
http://www.diva-portal.org

This is the published version of a paper published in *Genomics*.

Citation for the original published paper (version of record):

Rahman, A., Nahar, N., Nawani, N N., Jass, J., Ghosh, S. et al. (2015)
Comparative genome analysis of Lysinibacillus B1-CDA, a bacterium that accumulates arsenics.
Genomics, (6): 384-392
http://dx.doi.org/10.1016/j.ygeno.2015.09.006

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-11575

Comparative genome analysis of *Lysinibacillus* B1-CDA, a bacterium that accumulates arsenics

Aminur Rahman[1,3], Noor Nahar[1], Neelu N. Nawani[2], Jana Jass[3], Sibdas Ghosh[4], Björn Olsson[1]* and Abul Mandal[1]*

[1]Systems Biology Research Center, School of Bioscience, University of Skövde, P.O. Box 408, SE-541 28 Skövde, Sweden

[2]Dr. D. Y. Patil Biotechnology and Bioinformatics Institute, Dr. D. Y. Patil Vidyapeeth, Tathawade, Pune-411033, India

[3]The Life Science Center, School of Science and Technology, Örebro University, SE-701 82 Örebro, Sweden

[4]School of Arts and Science, Iona College, New Rochelle, NY 10801, USA

* Authors have the equal contribution

Address correspondence to Abul Mandal, System Biology Research Center

School of Bioscience, University of Skövde, P. O. Box 408, SE-541 28 Skövde, Sweden

Phones +46-500448608 (direct), +46-739-876839 (mobile)

E-mail: abul.mandal@his.se

Abstract

Previously, we reported an arsenic resistant bacterium Lysinibacillus sphaericus B1-CDA, isolated from an arsenic contaminated lands. Here, we have investigated its genetic composition and evolutionary history by using massively parallel sequencing and comparative analysis with other known Lysinibacillus genomes. Assembly of the sequencing reads revealed a genome of ~4.5 Mb in size encompassing ~80% of the chromosomal DNA. We found that the set of ordered contigs contains abundant regions of similarity with other Lysinibacillus genomes and clearly identifiable genome rearrangements. Furthermore, all genes of B1-CDA that were predicted be involved in its resistance to arsenic and/or other heavy metals were annotated. The presence of arsenic responsive genes was verified by PCR in vivo conditions. The findings of this study highlight the significance of this bacterium in removing arsenics and other toxic metals from the contaminated sources. The genetic mechanisms of the isolate could be used to cope with arsenic toxicity.

Keywords: Toxic metals, Bioremediation, Lysinibacillus sphaericus B1-CDA, Genome sequencing, de novo assembly, Gene prediction.

Introduction

Worldwide various anthropogenic activities such as mining, chemical industries, use of arsenic-based pesticides, and natural occurrences continue to cause major environmental and health problems by releasing heavy metals into the soil and water alike [1], and exposing millions of people directly or indirectly to toxic metals including arsenic (As). Long-term exposure to As leads to skin diseases, such as hyper- and hypo-pigmentation, hyperkeratosis and melanosis, as well as gangrene, skin cancer, lung cancer and bladder cancer [2]. Poisoning occurs from drinking of contaminated water and/or consuming crops cultivated by irrigation with As-contaminated groundwater [3-5]. Studies by the Food and Agricultural Organization of the United Nations (FAO) indicate that arsenic is accumulated in different parts of the cultivated crops, such as the grain and straw of rice, a major staple food [6]. It is therefore important that we develop efficient, yet affordable, technologies to clean arsenic from soil and water.

Remediation of toxic metal using microbes has been shown to be more proficient than physical and chemical methods [7]. In fact, bacteria have developed several metabolic processes and strategies to transform As, including respiratory arsenate reduction, cytoplasmic arsenate reduction and arsenite methylation [8]. Furthermore, certain bacteria have evolved the necessary genetic components that confer resistance mechanisms, allowing them to survive and grow in environments containing levels of As that would be toxic to most other organisms. The high-level resistance to As in bacteria is conferred by the arsenical resistance (ars) operon comprising either three (*arsRBC*) [9] or five (*arsRDABC*) genes arranged in a single transcriptional unit located on plasmids [10] or chromosomes [11]. ArsB, an integral membrane protein that pumps arsenite out of the cell, is often associated with an ATPase subunit, arsA [12]. The *arsC* gene encodes the enzyme for arsenate reductase, which is responsible for the biotransformation of arsenate [As(V)] to arsenite [As(III)] prior to

efflux. ArsR is a trans-acting repressor involved in the basal regulation of the ars operon, while arsD is a second repressor controlling the upper levels of expression of ars genes [13]. Several researchers have studied As transformation mechanisms using genetic markers such as *arsB* and *arsC* genes in the ars operon for arsenic resistance [12, 14], the arrA gene for dissimilatory As(V) respiration (DAsR) [15-17], and the aoxB gene for As(III) oxidation [18-19]. Moreover, some studies detected that in spite of clear evidence of the As-transforming activity by microorganisms, no amplicon for arsenite oxidase (aoxB) or As(V) respiratory reductase (arrA) was attained using the reported polymerase chain reaction (PCR) primers and protocols [17, 20-21]. Here we report a bacterial strain, *Lysinibacillus sphaericus* B1-CDA as potential candidate for heavy metal bioremediation. This bacterial strain was isolated from cultivated land in the Chuadanga district of Bangladesh, where soil, sediment, and ground water have been contaminated with arsenics for many years.

In this study, we provide *in vivo* findings of potential arsenic responsive genes in B1-CDA and summarize a set of phenotype features for *Lysinibacillus sphaericus* B1-CDA, together with a description and annotation of its genome sequence. To improve our understanding of genes involved in metal binding activity and reduction of metal by the B1-CDA strain, we performed massively parallel genome sequencing to investigate the metal responsive genes, predicted by RAST and/or Blast2GO. Employing comparative analyses with other available *Lysinibacillus* genome sequences, we investigated genetic composition and evolutionary history of strain B1-CDA and characterized the genetic differences among the various lineages to understand the evolutionary processes involved in shaping the genomes of these bacteria.

Methods

Strain isolation

The soil samples were collected from cultivated land in the Chuadanga district of Bangladesh, a highly arsenic-contaminated region located in the south-west region of this country. The soil was collected from the surface at 0-15 cm in depth, retained in plastic bags and kept at 4°C until further analysis. Isolation of bacteria from the collected soil, the characterization of the soil samples and the content of metal ions has been described previously [22]. Previously, we have reported that the strain *Lysinibacillus sphaericus* B1-CDA is highly resistant to arsenic and it accumulates arsenic inside the cells [22].

Genomic DNA extraction and electrophoresis

Genomic DNA was extracted from the isolate, B1-CDA using Master pure™ Gram positive DNA purification kit (Epicentre, USA) with a minor modification. Bacteria were cultured in Luria Bertani (LB) medium and pellets were collected from 1.0 ml of bacterial cultures by centrifugation at 8000 rpm for 10 min. The pellets were resuspended in TE buffer (10 mM Tris- HCl, 1 mM EDTA [pH 8.0]) containing lysozyme (50 mg/ml) and RNase (50 mg/ml) and incubated at 37°C for 2 h. The suspension was then mixed with proteinase K (50 μg/μl) and the cells were lysed following incubation at 65-70°C for 15 minutes. Protein precipitation solution was added to remove the proteins, and the DNA was precipitated with isopropanol and washed with 70% ethanol. The DNA was resuspended in TE buffer and quantified by Nanodrop® ND-1000 Spectrophotometer (Saveen Werner, USA).

Agarose gel electrophoresis was performed based on the established protocol [23]. The gel was prepared with 0.8% agarose in 1X TAE buffer. The *Saccharomyces cerevisiae* chromosomal DNA (Bio-Rad, USA) was used as the size standard to estimate the molecular weight of the *L. sphaericus* B1-CDA chromosome.

Primer design and PCR amplification of arsenic-responsive genes

Primers were designed based on the multiple sequence alignment of target genes in a variety of arsenic resistant bacilli. Sequences of the *arsR* gene from 25 *Bacillus species* were randomly selected from GenBank. The multiple sequences of the *arsR* genes were aligned by ClustalW [24] in order to find the longest region of conserved homology. Seven bacterial strains exhibiting highest homology to the conserved region with each other were chosen for designing PCR primers. Two degenerate primer pairs were used to amplify the As marker genes *arsB* [12] and *arsC* [14]. Primers for the *acr3* gene were designed by using the Primer3Plus web tool [25]. PCR amplification of arsenic-related marker genes was performed by using bacterial genomic DNA as a template. All PCR reaction mixtures contained approximately 50 ng DNA template, 1X PCR buffer, 0.2 mM of each deoxyribonucleoside triphosphate, 0.5 mM of each primer and 1 U Taq DNA polymerase in 50 µl volume. Amplifications were performed in a piko thermal cycler (Finzymes). Cycling conditions for all PCRs consisted of 5 min of denaturation at 95°C followed by 34 cycles of 1 min of denaturation at 95°C, 45 s of annealing at 57.7°C and primer extension at 72°C for 1 min of each Kb product size. This was followed by a final extension reaction at 72°C for 10-15 min. PCR products were purified with a QIAquick gel extraction kit (Qiagen, Cat No 28706).

Genome sequencing

Sequencing of the genomic DNA of *Lysinibacillus sphaericus* B1-CDA was performed by the Otogenetics Corporation (GA, USA). Purified 0.5-1 µg of genomic DNA sample was sheared into smaller fragments with a Covaris E210 ultrasonicator. Genomic DNA library was constructed by using the NEB library preparation kit (New England Biolabs) for the Illumina sequencer with a single sequencing index and sequencing was performed with the help of HiSeq2500 PE100 read format. Properly paired reads (≥30bp) were extracted from the

corrected read pool and the remaining singleton reads were combined as single-end reads. Both corrected paired-end and single-end reads were used in the subsequent *de novo* assembly.

Genome assembly

The genome assembly started with Illumina 100-bp paired-end reads of genomic DNA with insert length 300 bp. The read quality was checked by using FastQC, version 1.10.1 [26]. The raw reads were quality trimmed and corrected using Quake version 0.3.4 [27]. Properly paired reads of length ≥30 bp were selected from the pool of corrected reads and the remaining singleton reads were considered as single-end reads. Both the paired-end and single-end corrected reads were then used in k-mer-based *de novo* assembly employing SOAPDenovo, version 2.04 [28]. The set of scaffolds with largest N50 was identified by evaluating k-mers in the range 29-99. The optimal scaffold sequences were further subjected to gap closing by utilizing the corrected paired-end reads. The resulting scaffolds of length ≥300 bp were chosen as the final assembly. The scaffolds were ordered by finding the location of the best Blastn hit for each scaffold on the reference genome *Lysinabacillus sphaericus* C3-41. A total of 31 scaffolds were used for prediction of the genome size and it was performed by following the Mauve Contigs Mover (http://darlinglab.org/mauve/user-guide/reordering.html).

Phylogenetic inference of Lysinibacillus sp.

In this study, a phylogenetic tree was inferred from the 16S rRNA genes of B1-CDA and other related bacteria [29] by using the Neighbor-Joining method [30] presented in the MEGA6 software [31]. The analysis involved nucleotide sequences from 27 bacteria in the

Bacillaceae family. The evolutionary distances were computed using the Kimura 2-parameter method [32] in the units of the number of base substitutions per site, including all codon positions (1st, 2nd, 3rd and noncoding). Positions with < 95% site coverage were eliminated, thereby allowing fewer than 5% alignment gaps, missing data, and ambiguous bases at any position. There were a total of 1227 positions in the final dataset.

Comparative analysis with other Lysinibacillus *genomes*

To study genome rearrangements in *Lysinibacillus* and related bacteria, the progressive MAUVE algorithm in the MAUVE genome alignment software version 2.3.1 [33] was employed. The main purpose of using this method was to compare the possible rearrangements that may occur in *L. sphaericus* B1-CDA [GenBank accession number PRJEB7750, http://www.ebi.ac.uk/ena/data/view/PRJEB7750], *L. sphaericus* OT4b.31 [accession number AQPX00000000] and L. *sphaericus* C3-41 [accession number CP000817.1]. A nucleotide-based dot plot analysis was performed with the Gepard software [34] to compare the 4.09 Mbp chromosomal scaffolds of *L. sphaericus* B1-CDA with the 4.6 Mbp chromosome of *L. sphaericus* C3-41 and the genome rearrangements were studied.

Gene prediction and annotation of metal resistant genes

Circular plot of ordered contigs of B1-CDA was generated with DNAPlotter [35] to predict the graphical map of the genome. The assembled genome sequence was annotated with Rapid Annotations using Subsystems Technology, RAST [36]. The RAST analysis pipeline uses the tRNAscan-SE to predict tRNA genes [37] and the GLIMMER algorithm to predict protein-coding genes [38]. In addition, it uses an internal script for identification of rRNA genes [36]. It then infers putative function(s) of the protein coding genes based on homology to already known protein families in phylogenetic neighbor species. Finally, RAST identifies

subsystems represented in the genome, and uses this information to reconstruct the metabolic networks. The GeneMark [39] and the FGenesB [40] algorithms were applied for verification of the RAST results obtained in prediction of protein coding genes. Prediction of rRNA genes was also done through the RNAmmer prediction server version 1.2. [41]. Annotation of all genes that were predicted to be metal responsive was manually curated, with a particular focus on genes responsive to As. Functional annotation analysis was also carried out by the Blast2GO pipeline [42] using all translated protein coding sequences resulting from the GeneMark. In Blast2GO the BlastX option was chosen to find the closest homologs in the non-redundant protein databases (nr), followed by employing Gene Ontology (GO) annotation terms [43] to each gene based on the annotation of its closest homologs. An InterPro scan [44] was then performed through the Blast2GO interface and the InterPro IDs merged with the Blast-derived GO-annotation for obtaining integrated annotation results. The GO annotation of all putative metal responsive genes was manually curated.

Results and discussion

Detection of arsenic marker genes

Several studies have used genetic markers to study As transformation mechanisms [9, 12, 14-19]. In this study, we present a genetic mechanism for As resistance and As transformation in the bacterial isolates. This mechanism was examined via PCR amplification of As responsive genes. The strain B1-CDA was found to harbor *acr3, arsR, arsB* and *arsC* arsenic marker genes (Supplementary Figure 1). The *arsC* gene consists enzyme for arsenate reductase, which is responsible for the biotransformation of arsenate [As(V)] to arsenite [As(III)] prior to efflux. ArsB, an integral membrane protein that pumps arsenite out of the cell, is often associated with an ATPase subunit, arsA [12]. It is hypothesized that the *arsB/acr3* genes are the primary determinants in arsenite resistance [12]. These genetic mechanisms of the isolate

could be used to cope with arsenic toxicity. Such mechanisms could comprise arsenite methylation that results in volatile products which having very less toxicity that escape from the cells [45].

Sequencing and de novo genome assembly

A total of 11,105,899 pairs of reads were generated by Illumina deep sequencing. Analysis of the raw reads with FastQC showed that the average per base Phred score was ≥32 for all positions and that the mean per sequence Phred score was 38. The overall GC content was 38%. After removal of the TruSeq adaptor sequence (which was found in 13,435 reads, 0.12%) and error correction and trimming done by using the Quake software, 10,940,654 read pairs (98.5%) and 145,888 single end sequences remained for further analysis. Trimming was performed as the trimming of sequences is an important step for improving mapping efficiency. SOAPDenovo was utilized to perform *de novo* assembly optimization with the error corrected reads. The set of scaffold sequences with maximal N50 (507,225 bp) was produced at k-mer 91. The corresponding scaffold sequences were subjected to gap closure using the corrected paired-end reads and the resulting set of scaffolds (≥300 bp) was defined as the final assembly. The final assembly consists of 31 scaffolds, with lengths ranging from 314 bp to 1,145,744 bp, resulting a total length of 4,509,276 bp. It contained only 25 bp of unknown nucleotides, i.e. the error rate was less than 1 in 1,000,000. The summary of the genome with nucleotide content and gene count levels are described in Table 1.

Phylogenetic analysis

Peña-Montenegro and Dussán [29] evaluated phylogenetic tree with native *Bacillaceae* isolates along with *Lysinibacillus sphaericus* OT4b.31, a heavy metal tolerant bacterium. Strains of *L. sphaericus* can be divided into seven DNA similarity subgroups (I–VII), with a

clear separation between the groups I-V and groups VI-VII [29, 46]. Phylogenetic analysis based on 16S ribosomal RNA gene sequences did not place our strain *L. sphaericus* B1-CDA into any of the existing DNA similarity groups (Figure 1). The placement of *L. sphaericus* B1-CDA in the phylogeny does, however, indicate higher similarity to groups I-V than to groups VI-VII. In agreement with earlier studies [29, 46] *Bacillus silvestris* was also placed between groups I-V and groups VI-VII. This indicated that B1-CDA could also belong to *B. silvestris* but based on the branch lengths it was confirmed that B1-CDA belongs to the species *L. sphaericus* rather than *B. silvestris*.

Comparative genome analysis

Using the Gepard dot plot software [34] and progressiveMauve from the Mauve software [33], we compared the chromosomal assembly of B1-CDA with that of the *L. sphaericus* C3-41 and *L. sphaericus OT4B.31*. The alignment of B1-CDA, *L. sphaericus* C3-41 and *L. sphaericus* OT4B.31 showing the same chromosomal rearrangement but B1-CDA consists mostly the inversions with *L. sphaericus* C3-41 and *L. sphaericus* OT4B.31 (Figure 2A). The chromosomal alignments of *L. sphaericus* C3-41 and *L. sphaericus* OT4B.31 are about to identical (Figure 2A). The dot plot was performed with B1-CDA and *L. sphaericus* C3-41 since *L. sphaericus* OT4B.31 and *L. sphaericus* C3-41 were similar. The dot plot shows that the genome rearrangements consist mostly of inversions (Figure 2B). There are large segments of high similarity when most parts of the two chromosomes are mapped onto each other. However, a region comprising around 3 Mbp in the C3-41 chromosome and the contigs 15 to 19 in the B1-CDA chromosomal scaffold were somewhat scattered in the dot plot, revealing lower similarity levels and different syntenial relationships to the reference sequence.

Gene predictions

Prediction of tRNA-, rRNA- and protein coding genes were performed through the RAST server. The graphical map of the genome and the locations of all predicted genes are shown in the circular genome plot in Figure 3. The search by tRNAscan-SE (which is the first step in the RAST pipeline) located 77 tRNA genes. A confirmatory scan with the algorithm ARAGORN [47] predicted the identical number, and all predictions overlapped in location, although with slight variation regarding the start or end point for a few genes. The predictions included tRNAs for 19 amino acids, ranging in number from one gene for the cysteine tRNA to six genes for the arginine and glutamic acid tRNAs. The only tRNA gene missing in these predictions was for the amino acid serine. However, three pseudo-tRNA genes were predicted, and two of these contained anticodons for serine. In an ARAGORN scan the total number of predicted tRNA genes in *L. sphaericus* B1-CDA was 77, which was similar to that (83) predicted in the genome of *L. sphaericus* C3-41.

The rRNA prediction in RAST resulted in 11 rRNA genes, including seven 5S, one 16S and three 23S genes. Whereas, another related strain *Lysinibacillus sphaericus* C3-41 containing 31 rRNA genes, including eleven 5S, ten 16S and ten 23S genes. Due to the surprisingly low number of 16S and 23S genes, RNAmmer [41] scans were performed on the genomes of *L. sphaericus* B1-CDA as well as two unrelated bacteria (*Enterobacter cloacae* and *Salmonella bongori*). The results of these scans were compared to each other prior to making any conclusion. These results confirmed that the *L. sphaericus* B1-CDA genome seems to contain approximately the same number of 5S rRNA genes as the other bacteria, but substantially fewer 16S and 23S rRNA genes. Previously, Pei et al. [48] have shown 143 bacterial species contain only a single 16S rRNA whereas Pei et al. [49] have shown 184 genomes had a median of 4.57 23S rRNA genes/ genome (range 1 to 15). Therefore, the lower number of 16S- and 23S rRNA is not an uncommon feature of bacterial genome.

For prediction of the protein coding genes, RAST uses the GLIMMER algorithm [38]. A total of 4513 protein coding genes were predicted using GLIMMER algorithm), of which 2671 could be annotated by RAST's automated homology analysis procedure and assigned to functional categories (Figure 4). For confirmation of the number of protein coding genes, the GeneMark [39] and FGenesB [40] algorithms were also applied, yielding 4562 and 4323 genes, respectively. We observed that *L. sphaericus* B1-CDA contains many specific metal resistant genes, such as arsenic, nickel, cobalt, iron, manganese, chromium, cadmium, lead and zinc (Table 2). Further, the functional annotation carried out by the Blast2GO pipeline also indicates that B1-CDA contains many genes which are responsive to specific metal ions like arsenic, cobalt, copper, iron, nickel, potassium, manganese and zinc. (Table 2). The annotations by RAST and Blast2GO remind in agreement. Prediction by RAST and Blast2GO (Supplementary Table 1) revealed that the B1-CDA genome contains additionally a total of 123 proteins which are involved in binding and transport of metal ions. This prediction also indicated that B1-CDA contains many other proteins (approximately 30) that catalyze binding and transport of the metal ions such as metalloendopeptidase, metalloexopeptidase, metallopeptidase, metallocarboxypeptidase and metallochaperone (Supplementary Table 2). Overall statistics of the Blast hits in the Blast2GO annotation process confirmed assignment of the new bacterium as *L. sphaericus*, since this was the most frequent species in the overall list of protein homologs (Figure 5). The functional assignment of genes into subsystem categories by RAST was compared between B1-CDA and C3-41 (Figure 5). The large categories of "housekeeping" genes, such as those coding for amino acids, carbohydrates, RNA metabolism, generally contained very similar numbers of genes in the two genomes.

By using InterPro [44] the arsenic responsive genes of B1-CDA genome were compared with the genomes of validly named and sequenced species of Lysinobacter as well as with other closely related arsenic tolerant or resistant species. These results are presented in Table 3. All

seven analyzed genes of B1-CDA showed very high similarity with the genes of other Lysinibacillus species. The minimum identity level (97%) was observed in the *arsC* gene (arsenic reductase) of *Lysinibacillus fusiformis,* whereas the highest similarity (100%) in the *arsC* gene (arsenate reductase regulatory protein Spx) of *Lysinibacillus sphaericus* C3-41.

The origin of replication was estimated to be located in the region between 4.3 and 4.4 Mbp, based on homology search in the DoriC database ver 5.0 [50]. There were significant hits to both *L. sphaericus* oriC regions on scaffold 7. The first covers 772 of the 972 nucleotides of ORI92310378, with 94% identities (and 105 of the remaining nucleotides also match with 94% id 84 bp further downstream), while the second covers 167 of the 170 nucleotides of ORI92310377, again with 94% identities. The corresponding region in the full sequence of the ordered contigs is 4,303,598 – 4,302,608 bp. In the prediction and annotation of genes, RAST predicted the chromosomal replication initiator protein DnaA at location 4,303,961 – 4,302,612. Several other replication-related genes were predicted in the near vicinity, such as DNA gyrase subunit A (at bp 4,297,795 – 4 295 330) and subunit B (at bp 4,299,7554 – 4,297,821), as well as the DNA recombination and repair protein RecF (at bp 4,300,964 – 4,299,849).

Conclusions

The native strain *Lysinibacillus sphaericus* B1-CDA, isolated from a cultivated land, was characterized and found to be a metal including arsenic resistant bacterium. A comparison of the genomic sequences of strains B1-CDA with *L. sphaericus* C3-41 and *L. sphaericus* OT4B.31 demonstrated the presence of only a few similar regions with syntenial rearrangements. By using RAST and Blast2GO analyses we have found genes responsive to several metals such as arsenic, nickel, cadmium, iron, manganese, chromium, cadmium, lead,

cobalt, zinc, silver and mercury. Therefore, our findings in this study may be useful in bioremediation of toxic metals like arsenic, nickel, cadmium, iron, manganese, chromium, cadmium, lead, cobalt, zinc, and silver mercury in polluted environments. In conclusion, our study demonstrates that it is possible to speed up molecular biology research by using bioinformatics tools.

Acknowledgements

This research has been funded primarily by the Swedish International Development Cooperation Agency (SIDA, grant number: AKT-2010-018) and partly by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS, grant number: 229-2007-217). We also acknowledge Nilsson-Ehle (The Royal Physiographic Society in Lund) foundation in Sweden for a mini grant.

References

1. Chen Y, Parvez F, Gamble M, Islam T, Ahmed A, Argos M, Graziano JH, Ahsan H. Toxicol Appl Pharmacol. 2009; 239:184–192.

2. Zhao FJ, McGrath SP, Meharg AA. Arsenic as a food chain contaminant: mechanisms of plant uptake and metabolism and mitigation strategies.Annu Rev. Plant Biol. 2010; 61:535-559. doi: 10.1146/annurev-arplant-042809-112152.

3. Bundschuh J, Nath B, Bhattacharya P, Liu CW, Armienta MA, Moreno Lopez MV, Lopez DL, Jean JS, Cornejo L, Lauer Macedo LF, Filho AT. Arsenic in the human food chain: the Latin American perspective. Sci. Total Environ. 2012; 429:92−106.

4. Halder D, Bhowmick S, Biswas A, Mandal U, Nriagu J, Guha Mazumder DN, Chatterjee D, Bhattacharya P. Consumption of brown rice: A potential pathway for arsenic exposure in rural Bengal. Environmental Science & Technol. 2012; 46:4142 –4148.

5. Neidhardt H, Norra S, Tang X, Guo H, Stuben D. Impact of irrigation with high arsenic burden groundwater on the soil-plant system: result from a case study in the Inner Mongolia, China. Environ Poll. 2012; 163:8-13.

6. Abedin J, Feldman J. Meharg A. Uptake kinetics of arsenic species in rice plants. Plant Physiology. 2002; 128:1120-1128.

7. Valls M, de Lorenzo V. Exploiting the genetic and biochemical capacities of bacteria for the remediation of heavy metal pollution. FEMS Microbiol Rev. 2002; 26(4):327-338.

8. Simeonova DD, Micheva K, Muller DA, Lagarde F, Lett MC, Groudeva VI, Lievremont D. Arsenite oxidation in batch reactors with alginate-immobilized ULPAs1 strain. Biotechnol Bioeng. 2005; 91(4):441-446.

9. Liao VHC, Chu YJ, Su YC, Hsiao SY, Wei CC, Liu CW, Liao CM, Shen WC, Chang FJ. Arsenite-oxidizing and arsenate-reducing bacteria associated with arsenic-rich groundwater in Taiwan. J Contam Hydrol 2011; 123:20-29. doi:10.1016/j.jconhyd.2010.12.003.

10. Owolabi JB, Rosen BP. Differential mRNA stability controls relative gene expression within the plasmid-encoded arsenical resistance operon. J Bacteriol. 1990; 172:2367–2371.

11. Diorio C, Cai J, Marmor J, Shinder R, DuBow MS. An Escherichia coli chromosomal ars operon homolog is functional in arsenic detoxification and is conserved in gram-negative bacteria. J Bacteriol. 1995; 177:2050–2056.

12. Achour AR, Bauda P, Billard P. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. Res Microbiol. 2007; 158:128–137.

13. Silver S, Phung LT. Genes and enzymes involved in bacterial oxidation and reduction of inorganic arsenic. Appl Environ Microbiol. 2005; 71:599-608.

14. Sun Y, Polishchuk EA, Radoja U, Cullen WR. Identification and quantification of arsC genes in environmental samples by using real-time PCR. J Microbiol Meth. 2004; 58:335–339.

15. Malasarn D, Saltikov CW, Campbell KM, Santini JM, Hering JG, Newman DK. arrA is a reliable marker for As(V) respiration. Science. 2004; 306:455.

16. Kulp TR, Han S, Saltikov CW, Lanoil BD, Zargar K, Oremland RS. Effects of imposed salinity gradients on dissimilatory arsenate reduction, sulfate reduction, and other microbial processes in sediments from two California soda lakes. Appl Environ Microbiol. 2007; 73:5130–5137.

17. Song B, Chyun E, Jaffé PR, Ward BB. Molecular methods to detect and monitor dissimilatory arsenate-respiring bacteria (DARB) in sediments. FEMS Microbiol Ecol. 2009; 68:108–117.

18. Rhine ED, Ní Chadhain SM, Zylstra GJ, Young LY. The arsenite oxidase genes (aroAB) in novel chemoautotrophic arsenite oxidizers. Biochem Biophys Res Commun. 2007; 354:662–667.

19. Hamamura N, Macur RE, Korf S, Ackerman G, Taylor WP, Kozubal M, Reysenbach AL, Inskeep WP. Linking microbial oxidation of arsenic with detection and phylogenetic analysis of arsenite oxidase genes in diverse geothermal environments. Environ. Microbiol. 2009; 11:421–431.

20. Kulp TR, Hoeft SE, Asao M, Madigan MT, Hollibaugh JT, Fisher JC, Stolz JF, Culbertson CW, Miller LG, Oremland RS. Arsenic(III) fuels anoxygenic photosynthesis in hot spring biofilms from Mono Lake, California. Science. 2008; 321:967–970.

21. Handley KM, Héry M, Lloyd JR. Redox cycling of arsenic by the hydrothermal marine bacterium Marinobacter santoriniensis. Environ Microbiol. 2009; 11:1601–1611.

22. Rahman A, Nahar N, Nawani NN, Jass J, Desale P, Kapadnis BP, HossainK, Saha AK, Ghosh S, Olsson B, Mandal A. Isolation of a *Lysinibacillus* strain B1-CDA showing potentials for arsenic bioremediation. J Environ Sci and Health, Part A. 2014; 49:1349–1360.

23. Reece RJ. Analysis of Genes and Genomes. John Willey and Sons Ltd. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England. 2004

24. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947-2948.

25. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Jack AM. Leunissen: Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007; 35:W71-W74.

26. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

27. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 2010;11(11):R116. Epub 2010 Nov 29.

28. Li R. Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20(2):265-272.

29. Peña-Montenegro TD, Dussán J. Genome sequence and description of the heavy metal tolerant bacterium *Lysinibacillus sphaericus* strain OT4b.31. Standards in Genomic Sciences. 2013; 9:42-56. DOI:10.4056/sigs.4227894

30. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol & Evol. 1987; 4:406-425.

31. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol & Evol. 2013; 30:2725-2729.

32. Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16:111-120.

33. Darling AE, Mau B, Perna NT. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE. 2010; 5:e11147

34. Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating dot plots on genome scale. Bioinformatics. 2007; 23(8):1026-1028.

35. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics (Oxford, England) 2009; 25(1):119-120.

36. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST Server: Rapid Annotations using Subsystems Technology BMC Genomics. 2008; 9:75. doi:10.1186/1471-2164-9-75.

37. 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.

38. Salzberg SL; Delcher A L; Kasif S; White O. Microbial gene identification using interpolated Markov models. Nucleic acids Res. 1998; 26(2):544–548

39. Borodovsky M, McIninch J. GeneMark: parallel gene recognition for both DNA strands. Comput Chem. 1993; 17:123-133.

40. Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. Genome Res. 2000; 10:516–522.

41. Lagesen K, Hallin P, Rødland EA, Stærfeldt 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

42. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008; 36:3420-3435.

43. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Tarver LI, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000; 25(1):25–29.

44. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–848.

45. Qin J, Rosen BP, Zhang Y, Wang G, Franke S, Rensing C. Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:2075-2080.

46. Nakamura LK. Phylogeny of *Bacillus sphaericus*-like organisms. Int J Systematic and Evol Microbiol. 2000; 50:1715–1722.

47. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research. 2007; 32(1):11-16.

48. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, DeSantis T, Brodie E, Nelson K, Pei Z. Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes. Applied & Environmental Microbiology. 2010; 76(12): 3886–3897.

49. Pei A, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM, Pei Z. Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. PLoS ONE. 2009; 4(5):e5437. doi:10.1371/journal.pone.0005437.

50. Gao F, Luo H, Zhang CT. DoriC 5.0: an updated database of *oriC* regions in both bacterial and archaeal genomes. Nucleic Acids Res. 2013; 41:90-93.

Figure legend

Figure 1. Phylogenetic tree depicting the position of *Lysinibacillus sphaericus* B1-CDA relative to the available type strains and other non-assigned species within the family *Bacillae*. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) is shown next to the branches.

Figure 2. (A) Nucleotide-based alignment of a 4.5 Mbp chromosomal assembly of *L. sphaericus B1-CDA* (upper) to a 4.6 Mbp chromosome of *L. sphaericus* C3-41 (middle) and 4.09 Mbp chromosome of *L. sphaericus* OT4b.31 (lower). A total of 27 homologous blocks are shown as identically colored regions and linked across the sequences. Regions that are inverted relative to *L. sphaericus* B1-CDA are shifted to the right of center axis of the sequence. (B) Dot plot of nucleotide sequences of *L. sphaericus B1-CDA* (x-axis) and *L. sphaericus* C3-41 (y-axis). Aligned segments are represented as dots, with regions of conservation appearing as lines.

Figure 3. Circular plot of ordered contigs, generated with DNAPlotter. Tracks indicate (from outside inwards) predicted protein coding genes on forward strand (light blue) and reverse strand (dark blue), metalloproetin genes (red) listed in Table 2, tRNA and rRNA genes (both green), origin of replication (black), GC content and GC skew.

Figure 4. RAST analysis of genes connected to subsystems and their distribution in different functional categories.

Figure 5. Schematic representation of species distributions for the homologous proteins found by BlastP in the Blast2GO annotation process.