# Monitoring of Biotechnological Strains of the *Bacillus subtilis* / *Bacillus amyloliquefaciens* Group in Natural Habitats by using Multilocus Genetic Barcodes

Oleg N. Reva
Bioinformatics and Computational Biology Unit, Dep. Biochemistry,
University of Pretoria
Pretoria, South Africa
e-mail: oleg.reva@up.ac.za

***Abstract*—*Bacillus subtilis, B. amyloliquefaciens* and other related strains isolated from rhizosphere are widely used in biotechnology as Plant Growth Promoting Rhizobacteria (PGPR) showing variable activity in different soils and on different plants. Despite diverse bioactivity, they are almost indistinguishable by phenotype and by 16S rRNA that makes it problematic to trace them down in nature. Spores of these microorganisms are abundant in many habitats, but it remains unclear whether these organisms thrive in these habitats or simply contaminated the samples? In this work, several new computational approaches were proposed for design, evaluation and application of multilocus genetic barcodes for identification and monitoring of biotechnological strains. The barcodes containing 150 marker genes were designed for 35 strains of the *B. subtilis* group and tested on publically available metagenomic datasets. Clear habitat preferences were observed for different subspecies and even individual strains. Also, an approach of evaluation and improving of sensibility and sensitivity of barcodes was proposed.**

***Keywords-genetic barcoding; plant growth promoting; Bacillus; metagenomics.***

## I. INTRODUCTION

*Bacillus subtilis* was one of the first known bacterial species described in 1835 by Ehrenberg as *Vibrio subtilis* and then renamed to *Bacillus subtilis* in 1872 by Cohn [1]. This spore-forming bacterium caught eye because of its abundance and ease of isolation. From then on *B. subtilis* has become one of the most popular model organism used in bacteriological and genetic laboratories. At dawn of the genomic era it became obvious that *B. subtilis* in fact is a conglomerate of several closely related taxa vernacularly termed the *B. subtilis / B. amyloliquefaciens* group. Many of these organisms found a wide application in biotechnology as biopesticides and plant growth promoters [2], lytic enzyme producers [3] and in decontamination tests [4]. Activities of a biotechnological interest often are associated with genetically isolated groups of these organisms [5]. For example, many commercial plant growth promoters belong to *B. amyloliquefaciens* ssp. *plantarum* [2]; however, several strains of *B. subtilis* and *B. atrophaeus* also were introduced as successful biopesticides [6].

Surprisingly, despite this long history of study of this group of microorganisms, ecology and preferable habitats of species and subspecies of this group remain unclear. This is a back side of the abundance of these microorganisms. Their spores may tolerate even several minutes of boiling and they easily can be spread by water and wind to any environment. This is why these organisms are the most frequent contaminants in microbiological laboratories and an isolation of *B. subtilis* related organisms from any eco-niche cannot a proof of preferable inhabiting this biotope. Our ignorance of habitat preferences of these biotechnologically important microorganisms limits their application and may be an explanation why *Bacillus* based plant growth promoters and biopesticides sometimes show an unstable efficacy in field conditions and on different plants [8]. Strains for new biopesticides often are selected based on their antagonistic activity against pathogens, but it was shown that the bacteria with extraordinary antagonistic activities not always were able to survive in rhizosphere and to colonize plants [10].

Next generation sequencing and metagenomics allowed studying complex microbial populations without isolation of individual strains. The most common approach is amplification of fragments of 16S rRNA from the whole DNA sample using the universal primers followed by massive parallel sequencing of the amplified fragments. Application of this approach is limited in our case as these bacteria are almost indistinguishable by 16S rRNA [5]. Another metagenomic approach consists in sequencing of randomly generated genomic fragments. As there are plenty of bacterial species in complex habitats, any single marker gene for a species of interest may be absent in the generated metagenomic set of reads. To avoid overlooking of species of interest, the genetic barcode sequence should contain multiple marker genes. The aim of this study was to develop computational approaches for a standardized selecting of marker genes and evaluation of the barcode efficacy. An approach of identification of suitable marker genes by whole genome comparison was proposed in our previous publication [11]. Here, the prepared barcodes were used for identification of closely related *Bacillus* species in publically available metagenomic datasets. Then, an in-house Python script was used for evaluation of performance of different loci in the barcode sequences for further improvement of the barcodes. In Section III.A, the results of identification of barcoded strains of *Bacillus* in plant and rhizosphere associated habitats were presented that followed by the analysis of several gut microbiomes (Section III.B) and estuarine habitats (Section III.C). The approaches of improving of multi-locus barcode sequences to achieve a higher sensitivity and specificity were discussed in the Section III.D. The paper ends up with conclusive remarks.

## II. MATERIALS AND METHODS

### A. *Complete genome sequences and metagenomic datasets used in this study.*

Genetic barcodes were developed for 34 strains of the *B. subtilis / B. amyloliquefaciens* group representing different species and subspecies including commercial and biotechnologically potential strains (Table 1).

TABLE I. BACILLUS GENOME SEQUENCES USED IN THIS STUDY

| Species and strain | NCBI ID or BioProject |
|---|---|
| *B. subtilis* ssp. *subtilis* | |
| 1 168W | NC_000964 |
| 2 6051 | NC_020507 |
| 3 QB928 | NC_018520 |
| 4 UCMB5121 | Newly sequenced |
| 5 UCMB5021 | Newly sequenced |
| 6 BSP1 | NC_019896 |
| 7 BSn5 | NC_014976 |
| 8 BAB1 | NC_020832 |
| 9 XF1 | NC_020244 |
| 10 RO_NN_1 | NC_017195 |
| 11 *B. subtilis* ssp. *natto* BEST195 | NC_017194 |
| 12 *Bacillus* sp JS | PRJNA79217 |
| *B. subtilis* ssp. *spizizenii* | |
| 13 At3 | Newly sequenced |
| 14 UCMB5014 | PRJNA176696 |
| 15 At2 | PRJNA176701 |
| 16 TU-B-10 | NC_016047 |
| 17 W23 | NC_014479 |
| 18 *B. mojavensis* UCMB5075 | Newly sequenced |
| *B. atrophaeus* | |
| 19 1942 | NC_014639 |
| 20 UCMB5137 | PRJNA176685 |
| *B. amyloliquefaciens* ssp. *amyloliquefaciens* | |
| 21 TA208 | NC_017188 |
| 22 XH7 | NC_017191 |
| 23 LL3 | NC_017189 |
| 24 DSM7 | NC_014551 |
| 25 IT-45 | NC_020272 |
| *B. amyloliquefaciens ssp. plantarum* | |
| 26 CAU_B946 | NC_016784 |
| 27 YAU_B9601_Y2 | NC_017061 |
| 28 UCMB5007 | PRJNA176687 |
| 29 UCMB5044 | Newly sequenced |
| 30 UCMB5036 | Newly sequenced |
| 31 UCMB5140 | PRJNA176688 |
| 32 At1 | PRJNA176703 |
| 33 FZB42 | NC_009725 |
| 34 AS43.3 | NC_019842 |

Barcodes were designed as it was explained in our previous publication [11]. Shortly: complete genome comparison revealed a set of 150 rather conserved core genes, which have accumulated amino acid substitution with a relatively higher rate than other genes. It implied an evolutionary positive selection of amino acid substitutions leading to adaptation of organisms to specific habitat conditions.

Metagenome datasets representing different eco-niches were obtained from NCBI and MG-RAST databases [12] (Table 2).

TABLE II. METAGENOMIC SUBSETS

| Biotope | # reads | Types of reads |
|---|---|---|
| Rice phyllosphere | 1,026,982 | Roche 454, ~500 bp. |
| Meadow grassland , USA | 976,268 | Roche 454, ~300 bp. |
| Rain forest soil | 782,404 | Roche 454, ~300 bp. |
| Tropical soil | 5,235,352 | Illumina, 100 bp. |
| Soybean rhizosphere from Amazon soils | 578,060 | Roche 454, ~500 bp |
| Mediterranean oak forest rhizosphere | 561,526 | Roche 454, ~150 bp |
| Mangrove estuarine mud | 481,226 | Roche 454, ~300 bp. |
| Anthropogenic estuarine mud | 526,919 | Roche 454, ~300 bp. |
| Termite gut | 99,776 | Roche 454, ~600 bp. |
| Cow gut | 264,849 | Roche 454, ~100 bp. |
| Human gut | 1,000,000 | Sanger, ~1300 bp. |
| Canine gut | 583,523 | Roche 454, ~500 bp. |

### B. *Application of the barcodes.*

DNA reads of metagenomic datasets were aligned against the barcode sequences by BLASTN. Hits with evalue smaller than 0.00001 and the scores above 75 for short Illumina and Roche 454 reads; above 100 for Roche 454 reads and short contigs around 500 bp; and above 150 for Sanger reads were selected for barcode scoring according to the (1):

$$score = \frac{\sum \frac{b\_score}{r\_length} \times \frac{G - N_g}{G - 1}}{barcode\_length} \qquad (1)$$

where *b_score* – blastn score of a read; *r_length* – length of the read; *G* – number of barcodes in the set, in our case – 34; $N_g$ – number of barcodes with which the read gives reliable blast hits. The logic of this score is that the reads aligned through its whole length contribute more to the score than those aligned partially; and the reads with specific hits against one single barcode sequence contribute more than those with multiple hits against many barcodes. An assumption was that the strain specific barcodes with higher scores would indicate presence of similar organisms in the habitat specific metagenome.

Local blast search and statistics were implemented in Python 2.5 scripts.

### C. *Phylogenomic tree inferring.*

Phylogenomic tree was inferred by the maximum likelihood algorithm based on a super-alignment of amino acid sequences of 2,318 identified core genes common for all tested *Bacillus* genomes.

## III. Results and Discussion

### A. *Analysis of the rhizosphere and phyllosphere associated barcodes*

Results of the blast search of reads of 5 metagenomic datasets associated with rhizosphere, soil and plants are shown in Figure 1.

In Figure 1 and in the following figures of mapping of the metagenomic reads, the barcodes are numbered in the same order as in Table 1.

The profiles of strains were quite similar in the rhizosphere associated biotopes with a general predominance of *B. amyloliquefaciens* over *B. subtilis* and other species except for the strain *B. subtilis* ssp. *spizizenii* At3. The latter one was frequent in all metagenomic datasets and in two of them it was the most abundant strain. The bacillary microflora of rice phyllosphere was quite similar to those of rhizosphere samples, but an interesting observation was that *B. amyloliquefaciens* ssp. *amyloliquefaciens* were more frequent in the phyllosphere sample, while their closest relatives of the subspecies *plantarum* were more common in the rhizosphere samples.

*Bacillus* species profile in the oak forest rhizosphere was quite different with *B. subtilis* ssp. *subtilis* and *B. mojavensis* to be the dominant organisms.

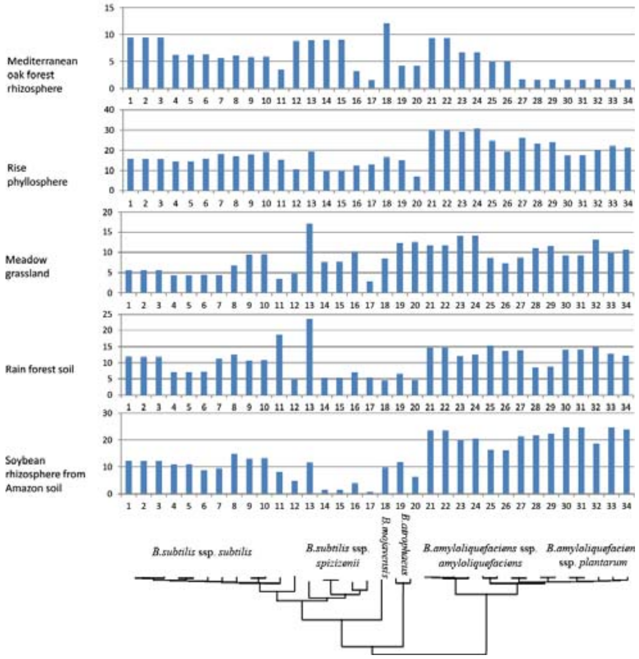

Figure 1. Scores calculated for the barcodes by BLASTN mapping of DNA reads are shown in histograms. Phylogenomic tree of tested strains is shown below.

*B. amyloliquefaciens* ssp. *plantarum* are widely used in many commercial biopesticides and as plant growth promoting agents. The observation of distribution of *Bacillus* species shown in Figure 1 suggests that strains of *B. amyloliquefaciens* ssp. *amyloliquefaciens* may be a better choice for applying biocontrol agents on leaves; and *B. subtilis* based biopesticides may be more effective in forestry.

### B. *Analysis of metagenomes of gut microflora*

It was interesting to investigate whether the microflora of herbivorous organisms would resemble that one of plants and would there be differences in microbiota of herbivorous and carnivorous organisms. In this study, the metagenomes of gut microflora of cow, human, dog and termite were analysed (Figure 2).

In termite gut, *B. amyloliquefaciens* and *B. atrophaeus* were absolutely dominants. These species were abundant also in guts of cow, but *B. subtilis* ssp. *spizizenii* was also frequent in this microbiota, especially the linage At3 that was although frequent in the rhizosphere (Figure 1). *Bacillus* species were more or less equally distributed in human intestines with minority in *B. subtilis* ssp. *spizizenii* and *B. mojavensis*. Contrary, in canine guts *B. mojavensis* and organisms similar to *B. subtilis* ssp. *spizizenii* W23 were the only present *Bacillus*.

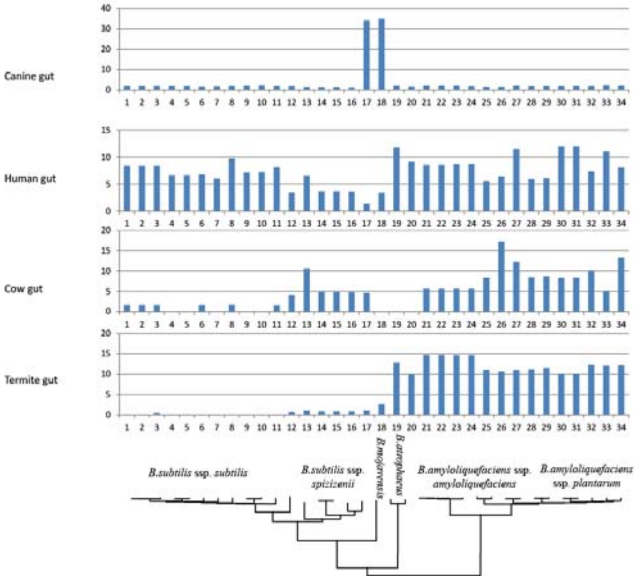

Figure 2. *Bacillus* species profiles in metagenomes of cow, human, canine and termite gut microflora.

Strains of the *B. subtilis* group are used in several medicinal probiotics to prevent gastro-intestinal diseases and dysbacteriosis; however, it has been never reported was there any difference between the strains used for plant protection and those used in probiotics, as they all belonged to the same species [13]. The hypothesis was that although species of *Bacillus* did not belong to human and animals' resident microflora, they constantly had been arriving with the food and might interfere with pathogenic bacteria and the immune system of the higher organism. This research confirmed that *Bacillus* are common in the gut microflora of herbivorous and omnivorous organisms including the human gut microflora. However, a significant difference in species profiles was observed suggesting that human probiotics may not necessary be effective for domestic and farm animals.

### C. Analysis of other metagenomes

To check how specific the profiles of species of the *B. subtilis* group are in the habitats observed above, several other metagenomic datasets were analysed. In Figure 3, the results of analysis of metagenomes of natural and anthropogenic estuarine mud are shown.

In the natural estuarine mud the species of *B. subtilis* / *B. amyloliquefaciens* were poorer represented than it was in the rhizosphere associated habitats. There were more representatives of *Bacillus* in the estuarine affected by industrial pollution with *B. atrophaeus* became the most abundant species comparing to other species of this group.

Several more metagenomic datasets not mentioned in Table 2 were tried in this study: cliff soil, acid mine drainage, rock biofilm, activated sludge and hydrothermal vent. There were no hits against *Bacillus* specific barcodes indicating that these habitats were not occupied by representatives of the *B. subtilis* and *B. amyloliquefaciens* lineages.

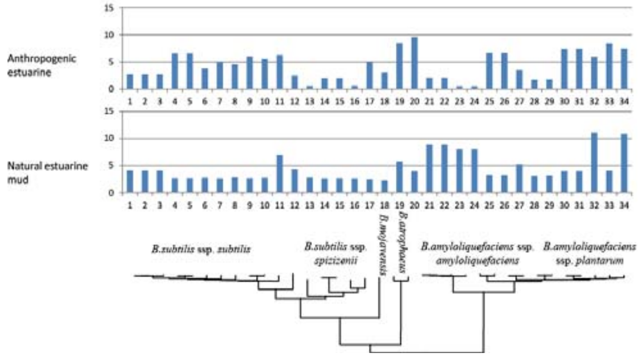

Figure 3. *Bacillus* species profiles in estuarine mud metagenomes.

### D. Further improvement of barcode specificity

In addition to getting information regarding distribution of *B. subtilis* related species in different habitats, the conducted study aided in identification of loci in the barcode sequences, which contributed to species distinguishing and those which created noise of were silent. A developed Python script returned a graphical representation of efficacy of different barcode loci as shown in Figure 4. In this figure red to brown areas were the most effective in separating even closely related strains. Green areas were species and subspecies specific. Blue areas were not specific and created informative noise, and the white areas never have been hit by any read in this study.

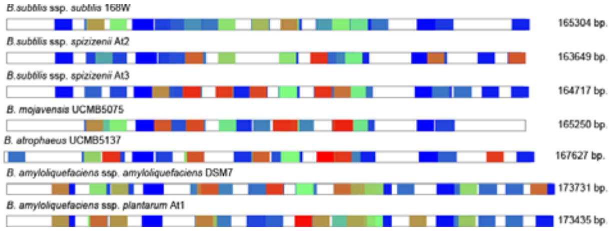
B.subtilis ssp. subtilis 168W
165304 bp.
B.subtilis ssp. spizizenii At2
163649 bp.
B.subtilis ssp. spizizenii At3
164717 bp.
B. mojavensis UCMB5075
165250 bp.
B. atrophaeus UCMB5137
167627 bp.
B. amyloliquefaciens ssp. amyloliquefaciens DSM7
173731 bp.
B. amyloliquefaciens ssp. plantarum At1
173435 bp.

Figure 4. Per locus analysis of discriminative power of several selected barcode sequences.

The barcodes may be improved by removal non-specific loci (in Figure 4 shown in blue) and replacing them with other marker sequences. Than the blastn mapping may be repeated to check whether the specificity of barcodes had been improved.

## IV. Conclusion

The secret of success of a biotechnological culture is not in its species name but in the strain-specific biological activities. Recent advance in sequencing technologies made it affordable to sequence and compare whole genomes of closely related organism in infancy of clonal segregation and speciation. The most popular next generation sequencing techniques are Roche 454, Illumina and Ion Torrent. They produce short DNA reads of several hundred nucleotides depending on the used technology, which are randomly generate from genome sequences in a sample. There is an urgent need in new computational techniques for mining of huge amounts of sequence data generated by the next generation sequencers to identify and highlight marker sequences suitable for barcoding of strains of interests and their biological activities. Multilocus barcoding seems to be a promising approach for a reliable identification of strains of closely related bacteria in environmental samples. Prior to this study the genomes of 34 organisms of the *B. subtilis* / *B. amyloliquefaciens* group were compared and 150 core genes with traces of positive selection were chosen for genetic barcode design [11]. The aim of the current research was to evaluate the prepared barcodes on publically available metagenomic datasets and to develop an approach of estimation of efficacy of barcode sequences per individual loci for further improvement of the selectivity and specificity of the method.

Finding of this research was that the species and subspecies of the *B. subtilis* / *B. amyloliquefaciens* group, and even several individual strains of this group, have had preferences in distribution among different biotopes. Rhizosphere biotopes were populated mostly with *B. amyloliquefaciens* ssp. *plantarum* (Figure 1). Representatives of this taxonomic unit are promising biotechnological strains used as components of plant growth promoting preparations and biopesticides. Interestingly, that the strains of *B. amyloliquefaciens* ssp. *amyloliquefaciens*, which are characterized by a stronger enzymatic activity, were more dominant in rice phyllosphere and in the gut microbiota of termites (Figure 2), where they might contribute to enzymatic digestion of complex hydrocarbons [14]. Unexpectedly, the microflora of the oak forest rhizosphere was quite different from that of the grassland and tropical soils with dominance of *B. subtilis* ssp. *subtilis* strains in the former one. This observation may explain why several biotechnological formulations for plant growth promoting and protection not always were equally efficient in laboratory applications and in different field conditions.

One serious problem with multilocus barcode application is that the barcodes comprising different sets of marker sequences may return incomparable results. A Python script was introduced in this work that may be used for evaluation of efficacy of different barcodes and even individual loci in their sequences. This algorithm allows a critical consideration of results of barcode based identification and makes it possible to determine the loci causing noise or false signals. Sensitivity and specificity of barcodes may be improved by removal of these fragments. An ultimate goal of further research on this project will consist in development of a standardized computer based system for a recurrent design, evaluation and improving of barcode sequences for identification and monitoring of biotechnological and pathogenic microorganisms in nature on the level of subspecies or individual strains.

## ACKNOWLEDGMENT

This project is supported by the South African National Research Foundation (NRF) grant #93134.

## REFERENCES

[1] R. A. Slepecky and H. E. Hemphill, “The genus Bacillus – nonmedical” in The Prokaryotes, vol. 4, M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, Eds. Singapore: Springer, pp. 530-562, 2006.

[2] X. H. Chen, et al., “Comparative analysis of the complete genome sequence of the plant growth–promoting bacterium Bacillus amyloliquefaciens FZB42,” Nat. Biotechnol., vol. 25, pp. 1007-1014, Aug. 2007, doi:10.1038/nbt1325.

[3] G. R. Castro, B. S. Méndez, and F. Siñeriz, “Amylolytic enzymes produced by Bacillus amyloliquefaciens MIR ‐ 41 in batch and continuous culture,” J. Chem. Tech. Biotechnol., vol. 56, pp. 289-294, Apr. 1993, doi: 10.1002/jctb.280560312.

[4] H. S. Luftman and M. A. Regits, “B. atrophaeus and G. stearothermophilus biological indicators for chlorine dioxide gas decontamination,” Applied Biosafety: Journal of the American Biological Safety Association, Vol. 13, pp. 143-157, Mar. 2008.

[5] L. A. Safronova, L. B. Zelena, V. V. Klochko, and O. N. Reva, “Does the applicability of Bacillus strains in probiotics rely upon their taxonomy?” Can J Microbiol., Vol. 58, pp. 212-219, Feb. 2012, doi: 10.1139/w11-113.

[6] J. W. Kloepper, R. Lifshitz, and R. M. Zablotowicz, “Free-living bacterial inocula for enhancing crop productivity,” Trends in Biotechnology, Vol. 7, pp. 39-44, Feb. 1989, doi:10.1016/0167-7799(89)90057-7.

[7] W. Y. Chan, K. Dietel, S. V. Lapa, L. V. Avdeeva, R. Borriss, and O. N. Reva, “Draft genome sequence of Bacillus atrophaeus UCMB-5137, a plant growth-promoting rhizobacterium,” Genome Announc., Vol. 1, pp. e00233-13, Jun. 2013, doi: 10.1128/genomeA.00233-13.

[8] T. Vasiliki, J. O'Sullivan, A. C. Cassells, D. Voyiatzis, and G. Paroussi, "Comparison of AMF and PGPR inoculants for the suppression of Verticillium wilt of strawberry (Fragaria ananassa cv. Selva)," Applied Soil Ecology, Vol. 32, pp. 316-324, Jul. 2006, doi:10.1016/j.apsoil.2005.07.008.

[9] B. Fan, R. Borriss, W. Bleiss, and X. Wu, “Gram-positive rhizobacterium Bacillus amyloliquefaciens FZB42 colonizes three types of plants in different patterns,” J. Microbiol., Vol. 50, pp. 38-44, Feb. 2012, doi: 10.1007/s12275-012-1439-4.

[10] O. N. Reva, C. Dixelius, J. Meijer, and F. G. Priest, “Taxonomic characterization and plant colonizing abilities of some bacteria related to Bacillus amyloliquefaciens and Bacillus subtilis,” FEMS Microbiol. Ecol., Vol. 48, pp. 249-259, May 2004, doi: 10.1016/j.femsec.2004.02.003.

[11] O. N. Reva, et al, “Genetic barcoding of bacteria and its microbiology and biotechnology applications,” In Bioinformatics and Data Analysis in Microbiology, O. Tastan Bishop Ed., Caister Academic Press, pp. 229-244, Mar. 2014, ISBN 978 190823 0737.

[12] F. Meyer, et al., “The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes,” BMC Bioinformatics, Vol. 9, p. 386, Sep. 2008, doi: 10.1186/1471-2105-9-386.

[13] I. B. Sorokulova, et al, “The safety of two Bacillus probiotic strains for human use,” Dig. Dis. Sci.,Vol. 53, pp. 954-963, Apr. 2008, doi: 10.1007/s10620-007-9959-1.

[14] S. Um, A. Fraimout, P. Sapountzis, D. C. Oh, and M. Poulsen, “The fungus-growing termite Macrotermes natalensis harbors bacillaene-producing Bacillus sp. that inhibit potentially antagonistic fungi,” Sci. Rep., Vol. 3, p. 3250, Nov. 2013, doi: 10.1038/srep03250.