Resolution of MALDI-TOF compared to whole genome sequencing for identification of *Bacillus* species isolated from cleanrooms at NASA Johnson Space Center

Farnaz Mazhari[1,2], Aaron B. Regberg[3]*, Christian L. Castro[4] and Michael G. LaMontagne[1]*

[1]Department of Biology and Biotechnology, University of Houston-Clear Lake, Houston, TX, United States, [2]Jacobs, JETS II Contract, NASA Johnson Space Center, Houston, TX, United States, [3]Astromaterials Research and Exploration Science (ARES) Division, NASA Johnson Space Center, Houston, TX, United States, [4]JES Tech, NASA Johnson Space Center, Houston, TX, United States

Introduction: Bacteria are frequently isolated from surfaces in cleanrooms, where astromaterials are curated, at NASA's Lyndon B. Johnson Space Center (JSC). *Bacillus* species are of particular interest because endospores can endure extreme conditions. Current monitoring programs at JSC rely on culturing microbes from swabs of surfaces followed by identification by 16S rRNA sequencing and the VITEK 2 Compact bacterial identification system. These methods have limited power to resolve *Bacillus* species. Whole genome sequencing (WGS) is the current standard for bacterial identification but is expensive and time-consuming. Matrix-assisted laser desorption - time of flight mass spectrometry (MALDI-TOF MS), provides a rapid, low-cost, method of identifying bacterial isolates and has a higher resolution than 16S rRNA sequencing, particularly for *Bacillus* species; however, few studies have compared this method to WGS for identification of *Bacillus* species isolated from cleanrooms.

Methods: To address this, we selected 15 isolates for analysis with WGS and MALDI-TOF MS. Hybrid next-generation (Illumina) and 3rd-generation (nanopore) sequencing were used to draft genomes. Mass spectra, generated with MALDI-TOF MS, were processed with custom scripts to identify clusters of closely related isolates.

Results: MALDI-TOF MS and WGS identified 13/15 and 9/14 at the species level, respectively, and clusters of species generated from MALDI-TOF MS showed good agreement, in terms of congruence of partitioning, with phylotypes generated with WGS. Pairs of strains that were > 94% similar to each other, in terms of average amino acid identity (AAI) predicted by WGS, consistently showed cosine similarities of mass spectra >0.8. The only discordance was for a pair of isolates that were classified as *Paenibacillus* species. This pair showed relatively high similarity (0.85) in terms of MALDI-TOF MS but only 85% similarity in terms of AAI. In addition, some strains isolated from cleanrooms at the JSC appeared closely related to strains isolated from spacecraft assembly cleanrooms.

Discussion: Since MALDI-TOF MS costs less than whole genome sequencing and offers a throughput of hundreds of isolates per hour, this approach appears to offer a cost-efficient option for identifying *Bacillus* species, and related microbes, isolated during routine monitoring of cleanrooms and similar built environments.

1 Introduction

NASA has maintained cleanrooms at the Johnson Space Center (JSC) for curating extraterrestrial samples from the moon, meteorites, cosmic dust, asteroids, comets, solar wind particles, and micrometeorite impacts on space-exposed hardware starting with the lunar samples from the Apollo missions in 1969 (McCubbin et al., 2019). Oligotrophic, low humidity conditions, regular cleaning and air filtration render these facilities inhospitable to microbial life. Despite these fastidious controls, the cleanrooms contain bacteria and fungi (Regberg et al., 2018), which could alter the composition of astromaterials and confound searches for extraterrestrial life (Rummel, 2001). The presence of bacteria and fungi in the curation cleanrooms is acceptable because current astromaterials collections do not have biological contamination control requirements; however, future sample collections, such as the Mars sample return missions, will have these requirements (Carrier et al., 2021). To prepare for these missions, NASA has developed a routine microbial monitoring program for existing collections (McCubbin et al., 2019). *Bacillus* species comprise, on average, 45% of the microbes cultured in this curation program at the JSC (Mazhari, 2021) and are frequently isolated from cleanrooms at the NASA Jet Propulsion Laboratory (JPL) (Tirumalai et al., 2013; Tirumalai et al., 2018).

Strain-level identification of microbes recovered from cleanrooms is important for developing a robust microbial source tracking program and overall contamination control strategy (Song et al., 2024); however, 16S rRNA sequencing, which is widely used to identify bacterial isolates (Church et al., 2020), lacks the resolution required to differentiate closely related *Bacillus* species. For example, 16S rRNA gene sequences of many *Bacillus* species, that occupy fundamentally different environmental niches, are over 99% identical (Yamada et al., 1999). This limits the utility of 16S rRNA gene sequencing for monitoring diversity within curation cleanrooms (Espariz et al., 2016). Whole genome sequencing (WGS) has emerged as the definitive method for microbial identification (Kwong et al., 2015) and is widely used for tracking food-borne pathogens and disease outbreaks (Brown et al., 2019). However, building a library for microbial identification by WGS costs at least $400 per isolate (Brown et al., 2021) and requires highly trained personnel to generate and interpret the data.

Matrix-assisted laser desorption – time of flight mass spectrometry (MALDI-TOF MS) systems provide strain-level identification of microbes, for less than a dollar an isolate, in seconds (Gagné-Bourque et al., 2015). For example, MALDI-TOF can differentiate strains with different functional properties that produce and are resistant to antibiotics (Flores-Treviño et al., 2019) and differentiate between pathogenic and virulent strains of *Bacillus* species (Celandroni et al., 2016). MALDI-TOF MS systems use pattern matching between mass spectra generated from isolates and mass spectra generated from reference strains (Singhal et al., 2015). This high-throughput method is cheaper and more accurate than conventional biochemical systems for bacterial identification (Lévesque et al., 2015; Seng et al., 2009) and for identifying *Bacillus* species (Lasch et al., 2009) and related genera

(Celandroni et al., 2016). MALDI-TOF MS appears comparable to WGS for identification of the specific pathogenic bacteria (da Silva et al., 2020; Nithimongkolchai et al., 2023; Rudolph et al., 2019; Werinder et al., 2021); however, these systems have limitations (Haider et al., 2023; Rychert, 2019). For example, MALDI-TOF MS systems can struggle to differentiate some species related to *Bacillus cereus* (Muigg et al., 2022) and lack of suitable reference spectra has limited the application of MALDI-TOF MS outside of the field of clinical microbiology (Rahi et al., 2016). Further, variation between strains of bacteria, can influence the performance of MALDI-TOF systems, so there is a need to evaluate this approach for bacteria isolated from different environments (Emami et al., 2016; Topić Popović et al., 2023). This need is particularly pressing for extreme environments (Kopcakova et al., 2014), including cleanrooms (da Costa et al., 2022) and similar facilities (Seuylemezian et al., 2018).

In this study, we compared the resolution of MALDI-TOF MS to WGS for a set of bacteria isolated from cleanrooms at the Johnson Space Center. Draft genomes of 14 bacteria were assembled with a hybrid of Illumina and Oxford Nanopore sequencing and genomic relationships were characterized using an estimated maximum-likelihood phylogenomic tree. The resolution of WGS, in terms of average amino acid identity, was then compared to MALDI-TOF MS with custom scripts following LaMontagne (LaMontagne et al., 2021). MALDI-TOF MS showed species-level resolution, which is comparable to WGS.

2 Materials and methods

2.1 Sample collection and genus-level identification

As part of routine microbial monitoring of cleanrooms, samples were collected from the Meteorite, Cosmic Dust, Star Dust, Lunar, Genesis, and Hayabusa labs, and a temporary Cold Curation facility at JSC (Mazhari, 2021). Puritan Brand, Sterile, DNA Free, Foam-Tipped-Applicator (Part Number: 25–1805 1PF RND FDNA) and Puritan Brand, Sterile, Polyester-Tipped-Applicators (Part Number: 25–1000 1PD) were used to sample 300 cm^2 areas of cleanroom surfaces described in Table 1. Air sampling of the meteorite lab was conducted using the SAS Super 180 air sampler at 180 L/ min for 2 min. (360 L). using one Tryptic Soy agar (TSA) and one Sabouraud Dextrose agar (SDA) plate. Samples were transported to a dedicated microbiology lab and the swabs were resuspended in 15 mL of phosphate-buffered saline (PBS) and vortexed for 10 s. Four Tryptic soy agar (TSA), two Blood agar (BA) were inoculated with 100 μL PBS suspensions. Two Reasoner's 2A agar (R2A), two Sabouraud Dextrose agar (SDA), one Sabouraud Dextrose Chloramphenicol agar (SDA + C), and one Potato Dextrose agar (PDA) plates were inoculated with a 300 μL PBS suspension. TSA plates were analyzed following a 48 h incubation at 35°C. BA plates were analyzed following a 48 h incubation at 35°C. R2A plates were analyzed following a seven-day incubation at 25°C. PDA plates were analyzed following a

TABLE 1 Summary of hybrid genome assemblies.

Sample name	Curation lab	Equivalent ISO level[a]	Location	BioSample ID	No. contigs	Genome size (bp)	N50 (bp)	GC content (%)	Coverage (x)[b]	Genome Completion (%)
1813sda1	Cold	N/A	Glove box	SAMN29043843	2	5,662,393	5,204,374	34	77	100
2987tsa1	Cosmic Dust	5	Floor	SAMN29043850	1	4,088,722	4,088,622	46	92	99
1735sda2	Hayabusa	5	Floor	SAMN29043840	4	4,055,684	4,035,936	38	55	100
2090tsa1	Hayabusa	5	Table	SAMN29043852	7	5,745,698	5,311,671	34	88	99
1780r2a1	Lunar	6	Floor	SAMN29043839	1	4,184,516	4,184,416	38	108	99
1781tsa1	Lunar	6	Floor	SAMN29043842	2	7,080,870	7,075,035	45	26	99
2933tsa1	Lunar	6	Air	SAMN29043848	3	5,356,151	5,043,971	40	67	98
2941sda1	Lunar	6	Floor	SAMN29043849	6	6,929,459	6,868,407	49	17	99
1943r2a1	Lunar	6	Floor	SAMN29043844	1	3,749,596	3,749,496	41	77	100
2069sda1	Meteorite	7	Air	SAMN29043846	3	5,795,782	5,505,279	38	55	98
1370ba1	Meteorite	7	Floor	SAMN29043836	1	3,659,009	3,658,909	42	141	99
1480ba3	Meteorite	7	Floor	SAMN29043837	6	4,111,671	3,935,672	36	152	98
1663tsa1	Meteorite	7	Floor	SAMN29043838	38	5,550,983	614,346	36	82	100
2047tsa1	Meteorite	7	Table	SAMN29043845	3	3,909,693	3,692,152	39	124	99
1570r2a1	Cosmic Dust	5	Microscope	N/A[c]	6	10,188,618	5,353,115	35	54	100

[a]International Standards Organization (ISO) level (ISO 14644-1).
[b]Calculated for MiSeq reads only.
[c]Assembly removed because of contamination.

7-day incubation at 30°C. Isolates were sub-cultured onto TSA plates and identified at the genus level using the Vitek 2 compact bacterial identification system (bioMérieux USA, St. Louis, MO) or Applied Biosystems Applied Biosystems 3500 Series Genetic Analyzer (Applied Biosystems, Waltham, MA) previously (Mazhari, 2021). Isolates were sub-cultured onto TSA plates and incubated for 24 h at 35°C. A 10 μL loopful of bacteria was transferred into a Microbank® tube (Pro-Lab Diagnostics, Round Rock, TX), which contains a proprietary cryopreservation solution, vortexed for 30 s and stored at −80°C.

2.2 DNA isolation

For WGS, isolates were removed from the −80°C freezer, sub-cultured onto TSA plates and incubated at 35°C for 24 h. Isolates were sub-cultured after 24 h. to fresh plates. Individual colonies were then inoculated into 20 mL Hardy dx (cat no. Q85) TSB tubes. Tubes were incubated at 35°C while shaking at 200 rpm with loosened lids. After 24 h. liquid cultures were centrifuged at 10,000×g for 10 min. The cell pellet was resuspended in 1 mL sterile PBS and then centrifuged at 13,000×g for 2 min. to pellet the cells. The cell pellets were then resuspended in 480 μL of 50 mM (pH 8.0) ethylenediaminetetraacetic acid (EDTA) and 120 μL of 10 mg/mL egg white lysozyme, ultra-pure grade (Amresco, Solon, OH) was added to the resuspended cell pellet for cell lysis. Isolates were incubated at 37°C for 30–60 min. and centrifuged at 13,000×g for 2 min. The supernatant was removed and the Promega Wizard™ Genomic DNA Purification Kit was used for DNA extraction following the manufacturer's protocol (Promega, Madison, WI). The Qubit 1X dsDNA High Sensitivity assay kit (Invitrogen, Waltham MA) was used to quantify DNA concentration from extracted DNA of samples. DNA size (~ 600 bp) and integrity was assessed with 4,200 TapeStation System with the Genomic DNA ScreenTape assay (Agilent, Palo Alto, CA).

2.3 Hybrid whole genome sequencing and post processing

A MinION Mk1C from Oxford Nanopore Technologies (ONT, Oxford, UK) was used for long-read DNA sequencing using the Rapid Barcoding Sequencing (SQK-RBK004,RBK_9054_v2_ revQ_14Aug2019) kit according to manufacturer's protocol. Two sequencing runs were conducted consisting of 8 and 12 DNA samples (100–400 ng). Samples were fragmented and barcoded then suspended in AMPure XP beads for cleaning and concentration. DNA concentration was measured using the Qubit Fluorometric Quantitation 1x dsDNA High Sensitivity Kit (ThermoFisher) and then added to 1 μL of rapid adapter buffer (ONT). The library was loaded onto a R.9.4.1 flow cell (ONT) under the MinKNOW program version 4.2.4 (ONT)[1] was used to control a sequencing run of 72 h. Sequence reads were base-called and demultiplexed using Guppy version 4.3.4 (ONT); the high accuracy configuration in

Guppy was used with a minimum quality score of 7 and barcode score of 40.

An Illumina MiSeq was used for short-read DNA sequencing using the MiSeq Reagent Kit v3 with paired-end reads, following the Illumina DNA prep reference guide # 1000000025416 v09; 400 ng of high molecular weight DNA was used for library prep. Tagmentation, the process of cleaving templated DNA and addition of adapters, was followed according to the protocol provided by Illumina. DNA concentration was measured for 15 libraries using the Qubit Fluorometric Quantitation 1x dsDNA High Sensitivity Kit (ThermoFisher). Each library was normalized to 4 nM and pooled to 12 pM in a total volume of 180 μL. Denature and dilution instructions were followed according to MiSeq® instrument protocol with libraries loaded onto MiSeq® Reagent v3 cartridge. The MiSeq Reporter software version 2.6.2.3 was used to demultiplex and generate the fastq files for the Illumina short reads.

For hybrid genome processing, FastQC v0.11.9 (Andrews, 2010) and MultiQC v1.10.1 (Ewels et al., 2016) were used to assess sequence quality of both Illumina short reads and Nanopore long reads. The Bioinformatics platform EDGE Bioinformatics (Li et al., 2016) was used to analyze and assemble the raw sequence data using the hybrid whole genome pipeline. Reads were trimmed at a quality level of 30. The minimum sequence length was set to 50. The "N" base cut off was set to 10. Reads with adapters or contamination sequences were trimmed using Porechop (Wick et al., 2017a) and homopolymers greater than 15 poly A were removed. De novo assembly was done using Unicycler v0.4.9. (Wick et al., 2017b) at a minimum contig length of 200 bp and a minimum of 2000 reads. Miniasm (Li, 2018) was used to find consensus sequences at a minimum of 3. Reads were mapped with bowtie2 v2.4.3 (Langmead et al., 2018) at a max clip (number of clipped read characters) of 50 and a min mapq (mapping quality score) of 42. Contig taxonomy was classified with BLAST (Camacho et al., 2009). The assembled draft genomes were submitted for annotation using NCBI Prokaryotic Genome Annotation Pipeline (PGAP). The closest bacterial strain identity was determined using the Type (Strain) Genome Server (TYGS) (© 2016–2022 Leibniz Institute DSMZ) by comparing all genomes with available strain genomes in TYGS database and by extracting the 16 s rRNA sequence and aligning against the TYGS database. Genome quality and completeness was assessed using CheckM version 1.0.18 (Parks et al., 2015). A pairwise similarity of amino acids was generated with EzAAI (Kim et al., 2021) with default parameters. A16S rRNA maximum likelihood phylogenetic tree was constructed using ETE3 3.1.2 (Huerta-Cepas et al., 2016) aligned using MAFFT v6.861b on the GenomeNet[2] and inferred using RAxML v8.2.11 on34 model GTRGAMMA with default parameters (Stamatakis, 2014). An estimated maximum likelihood phylogenomic tree using 119 single copy genes (SCGs) from the phylum Bacillota was created using GToTree (Lee, 2019) under default parameters. Reference sequences were found using the tool Megablast in NCBI. Phylogenomic tree was visualized in iTOL interactive tree of life (Letunic and Bork, 2021).

---

1 https://community.nanoporetech.com/docs/prepare/library_prep_protocols/experiment-companion-minknow/v/mke_1013_v1_revdc_11apr2016

2 https://www.genome.jp/tools/ete/

2.4 MALDI-TOF MS

Isolates were cultured for 24 hours at 35°C on tryptic soy agar (TSA) plates. Isolates were prepared for analysis by ethanol treatment and formic acid extractions as previously described (Freiwald and Sauer, 2009). Briefly, 300 μL of Ultra-Pure Water, High Performance Liquid Chromatography Mass Spectrometer (HPLC/MS) Grade 11, was added to each 1.5 mL microfuge tube (Eppendorf, Hamburg, Germany). A 10 μL loop of a single colony was inoculated from the plates into separate microfuge tubes for each isolate and vortexed until homogeneous before adding 900 μL of ethanol, 100% HPLC/MS Grade 12, and then vortexing for 30 s. Tubes were stored at 4°C and then centrifuged at 13,000×*g* speed for 2 min. to pellet the cells. The supernatant was decanted, and the cell pellet was centrifuged for another minute at 13,000×*g* before excess liquid was removed with a pipette and left to air dry. After 5 min. of drying, 50 μL of 70% HPLC/MS grade formic acid was added and tubes were vortexed for 30 s. and incubated for 5 min. at room temperature. An equal volume (50 μL) of 100% HPLC/MS Grade acetonitrile was added and the suspension was vortexed for 30 s. prior to being centrifuged at maximum speed (13,000×*g*) for 2 min. Taking care to avoid the pellet, 70 μL of supernatant was transferred to a fresh 1.5 mL microfuge tube and stored at −20°C. For analysis, 1 μL of supernatant was pipetted onto a MALDI steel target (Bruker p/n 8280800). Two size standards were prepared by applying 1 μL of bacterial test standard (BTS, Bruker p/n 8255343) solution and two spots were left blank. After the spots dried, 1 μL of matrix solution (HCCA matrix, Bruker) was added to each spot and the target was allowed to thoroughly air dry for 10 min. Targets were shipped overnight, with an ice pack, to the Proteomics and Mass Spectrometry Core Facility at Pennsylvania State University, (University Park, PA). Positive-ion mass spectra were acquired on a Bruker Ultraflextreme MALDI-TOF/TOF mass spectrometer as described previously (LaMontagne et al., 2021).

MALDI-TOF MS data was analyzed with a script modified from LaMontagne et al. (2021). Briefly, mass spectra were analyzed by cluster analysis using an R script that contains functions from MALDIquant (Gibb and Strimmer, 2012) for processing mass spectra. Pvclust (Suzuki and Shimodaira, 2006) was used to provide bootstrap probability values for clusters in the dendrogram. Philentropy (Drost, 2018) was used to calculate distance and similarity measures. iNEXT (Hsieh et al., 2016) was used for rarefaction analysis and ggplot2 (Wickham, 2016) was used to for biplots. RWeka (Hornik et al., 2009) was to test the coherence of MALDI-TOF taxonomic units (MTUs). This script first ran a preliminary cluster analysis to calculate cosine similarities between spectra generated from pairs of mass spectra calculate a signal-to-noise ratio (SNR) for mass spectra. Cosine similarities were calculated following (Strejcek et al., 2018) and a histogram of cosine similarities to define a threshold for MTUs. This threshold was set to a cosine similarity of 0.7. We then ran a loop that iteratively sampled random values for seven parameters (half-window for smoothing, baseline removal, half-window for alignment, alignment tolerance, SNR for alignment, half-window for peak detection, and peak detection SNR) that are used by MALDIquant to align spectra and detect peaks. This loop was run 1,200 times. Jaccard coefficients, calculated following (Starostin et al., 2015), were used as the optimization parameter. The output of the alignment loop was sorted by Jaccard coefficients and the set of parameters that gave the highest number of shared peaks were selected. A quality check was done by calculation of the SNR for the 10 largest peaks in each spectra and plotting these ratios against the number of peaks detected. Spectra with low SNRs (< 11) and few (< 15) peaks were removed after confirmation by visual inspection.

Significance of differences between pairwise similarities was calculated with an ANOVA, using base functions in R, and a Tukey test using the package agricolae (de Mendiburu and Yaseen, 2020). The level of alpha for the Tukey test was set to 0.0001. Agreement between partitioning of isolates was tested with the adjusted Wallace coefficient (Severiano et al., 2011), using an online tool.[3] Partitions for MALDI-TOF data were defined by MTUs. Partitions for WGS were defined by species identifications obtained with the TYGS (see text footnote 3). Isolates that were not typed to the species level by WGS, but showed AAI values greater than 90%, were put in the same partition.

Raw mass spectra are available in Supplementary file S1 and the script used to process this data is available, in R markdown format, in Supplementary file S2. Input and output files to and from this script, including mass spectra and MALDI Biotyper® scores, are available as Supplementary files S3–S6. Scripts used for the Tukey test are available in Supplementary file S7.
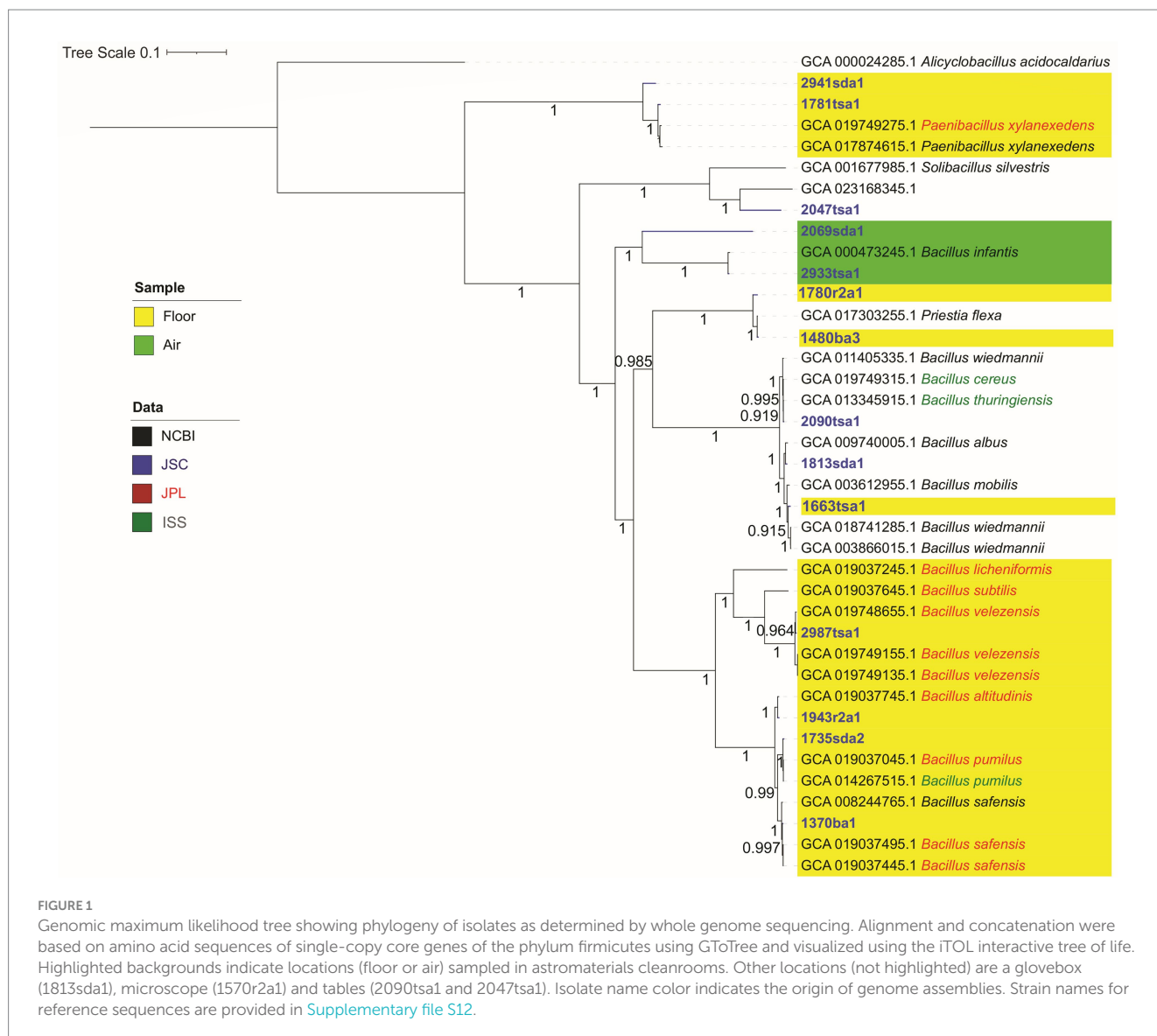
3 Results

The optimization loop showed a modal relationship between Jaccard coefficients and the number of peaks detected. The highest Jaccard coefficients were observed at 163 peaks (Supplementary file S8).

From 15 isolates, 14 draft genomes were *de novo* assembled with greater than 98% mean completion (Table 1). GC contents for these isolates ranged from 34% to 49%, which is consistent with the nucleotide compositions typical for these genera (Lightfield et al., 2011) and mesophilic bacteria in general (Hu et al., 2022). Genome sizes ranged from 7.1 to 3.7 Mbp, which is typical for the phylum Bacillota (Martinez-Gutierrez and Aylward, 2022). Isolate 1781tsa1, which was classified as a *Paenibacillus* species (Figure 1), showed the largest genome size; isolate 1370ba1, which was classified as a strain of *Bacillus safensis* (Figure 1), showed the smallest genome size.

Bacterial identification by WGS generally agreed with identifications by the MALDI system at the genus level (Table 2); however, two isolates were classified as different genera. Isolate 1480ba3 was identified as *Priestia flexus* and isolate 2069sda1 was identified as *Peribacillus frigoritolerans* by WGS. These two isolates were both classified as *Bacillus* species by the MALDI system (Table 2). This reflects the recent assignment of *Bacillus flexus* to the novel genera *Priestia* (Gupta et al., 2020) and the proposal of *Peribacillus* as a novel genus (Patel and Gupta, 2020). Apparently, these changes in nomenclature have not been updated in the proprietary Bruker MALDI Biotyper® database.

The MALDI system and WGS identified 13/15 and 9/14 at the species level, respectively, (Table 2). For 9 isolates identified at the species level by WGS, only 4 identifications were shared with identifications by MALDI-TOF MS. For example, isolate 2987tsa1 was identified as *B. amyloliquefaciens* and *B. velezensis* and isolate

---

3   http://www.comparingpartitions.info/

FIGURE 1
Genomic maximum likelihood tree showing phylogeny of isolates as determined by whole genome sequencing. Alignment and concatenation were based on amino acid sequences of single-copy core genes of the phylum firmicutes using GToTree and visualized using the iTOL interactive tree of life. Highlighted backgrounds indicate locations (floor or air) sampled in astromaterials cleanrooms. Other locations (not highlighted) are a glovebox (1813sda1), microscope (1570r2a1) and tables (2090tsa1 and 2047tsa1). Isolate name color indicates the origin of genome assemblies. Strain names for reference sequences are provided in Supplementary file S12.

2069sda1 was identified as *B. simplex* and *Peribacillus frigoritolerans* by MALDI-TOF MS and WGS and respectively (Table 2). In contrast, 16S rRNA gene taxonomy did not identify any of the isolates at the species level (Mazhari, 2021).

The topology of a phylogenomic tree generated from multiple single-copy core genes appeared largely coherent in terms of isolate identifications (Figure 1). That is, isolates generally formed monophyletic clades with representatives of their genera and the topology of this tree received strong bootstrap support (0.915–1.00). In contrast, a tree generated from 16S rRNA gene sequences (Supplementary file S9) showed several clades with relatively weak bootstrap support (0.24–0.41). Also, clades containing reference genomes for *Bacillus wiedmannii* were polyphylogenetic. This species clustered with two different *Bacillus* clades. Isolates did not appear to cluster by sample location.

Several clades in the phylogenomic tree showed that bacteria isolated from cleanrooms at JSC appeared closely related to bacteria isolated from cleanrooms at JPL, or similar built environments, like the International Space Station (ISS). For example, isolates

identified as *Paenibacillus* species (2941sda1 and 1781tsa1), isolated from the floor of the Lunar curation lab at JSC, clustered with two strains *Paenibacillus xylanexedens* including a strain (PL-73) isolated from an air sample collected in a cleanroom at the JPL (GCA_019749275). Similarly, isolate 2090tsa1, which was isolated from a table in the Hayabusa lab at JSC, showed similarity to a strain (I1-R4) of *Bacillus cereus* isolated from an air sample collected at JPL. Further, *Bacillus* species (2987tsa1, 1943r2a1, 1735sda2, and 1370ba1) isolated from swabs of the cleanroom floors from the JSC Cosmic Dust, Lunar, Hayabusa, and Meteorite labs respectively, clustered with seven *Bacillus* strains isolated from spacecraft assembly cleanrooms at JPL and the ISS with strong bootstrap support (> 0.95, Figure 1).

The topology of a dendogram generated by mass spectra appeared coherent for isolates that were reliably identified at the species level by the MALDI system (Figure 2). In other words, representatives of the same species clustered together with strong bootstrap support (Figure 2). These clusters also appeared in the phylogenomic tree generated by WGS (Figure 1). In particular, edge 3 contained four

TABLE 2 Identification of bacterial isolates by DNA sequencing and MALDI-TOF MS.

Isolate	Edge[a]	MALDI Biotyper® system	Biotyper® score[b]	Whole genome sequencing
2090tsa1	3	*Bacillus cereus*	2.44	*Bacillus cereus*
1570r2a1	3	*Bacillus cereus*	2.28	*N.A.*[c]
1813sda1	3	*Bacillus cereus*	2.32	*Bacillus* sp.
1663tsa1	3	*Bacillus cereus*	2.31	*Bacillus* sp.
1781tsa1	4	*P. amylolyticus*	2.34	*Paenibacillus* sp.
2941sda1	4	*Paenibacillus* sp.	1.96	*P. polysaccharolyticus*
1480ba3	5	*Bacillus flexus*	2.38	*Priestia flexa*
1780r2a1	5	*Bacillus flexus*	2.32	*Bacillus* sp.
1735sda2	6	*Bacillus pumilus*	2.06	*Bacillus* sp.
1943r2a1	6	*Bacillus altitudinis*	2.00	*Bacillus altitudinis*
1370ba1	7	*Bacillus safensis*	2.03	*Bacillus safensis*
2069sda1	8	*Bacillus simplex*	2.25	*Peribacillus frigoritolerans*
2047tsa1	8	*Lysinibacillus* sp.	1.83	*Lysinibacillus endophyticus*
2987tsa1	11	*B. amyloliquefaciens*	2.05	*Bacillus velezensis*
2933tsa1	12	*Bacillus infantis*	2.35	*Bacillus infantis*

[a]Nodes defined in Figure 2.
[b]Top score from MALDI Biotyper® system. According to the manufacturer, a score > 1.8 is "probable genus identification," > 2.0 is "secure species identification" and > 2.2 is "highly probably species identification."
[c]Assembly removed because of contamination.

isolates that classified as *Bacillus cereus* (Figure 2, Table 2). This clade corresponded to a clade, generated by WGS, that contained *B. cereus* and related species (Figure 1). Similarly, edge 5 contained two isolates that were classified as *Bacillus flexus* (Figure 2, Table 2). This clade corresponded to a clade, generated by WGS, that contained *Priestia flexa* and related species (Figure 1). A clade (edge 4) that contained two isolates that were classified as *Paenibacillus* species also received strong bootstrap support. Clades containing singletons generally appeared unreliable; however, edge 6 contained two *Bacillus* species that appeared closely related to each other in terms of mass spectra (Figure 2) and WGS (Figure 1).

Partitioning of isolates into MTUs, by MALDI-TOF, generally agreed with partitions defined by WGS. The adjusted Rand coefficient, for congruence of microbial typing method, was 0.95 and the adjusted Wallace coefficient was 0.90, with a 95% confidence interval of 0.80 to 1.00.

Pairs of isolates that were classified as members of the same species by the MALDI system, showed high similarity to each other in terms of WGS (Figure 3). This supports the hypothesis that MALDI-TOF MS and WGS are comparable. Specifically, pairwise similarity of the amino acid sequences from single-copy core genes, predicted from the draft genomes, ranged from 93% to 98% AAI for pairs of isolates from the same species. These pairs showed cosine similarities, generated from mass spectra, that ranged from 0.83 to 0.90 with an average of 0.88 (Figure 3). Pairwise comparisons of mass spectra between members of the same species were significantly ($p = 0.0001$) larger than pairwise comparisons for members of different species (Figure 3, Supplementary file S7).

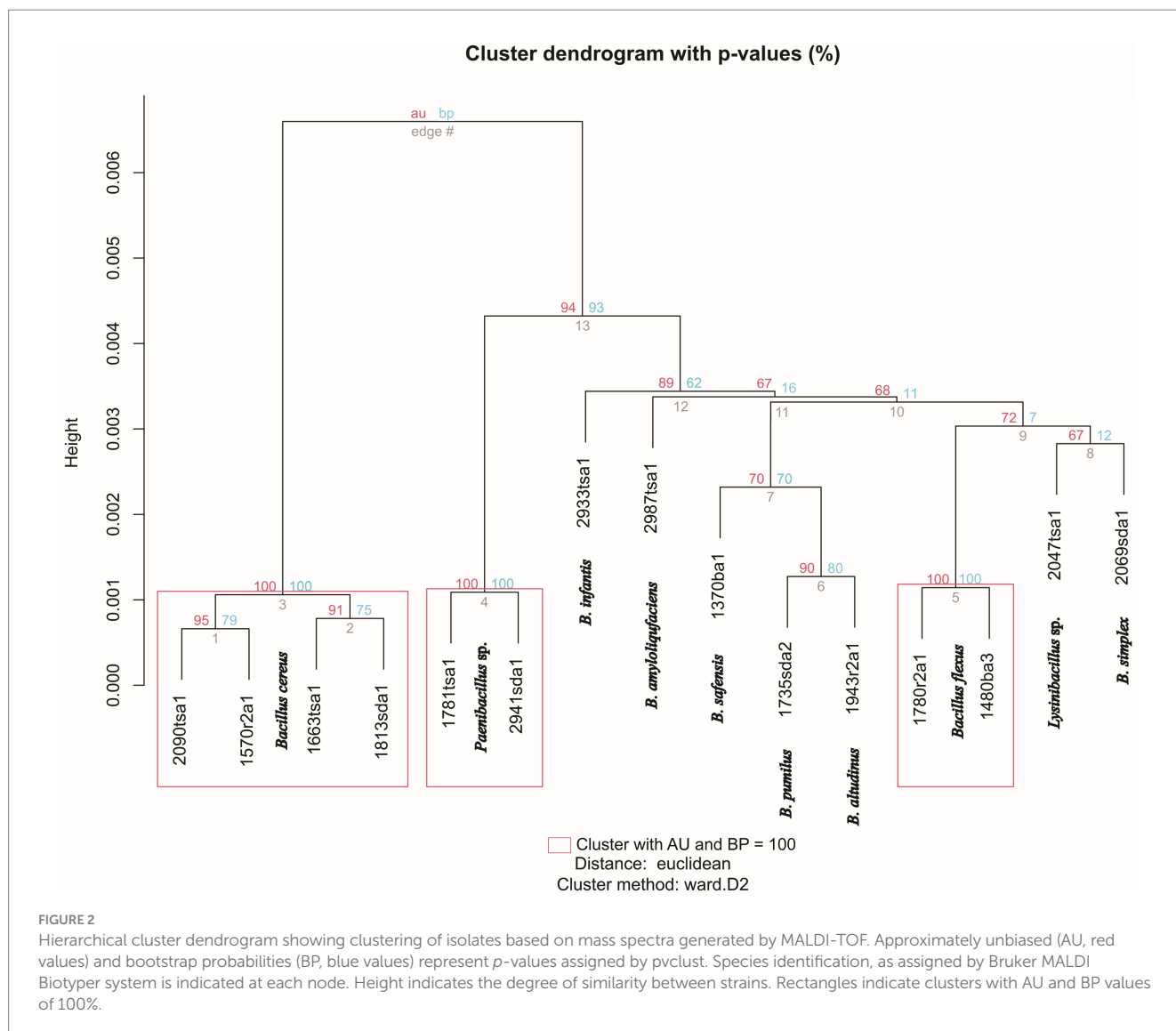Pairwise similarities of mass spectra at the genus level were on average lower (0.42); however, the pair of *Paenibacillus* species showed a high similarity to each other in terms of cosine similarity of mass spectra (0.85) but a relatively low AAI (Figure 3). Similarly, three *Bacillus* isolates (1370sda2, 1735sda2, and 1943r2a1), which were classified as different species by both WGS sequencing and mass spectra (Table 2), showed relatively high AAI similarities (93–95%). These three isolates showed cosine similarities of 0.68 to 0.74, which were lower than expected given trends in Figure 3.

Pairwise comparisons of isolates that were assigned to different families (Paenibacillaceae and Bacillaceae) of the order Bacillales, and different genera (*Lysinibacillus* and *Bacillus*) of the family Bacillaceae, were resolved by AAIs but not by cosine similarities (Figure 3). These comparisons at the order and family level averaged 57 (56–57)% and 61 (60–62)%, respectively, for AAIs and 0.40 (0.33–0.47), 0.41 (0.35–0.46) respectively for cosine similarities (Supplementary file S7).

At a cosine similarity of 0.8, which appeared to resolve genera and species of *Bacillus* (Figure 3), MALDI-TOF MS detected 10 MALDI-TOF taxonomic units (MTUs) in a library of 15 isolates. This library contained three MTUs with more than one isolate and seven singletons. This corresponds to a coverage of 59%; extrapolation to a library of 28 isolates would give 80% coverage (Supplementary file S10).

4 Discussion

MALDI-TOF showed a resolution comparable to that of WGS. Both MALDI-TOF MS and WGS identified *Bacillus* strains consistently at the genus level (Table 2) and the relationship between isolates as assessed by phylogenomic analysis (Figure 1) agreed with clustering generated by mass spectra (Figure 2). Comparison of the similarity between isolates as assessed by WGS, suggested a threshold

FIGURE 2
Hierarchical cluster dendrogram showing clustering of isolates based on mass spectra generated by MALDI-TOF. Approximately unbiased (AU, red values) and bootstrap probabilities (BP, blue values) represent *p*-values assigned by pvclust. Species identification, as assigned by Bruker MALDI Biotyper system is indicated at each node. Height indicates the degree of similarity between strains. Rectangles indicate clusters with AU and BP values of 100%.

for identifying species of *Bacillus* isolates by MALDI-TOF MS. Pairs of strains that showed cosine similarities of mass spectra of >0.80, were reliably identified at the species level by the MALDI system and showed AAI, as assessed by WGS, of >92% (Figure 3). The high cosine similarities for comparisons within *Bacillus* species relative to comparison between *Bacillus* species (Figure 3) confirms that MALDI-TOF MS provides an accurate representation of species diversity and can differentiate between related *Bacillus* species (Celandroni et al., 2016; Hotta et al., 2011; Lasch et al., 2009). This suggests MALDI-TOF provides a quick, cost-effective, and accurate method of identifying microbes that contaminate astromaterials curation facilities.

This threshold for defining a species by MALDI-TOF MS agrees with the cosine similarity threshold of 0.79 set by Strejcek et al. (2018). This proposed threshold of cosine similarity threshold (0.80) could result in the misclassification of *Paenibacillus* strains, which appear to be different species as assessed by AAI (Figure 3). AAI levels of >92%, observed for within-species comparisons is lower than the widely used threshold for defining a species for prokaryotes by AAI of 95%

(Konstantinidis and Tiedje, 2005); however, this higher threshold for AAI levels may be biased towards clinical isolates (Rosselló-Mora, 2005).

Phylogenomic analysis showed several isolates clustered with *Bacillus* spp. isolated from other studies of cleanrooms and the ISS (Figure 1). This suggests there is a cosmopolitan class of *Bacillus* strains associated with cleanrooms and similar built environments, including the ISS (Blaustein et al., 2019). Microbes in the ISS and cleanrooms face similar selective pressures, including low nutrient availability and humidity. These "extreme environments" (Checinska Sielaff et al., 2019; La Duc et al., 2003) may share many microbes; however, this biogeographical pattern was not consistent. For example, the isolate identified as *Lysinbacillus endophyticus* (2047tsa1) clustered with a *Lysinbacillus* species isolated from the ant microbiome and the isolate that classified as *Bacillus infantis* (2933tsa1) clustered with a *Bacillus* species isolated from a marine system. Clearly, these species are not exclusively observed in cleanrooms; however, their detection can be explained. *Lysinibacillus* species have been isolated from air samples (Wong et al., 2019) and novel species, that are closely related
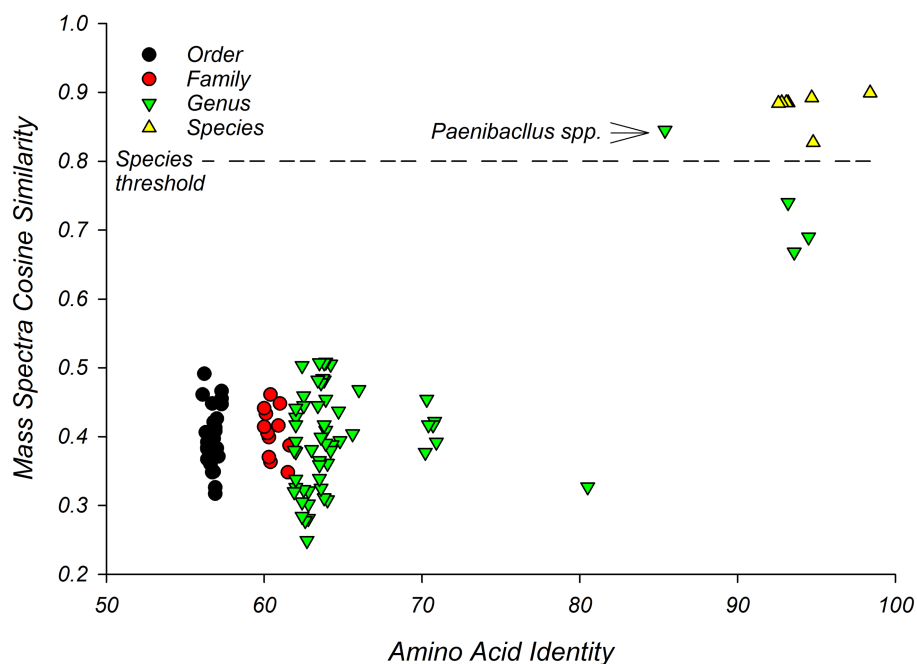
FIGURE 3
Pairwise similarity graph of MALDI-TOF and whole genome sequencing showing threshold for defining *Bacillus* species. Comparison of similarity between strains as assessed by MALDI-TOF MS and WGS shows that pairs of strains of the same *Bacillus* species showed cosine similarities of mass spectra of greater than 0.8. This threshold is indicated with a dashed line. X-axis presents amino acid identity between pairs of isolates. Y-axis presents pairwise similarity coefficients calculated from mass spectra. The comparison between the two *Paenibacillus* species is indicated.

to *Bacillus infantis,* have been isolated from cleanrooms (Seiler et al., 2013).

For routine monitoring, this proteomics approach could replace Vitek2 and 16S rRNA sequencing, which are widely used in environmental monitoring programs. The resolution of MALDI-TOF MS could also help investigators prioritize isolates for deeper analysis by WGS and biochemical tests. For example, genetic and functional characteristics of MTUs could be used to identify novel species with unique metabolic capabilities. The high resolution of MALDI-TOF also allows for tracking the source of microbial contamination, as has been applied to aquatic systems (Giebel et al., 2008; Siegrist et al., 2007).

4.1 Limitations

Sample size limits broader interpretation of the above results. Only 15 isolates were selected for this study. This corresponds to a coverage of a little more than half of the MTUs in the system sampled (Supplementary file S10). Accordingly, many of the isolates in the library were singletons (7/15), so in terms of their mass spectra we have little to compare them to internally. Despite the small sample size, precision of measurements provided the power to compare MALDI-TOF MS and WGS. Specifically, a Tukey test run at an alpha of 0.0001 showed that similarities of mass spectra generated by MALDI-TOF was significantly higher for comparisons within species than comparisons between species (Supplementary file S7).

The topology of the dendrogram generated from mass spectra (Figure 2) did not appear coherent for isolates that were not closely related, in terms of AAI or mass spectra. This may reflect the low similarity of mass spectra for isolates that are not members of the same species (Figure 3) and is consistent with previous studies that showed that dendrograms generated by hierarchical clustering of mass spectra do not consistently follow phylogenetically coherent topologies for *Bacillus* (Lasch et al., 2009) and *Lactobacillus* species (Douillard et al., 2013).

Lack of reference spectra can limit the ability of the MALDI-TOF system to identify bacteria isolated from built and natural environments (Popović et al., 2017). This can be addressed by developing internal databases, which has been initiated for *Bacillus* species isolated from cleanrooms (da Costa et al., 2022) and spacecraft assembly facilities (Seuylemezian et al., 2018). These custom databases improve with the addition of mass spectra from multiple strains. For example, Erler et al. (2015) generated a custom database for identifying *Vibrio* species isolated from aquatic environments; this expanded database of 997 mass spectra dramatically improved bacterial identification by MALDI-TOF MS.

5 Summary

MALDI-TOF MS can efficiently identify species of the genus *Bacillus* that are frequently isolated from facilities where astromaterials are curated, and similar built environments, with a resolution comparable to WGS. Implementation of this proteomics approach would require development of database of mass spectra.

Data availability statement

Author contributions

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2025.1499516/full#supplementary-material

References

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data.

Blaustein, R. A., McFarland, A. G., Maamar, S. B., Lopez, A., Castro-Wallace, S., and Hartmann, E. M. (2019). Pangenomic approach to understanding microbial adaptations within a model built environment, the international Space Station, relative to human hosts and soil. *mSystems* 4, e00281–e00218. doi: 10.1128/msystems.00281-18

Brown, B., Allard, M., Bazaco, M. C., Blankenship, J., and Minor, T. (2021). An economic evaluation of the whole genome sequencing source tracking program in the U.S. *PLoS One* 16:e0258262. doi: 10.1371/journal.pone.0258262

Brown, E., Dessai, U., McGarry, S., and Gerner-Smidt, P. (2019). Use of whole-genome sequencing for food safety and public health in the United States. *Foodborne Pathog. Dis.* 16, 441–450. doi: 10.1089/fpd.2019.2662

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10:421. doi: 10.1186/1471-2105-10-421

Carrier, B., Beaty, D., Hutzler, A., Smith, A., Kminek, G., Meyer, M., et al. (2021). Science and curation considerations for the design of a Mars sample return (MSR) sample receiving facility. *Astrobiology* 22, S-217–S-237. doi: 10.1089/ast.2021.0110

Celandroni, F., Salvetti, S., Gueye, S. A., Mazzantini, D., Lupetti, A., Senesi, S., et al. (2016). Identification and pathogenic potential of clinical *Bacillus* and *Paenibacillus* isolates. *PLoS One* 11:e0152831. doi: 10.1371/journal.pone.0152831

Checinska Sielaff, A., Urbaniak, C., Mohan, G. B. M., Stepanov, V. G., Tran, Q., Wood, J. M., et al. (2019). Characterization of the total and viable bacterial and fungal communities associated with the international space station surfaces. *Microbiome* 7:50. doi: 10.1186/s40168-019-0666-x

Church, D., Cerutti, L., Gürtler, A., Griener, T., Zelazny, A., and Emler, S. (2020). Performance and application of 16S rRNA gene cycle sequencing for routine identification of bacteria in the clinical microbiology laboratory. *Clin. Microbiol. Rev.* 33:e00053. doi: 10.1128/CMR.00053-19

da Costa, L. V., de Miranda, R. V. D. S. L., dos Reis, C. M. F., de Andrade, J. M., Cruz, F. V., Frazão, A. M., et al. (2022). MALDI-TOF MS database expansion for identification of *Bacillus* and related genera isolated from a pharmaceutical facility. *J. Microbiol. Methods.* 203:106625. doi: 10.1016/j.mimet.2022.106625

da Silva, D. A. V., Brendebach, H., Grützke, J., Dieckmann, R., Soares, R. M., de Lima, J. T. R., et al. (2020). MALDI-TOF MS and genomic analysis can make the difference in the clarification of canine brucellosis outbreaks. *Sci. Rep.* 10:19246. doi: 10.1038/s41598-020-75960-3

de Mendiburu, F., and Yaseen, M. (2020) Agricolae: Statistical procedures for agricultural research. 1.40.

Deutsch, E. W. (2010). Mass spectrometer output file format mzML. *Methods Mol. Biol.* 604, 319–331. doi: 10.1007/978-1-60761-444-9_22

Douillard, F. P., Ribbera, A., Kant, R., Pietilä, T. E., Järvinen, H. M., Messing, M., et al. (2013). Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet.* 9:e1003683. doi: 10.1371/journal.pgen.1003683

Drost, H.-G. (2018). Philentropy: information theory and distance quantification with R. *J. Open Source Softw.* 3:765. doi: 10.21105/joss.00765

Emami, K., Nelson, A., Hack, E., Zhang, J., Green, D. H., Caldwell, G. S., et al. (2016). MALDI-TOF mass spectrometry discriminates known species and marine environmental isolates of *Pseudoalteromonas*. *Front. Microbiol.* 7:104. doi: 10.3389/fmicb.2016.00104

Erler, R., Wichels, A., Heinemeyer, E.-A., Hauk, G., Hippelein, M., Reyes, N. T., et al. (2015). VibrioBase: a MALDI-TOF MS database for fast identification of Vibrio spp. that are potentially pathogenic in humans. *Syst. Appl. Microbiol.* 38, 16–25. doi: 10.1016/j.syapm.2014.10.009

Espariz, M., Zuljan, F. A., Esteban, L., and Magni, C. (2016). Taxonomic identity resolution of highly phylogenetically related strains and selection of phylogenetic

markers by using genome-scale methods: the *Bacillus pumilus* group case. *PLoS One* 11:e0163098. doi: 10.1371/journal.pone.0163098

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Flores-Treviño, S., Garza-González, E., Mendoza-Olazarán, S., Morfín-Otero, R., Camacho-Ortiz, A., Rodríguez-Noriega, E., et al. (2019). Screening of biomarkers of drug resistance or virulence in ESCAPE pathogens by MALDI-TOF mass spectrometry. *Sci. Rep.* 9:18945. doi: 10.1038/s41598-019-55430-1

Freiwald, A., and Sauer, S. (2009). Phylogenetic classification and identification of bacteria by mass spectrometry. *Nat. Prot.* 4, 732–742. doi: 10.1038/nprot.2009.37

Gagné-Bourque, F., Mayer, B. F., Charron, J.-B., Vali, H., Bertrand, A., and Jabaji, S. (2015). Accelerated growth rate and increased drought stress resilience of the model grass *Brachypodium distachyon* colonized by *Bacillus subtilis* B26. *PLoS One* 10:e0130456. doi: 10.1371/journal.pone.0130456

Gibb, S., and Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 28, 2270–2271. doi: 10.1093/bioinformatics/bts447

Giebel, R. A., Fredenberg, W., and Sandrin, T. R. (2008). Characterization of environmental isolates of *Enterococcus* spp. by matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry. *Water Res.* 42, 931–940. doi: 10.1016/j.watres.2007.09.005

Gupta, R. S., Patel, S., Saini, N., and Chen, S. (2020). Robust demarcation of 17 distinct *Bacillus* species clades, proposed as novel *Bacillaceae* genera, by phylogenomics and comparative genomic analyses: description of *Robertmurraya kyonggiensis* sp. nov. and proposal for an emended genus *Bacillus* limiting it only to the members of the subtilis and cereus clades of species. *Int. J. Syst. Evol. Microbiol.* 70, 5753–5798. doi: 10.1099/ijsem.0.004475

Haider, A., Ringer, M., Kotroczó, Z., Mohácsi-Farkas, C., and Kocsis, T. (2023). The current level of MALDI-TOF MS applications in the detection of microorganisms: a short review of benefits and limitations. *Microbiol. Res.* 14, 80–90. doi: 10.3390/microbiolres14010008

Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Comput. Stat.* 24, 225–232. doi: 10.1007/s00180-008-0119-7

Hotta, Y., Sato, J., Sato, H., Hosoda, A., and Tamura, H. (2011). Classification of the genus *Bacillus* based on MALDI-TOF MS analysis of ribosomal proteins coded in S10 and spc operons. *J. Agricult. Food Chem.* 59, 5222–5230. doi: 10.1021/jf2004095

Hsieh, T. C., Ma, K. H., and Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (hill numbers). *Methods Ecol. Evol.* 7, 1451–1456. doi: 10.1111/2041-210X.12613

Hu, E.-Z., Lan, X.-R., Liu, Z.-L., Gao, J., and Niu, D.-K. (2022). A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics* 23:110. doi: 10.1186/s12864-022-08353-7

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3:reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Kim, D., Park, S., and Chun, J. (2021). Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity. *J. Microbiol.* 59, 476–480. doi: 10.1007/s12275-021-1154-0

Konstantinidis, K. T., and Tiedje, J. M. (2005). Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* 187, 6258–6264. doi: 10.1128/JB.187.18.6258-6264.2005

Kopcakova, A., Stramova, Z., Kvasnova, S., Godany, A., Perhacova, Z., and Pristas, P. (2014). Need for database extension for reliable identification of bacteria from extreme environments using MALDI TOF mass spectrometry. *Chem. Papers* 68, 1435–1442. doi: 10.2478/s11696-014-0612-0

Kwong, J. C., McCallum, N., Sintchenko, V., and Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology* 47, 199–210. doi: 10.1097/PAT.0000000000000235

La Duc, M. T., Nicholson, W., Kern, R., and Venkateswaran, K. (2003). Microbial characterization of the Mars odyssey spacecraft and its encapsulation facility. *Environ. Microbiol.* 5, 977–985. doi: 10.1046/j.1462-2920.2003.00496.x

LaMontagne, M. G., Tran, P. L., Benavidez, A., and Morano, L. D. (2021). Development of an inexpensive matrix-assisted laser desorption—time of flight mass spectrometry method for the identification of endophytes and rhizobacteria cultured from the microbiome associated with maize. *PeerJ* 9:e11359. doi: 10.7717/peerj.11359

Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2018). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* 35, 421–432. doi: 10.1093/bioinformatics/bty648

Lasch, P., Beyer, W., Nattermann, H., Stämmler, M., Siegbrecht, E., Grunow, R., et al. (2009). Identification of *Bacillus anthracis* by using matrix-assisted laser desorption ionization-time of flight mass spectrometry and artificial neural networks. *Appl. Environ. Microbiol.* 75, 7229–7242. doi: 10.1128/AEM.00857-09

Lee, M. D. (2019). GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 35, 4162–4164. doi: 10.1093/bioinformatics/btz188

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Lévesque, S., Dufresne, P. J., Soualhine, H., Domingo, M.-C., Bekal, S., Lefebvre, B., et al. (2015). A side by side comparison of Bruker Biotyper and VITEK MS: utility of MALDI-TOF MS technology for microorganism identification in a public health reference laboratory. *PLoS One* 10:e0144878. doi: 10.1371/journal.pone.0144878

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, P.-E., Lo, C.-C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., et al. (2016). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* 45, 67–80. doi: 10.1093/nar/gkw1027

Lightfield, J., Fram, N. R., and Ely, B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One* 6:e17677. doi: 10.1371/journal.pone.0017677

Martinez-Gutierrez, C. A., and Aylward, F. O. (2022). Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLoS Genet.* 18:e1010220. doi: 10.1371/journal.pgen.1010220

Mazhari, F. (2021). Application of whole genome sequencing and MALDI-TOF to identification of *Bacillus* species isolated from cleanrooms at NASA Johnson Space Center, biology and biotechnology, University of Houston - Clear Lake, Houston, TX, pp. 63.

McCubbin, F. M., Herd, C. D. K., Yada, T., Hutzler, A., Calaway, M. J., Allton, J. H., et al. (2019). Advanced curation of astromaterials for planetary science. *Space Sci. Rev.* 215:48. doi: 10.1007/s11214-019-0615-9

Muigg, V., Cuénod, A., Purushothaman, S., Siegemund, M., Wittwer, M., Pflüger, V., et al. (2022). Diagnostic challenges within the *Bacillus cereus*-group: finding the beast without teeth. *New Microb. New Infect.* 49-50:101040. doi: 10.1016/j.nmni.2022.101040

Nithimongkolchai, N., Hinwan, Y., Kaewseekhao, B., Chareonsudjai, P., Reungsang, P., Kraiklang, R., et al. (2023). MALDI-TOF MS analysis of *Burkholderia pseudomallei* and closely related species isolated from soils and water in Khon Kaen, Thailand. *Infect. Genet. Evol.* 116:105532. doi: 10.1016/j.meegid.2023.105532

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Patel, S., and Gupta, R. S. (2020). A phylogenomic and comparative genomic framework for resolving the polyphyly of the genus *Bacillus*: proposal for six new genera of *Bacillus* species, *Peribacillus* gen. Nov., *Cytobacillus* gen. Nov., *Mesobacillus* gen. Nov., *Neobacillus* gen. Nov., *Metabacillus* gen. Nov. and *Alkalihalobacillus* gen. Nov. *Int. J. Syst. Evol. Microbiol.* 70, 406–438. doi: 10.1099/ijsem.0.003775

Popović, N. T., Kazazić, S. P., Strunjak-Perović, I., and Čož-Rakovac, R. (2017). Differentiation of environmental aquatic bacterial isolates by MALDI-TOF MS. *Env Res* 152, 7–16. doi: 10.1016/j.envres.2016.09.020

Rahi, P., Prakash, O., and Shouche, Y. S. (2016). Matrix-assisted laser desorption/ ionization time-of-fight mass-spectrometry (MALDI-TOF MS) based microbial identifications: challenges and scopes for microbial ecologists. *Front. Microbiol.* 7:1359. doi: 10.3389/fmicb.2016.01359

Regberg, A., Burton, A., McCubbin, F., Castro, C., Stahl, S., and Wallace, S. (2018) Microbial ecology of NASA curation clean rooms, 42nd COSPAR scientific assembly, Pasadena, California.

Rosselló-Mora, R. (2005). Updating prokaryotic taxonomy. *J. Bacteriol.* 187, 6255–6257. doi: 10.1128/JB.187.18.6255-6257.2005

Rudolph, W. W., Gunzer, F., Trauth, M., Bunk, B., Bigge, R., and Schröttner, P. (2019). Comparison of VITEK 2, MALDI-TOF MS, 16S rRNA gene sequencing, and whole-genome sequencing for identification of *Roseomonas mucosa*. *Microb. Pathog.* 134:103576. doi: 10.1016/j.micpath.2019.103576

Rummel, J. D. (2001). Planetary exploration in the time of astrobiology: protecting against biological contamination. *PNAS* 98, 2128–2131. doi: 10.1073/pnas.061021398

Rychert, J. (2019). Benefits and limitations of MALDI-TOF mass spectrometry for the identification of microorganisms. *J. Infectiol. Epidemiol.* 2, 1–5. doi: 10.29245/2689-9981/2019/4.1142

Seiler, H., Wenning, M., and Scherer, S. (2013). *Domibacillus robiginosus* gen. Nov., sp. nov., isolated from a pharmaceutical clean room. *Int. J. Syst. Evol. Microbiol.* 63, 2054–2061. doi: 10.1099/ijs.0.044396-0

Seng, P., Drancourt, M., Gouriet, F., La, S. B., Fournier, P. E., Rolain, J. M., et al. (2009). Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Clin. Infect. Dis.* 49, 543–551. doi: 10.1086/600885

Seuylemezian, A., Aronson, H. S., Tan, J., Lin, M., Schubert, W., and Vaishampayan, P. (2018). Development of a custom MALDI-TOF MS database for species-level identification of bacterial isolates collected from spacecraft and associated surfaces. *Front. Microbiol.* 9:e780. doi: 10.3389/fmicb.2018.00780

Severiano, A., Pinto, F. R., Ramirez, M., and Carriço, J. A. (2011). Adjusted Wallace coefficient as a measure of congruence between typing methods. *J. Clin. Microbiol.* 49, 3997–4000. doi: 10.1128/JCM.00624-11

Siegrist, T. J., Anderson, P. D., Huen, W. H., Kleinheinz, G. T., McDermott, C. M., and Sandrin, T. R. (2007). Discrimination and characterization of environmental strains of *Escherichia coli* by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS). *J. Microbiol. Methods* 68, 554–562. doi: 10.1016/j.mimet.2006.10.012

Singhal, N., Kumar, M., Kanaujia, P. K., and Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* 6:e791. doi: 10.3389/fmicb.2015.00791

Song, M., Li, Q., Liu, C., Wang, P., Qin, F., Zhang, L., et al. (2024). A comprehensive technology strategy for microbial identification and contamination investigation in the sterile drug manufacturing facility—a case study. *Front. Microbiol.* 15:1327175. doi: 10.3389/fmicb.2024.1327175

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Starostin, K. V., Demidov, E. A., Bryanskaya, A. V., Efimov, V. M., Rozanov, A. S., and Peltek, S. E. (2015). Identification of *Bacillus* strains by MALDI TOF MS using geometric approach. *Sci. Rep.* 5:16989. doi: 10.1038/srep16989

Strejcek, M., Smrhova, T., Junkova, P., and Uhlik, O. (2018). Whole-cell MALDI-TOF MS versus 16S rRNA gene analysis for identification and dereplication of recurrent bacterial isolates. *Front. Microbiol.* 9:e1294. doi: 10.3389/fmicb.2018.01294

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542. doi: 10.1093/bioinformatics/btl117

Tirumalai, M. R., Rastogi, R., Zamani, N., O'Bryant Williams, E., Allen, S., Diouf, F., et al. (2013). Candidate genes that may be responsible for the unusual resistances exhibited by *Bacillus pumilus* SAFR-032 spores. *PLoS One* 8:e66012. doi: 10.1371/journal.pone.0066012

Tirumalai, M. R., Stepanov, V. G., Wünsche, A., Montazari, S., Gonzalez, R. O., Venkateswaran, K., et al. (2018). *Bacillus safensis* FO-36b and *Bacillus pumilus* SAFR-032: a whole genome comparison of two spacecraft assembly facility isolates. *BMC Microbiol.* 18:57. doi: 10.1186/s12866-018-1191-y

Topić Popović, N., Kazazić, S. P., Bojanić, K., Strunjak-Perović, I., and Čož-Rakovac, R. (2023). Sample preparation and culture condition effects on MALDI-TOF MS identification of bacteria: a review. *Mass Spectrom. Rev.* 42, 1589–1603. doi: 10.1002/mas.21739

Werinder, A., Aspán, A., Söderlund, R., Backhans, A., Sjölund, M., Guss, B., et al. (2021). Whole-genome sequencing evaluation of MALDI-TOF MS as a species identification tool for *Streptococcus suis*. *J. Clin. Microbiol.* 59:e0129721. doi: 10.1128/JCM.01297-21

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017a). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* 3:e1099. doi: 10.1099/mgen.0.000132

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017b). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* 13:e1005595. doi: 10.1371/journal.pcbi.1005595

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag.

Wong, A., Junqueira, A. C. M., Uchida, A., Purbojati, R. W., Houghton, J. N. I., Chénard, C., et al. (2019). Complete genome sequence of *Lysinibacillus* sp. strain SGAir0095, isolated from tropical air samples collected in Singapore. *Microbiol. Resour. Announc.* 8:e00604. doi: 10.1128/mra.00604-19

Yamada, S., Ohashi, E., Agata, N., and Venkateswaran, K. (1999). Cloning and nucleotide sequence analysis of gyrB of *Bacillus cereus, B. thuringiensis, B. Mycoides,* and *B. Anthracis* and their application to the detection of *B. cereus* in rice. *Appl. Environ. Microbiol.* 65, 1483–1490. doi: 10.1128/AEM.65.4.1483-1490.1999