and *Bacillus mojavensis.* Evolution, 1995; 49, 1081-94.
2. Earl AM, Losick R, Kolter R. *Bacillus subtilis* genome diversity. J Bacteriol, 2007; 189, 1163-70.
3. 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, 2033-41.
4. Deng Y, Zhu Y, Wang P, et al. Complete genome sequence of *Bacillus subtilis* BSn5, an endophytic bacterium of *Amorphophallus konjac* with antimicrobial activity for the plant pathogen *Erwinia carotovora* subsp. *carotovora*. J Bacteriol, 2011; 193, 2070-1.
5. Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009; 10, R25.
6. Bi D, Liu L, Tai C, et al. SecReT4: a web-based bacterial type IV secretion system resource. Nucleic Acids Res, 2013; 41, D660-5.
7. Gioia J, Yerrapragada S, Qin X, et al. Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. PLoS One, 2007; 2, e928.
8. Nishito Y, Osana Y, Hachiya T, et al. Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data. BMC Genomics, 2010; 11, 243.
9. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol, 2010; 60, 708-20.
10. Sugawara M, Epstein B, Badgley BD, et al. Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. Genome Biol, 2013; 14, R17.