LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-516491

# Report for the NGFA-5 project.

C. Jaing, P. Jackson, J. Thissen, J. Wollard, S. Gardner, K. McLoughlin

November 28, 2011

## Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

# Report for the NGFA-5 project, comprehensive evaluation of the current genomic technologies including genotyping, TaqMan PCR, multiple locus variable tandem repeat analysis (MLVA), microarray and high-throughput DNA sequencing in the analysis of biothreat agents from complex environmental samples for DHS

**Contributors:**
James Thissen
Jessica Wollard
Shea Gardner
Kevin McLoughlin
Crystal Jaing
Paul Jackson
Lawrence Livermore National Laboratory (LLNL), Livermore, CA

Nadeem Bulsara, Viacheslav Fofanov and Heather Koshinski
Eureka Genomics, Hercules, CA

Sally Ellington, Loren Hauser and Tom Brettin
Oak Ridge National Laboratory, Oak Ridge, TN

**Principal Investigator and Correspondent**
Crystal Jaing
925-424-6574, jaing2@llnl.gov

Paul Jackson
(925) 424-2725, jackson80@llnl.gov

LLNL-TR-516491
**November 21, 2011**

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

# INTRODUCTION

The objective of this project is to provide DHS a comprehensive evaluation of the current genomic technologies including genotyping, TaqMan PCR, multiple locus variable tandem repeat analysis (MLVA), microarray and high-throughput DNA sequencing in the analysis of biothreat agents from complex environmental samples.

To effectively compare the sensitivity and specificity of the different genomic technologies, we used SNP TaqMan PCR, MLVA, microarray and high-throughput illumine and 454 sequencing to test various strains from *B. anthracis*, *B. thuringiensis*, BioWatch aerosol filter extracts or soil samples that were spiked with *B. anthracis*, and samples that were previously collected during DHS and EPA environmental release exercises that were known to contain *B. thuringiensis* spores. The results of all the samples against the various assays are discussed in this report.

# METHODS

*DNA extraction from BioWatch Filters*

Primary Sampling Unit (PSU) filters from the NCR Laboratory were received from the BioWatch group at LLNL. One quarter of each filter had previously been excised at NCR, so only ¾ of each filter was available. A set of filters used to collect samples over one week was obtained for each season: Spring (4/20-4/26/09), Summer (7/19-7/25/09), Fall (10/25-10/31/09), and Winter (1/22/09-1/28/09). For each day of the week, 7 to 11 "clean" filters were extracted (49 to 77 per week). Filters were determined to be "dirty" if they had an abundance of soot and dirt captured on their surface.

The ¾ PSU filters were cut into 5 roughly equal pieces. Up to 24 filters from a single week were placed into a sterile 50 mL conical tube. Thirty mL of 100 mM phosphate buffer (pH 7.4) and 0.05% (v/v) Tween 80 was added to each 50 mL tube. The conical tubes were closed, then vortexed for 30 seconds and placed on a rocking shaker for 15 additional minutes. The 30 second vortexing and 15 min shaking was repeated three additional times for a total of 1 hour of washing. The filters were then removed from each tube and the remaining solution was centrifuged at 3200 *x g* for 30 minutes at 5°C to collect material that had washed from the filters. Following centrifugation, the supernatant was removed and discarded.

To complete the DNA extraction and purification, components of the UltraClean Soil DNA Isolation Kit #12800 from MoBio (Carlsbad, CA) were used. The remaining pellet in each tube was suspended in the following solutions added in this order: 100 µL TE buffer, 350 µL MoBio Bead Solution, 60 µL MoBio Solution S1, and 200 µL MoBio Inhibitor Removal Solution. A 2 mL screw cap tube was loaded with 500 mg each of 106 and 500 mm zirconia/silica beads. The entire 700 µL of each suspended pellet was added to a 2 mL bead tube. The samples were bead-beaten at max speed for 2 minutes. Following this, the tubes were centrifuged at 10,000 *x g* for 30 seconds. The entire supernatant (~450 µL) was transferred to a fresh, sterile 2 mL tube for further extraction.

Two hundred fifty µL of MoBio Solution S2 was added to the supernatant in each tube, vortexed for 5 seconds, and incubated at 4°C for 5 minutes. The tubes were then centrifuged for 1 minute at 10,000 *x g* and the supernatant transferred to a clean 2 mL tube. Two volumes (~1.3 mL) of MoBio Solution S3 was added to each tube and vortexed for 5 seconds. The vortexed solution was added in 700µL aliquots to a MoBio spin filter until the entire sample was loaded, and the column was centrifuged for 1 minute at 10,000 *x g.* The column flow-through was discarded and the spin filter was washed 3 times by adding 300 µL MoBio Solution S4 and centrifuging for 30 seconds at 10,000 *x g*. The flow-through was discarded following each spin. The spin filter was centrifuged an additional 1 minute at 10,000 *x g* to dry the filter. It was then placed in a new 2 mL collection tube and 50 µL of MoBio Solution S5 was added to the membrane. The sample was centrifuged at 10,000 *x g* for 30 seconds and the eluted DNA was retained. The multiple elutions

for each season were combined into one large volume. Samples were subjected to vacuum in a Speed Vac vacuum centrifugation system to reduce the volume to ~50% of the starting volume, increasing the DNA concentration by a factor of ~2. The DNA concentration was determined using Picogreen dsDNA fluorescent stain and the resulting fluorescence was measured using an Invitrogen Qubit fluorometer (Carlsbad, CA).

*DNA Extraction from Soil*

Soil was collected in the downtown areas of Oakland and San Francisco, CA. Four samples were collected in each city at different sites. Samples were extracted using the MoBio Ultraclean Soil DNA Isolation Kit #12800. The manufacturer's Alternative Protocol (designed to provide maximum yields) was followed for this work. The only deviation from the protocol was to wash twice (Step 15) with Solution S4 instead of just once as the protocol suggested.

Following extraction, 1 ng of each extracted DNA was used as template in a Real-Time PCR assay to test for inhibition. All samples showed a high degree of PCR inhibition (data not shown). Based on these results, each sample was re-extracted starting from Step 12 of the MoBio Alternative Protocol. This additional extraction was intended to remove additional humic acid. The DNA concentration of each sample was determined as described above.

*DNA Extraction from EPA gauze wipes*

Gauze wipes were obtained from a group at LLNL that conducted tests with the EPA. Wipes were used to wipe dirty surfaces indoors and were then inoculated with $6.3 \times 10^7$ CFU of *B. thuringiensis* kurstaki spores. Following inoculation, DNA was extracted from the wipes using the Promega Blood Extraction Kit (Madison, WI) according to manufacturer's instructions. After extraction, $5.0 \times 10^4$ CFU were amplified and labeled for microarray assay.

*Addition of Bacillus anthracis Ames DNA to environmental samples*

*B. anthracis* Ames DNA was acquired from the LLNL select agent laboratory. Sample sterility was confirmed by plating 1/10 volume of the Ames DNA sample on blood agar plates and incubating for 48 hours at 37°C. No colonies were found after this incubation indicating the absence of viable *B. anthracis* Ames cells in the DNA preparation. The DNA concentration was determined using Picogreen as described above. Six solutions containing different concentrations of *B. anthracis* Ames DNA were prepared by 10-fold serial dilution to produce solutions containing ~1 to 100,000 copies of the Ames genome. Each concentration was mixed with 100 pg of DNA extracted from the Spring NCR filters or 1 ng of DNA extracted from the combination of Oakland and San Francisco soils.

*Whole Genome Amplification and Purification*

The environmental *B. anthracis* Ames spiked samples were amplified using the Qiagen REPLI-g Midi Kit #150043 (Valencia, CA). This kit is intended to provide uniform whole genome amplification using Multiple Displacement Amplification. Each copy number dilution of *B. anthracis* DNA spiked in either 1ng of soil or 100 pg of aerosol DNA was amplified using this kit according to manufacturer's instructions. Samples were allowed to amplify for 16 hours at 30°C. Amplified samples were purified using the Qiagen Qiaquick PCR Purification Columns #28106 according to manufacturer's instructions. Samples were eluted in 40μL of Buffer EB from the Qiagen kit.

*Bacillus anthracis Sterne ciprofloxacin selections*

A parental culture of avirulent *Bacillus anthracis* Sterne was streaked onto a nutrient broth agar plate. The wild-type ciprofloxacin minimum inhibitory concentration (MIC) value was determined for *B. anthracis* Sterne by picking a single colony to inoculate 5 mL nutrient broth and incubating overnight at 35°C, 160 rpm. A subculture containing 5 mL nutrient broth was inoculated with 200 μL of the overnight culture and incubated at 35°C, 160 rpm to an optical density at 600 nm of 0.8. A ciprofloxacin Etest (AB Biodisk) was applied to a nutrient broth agar plate swabbed for full coverage with the *B. anthracis* Sterne subculture, and the Etest plate was incubated overnight at 35°C. An approximate ciprofloxacin MIC was determined to be 0.047 μg/mL for the wild-type *B. anthracis* Sterne.

Cultures were prepared for first-round selections by inoculating 20 tubes containing 5 mL nutrient broth, each with a single *B. anthracis* Sterne colony. The tubes were incubated horizontally at 35°C, 160 rpm, overnight. Fresh subcultures were prepared by adding 500 μL of each overnight culture to 12 mL nutrient broth. The subcultures were incubated at 35°C, 160 rpm to an optical density at 600 nm of 0.8. The cells were concentrated by centrifugation at 4000*xg* for 10 min. Approximately 11.5 mL of the supernatant was discarded and each cell pellet was suspended in the remaining 1 mL nutrient broth. Each of the twenty 1 mL suspensions were plated on a nutrient broth agar plate containing 0.094 μg/mL ciprofloxacin (three times the wild-type MIC value). These 20 first-round selection plates were incubated at 35°C, up to 72 hours. Each of the small number of ciprofloxacin resistant colonies was picked into 5 mL nutrient broth containing 0.094 μg/mL ciprofloxacin (three times the wild-type MIC value) and incubated at 35°C, 160 rpm up to 72 hours. Subcultures were prepared from any passage cultures that grew in the presence of ciprofloxacin by adding 200 μL of the passage culture to 5 mL nutrient broth without ciprofloxacin and incubating at 35°C, 160 rpm to an optical density at 600 nm of 0.8. These subcultures were used for MIC value determinations (as indicated above) and for frozen stocks by adding 775 μL of the subculture to 225 μL sterile 80% glycerol followed by storage at -80°C.

Second and third round selections were performed by repeating this process to obtain mutants resistant to ciprofloxacin concentrations well above therapeutic levels. Second round selections of the first-round mutants were carried out by

increasing ciprofloxacin concentrations to approximately three-fold the parent generation MIC values at each step. Approximately 10 second-round mutants were carried on for each of 20 first-round mutants, and up to 5 third-round mutants were saved for each successful second-round mutant.

Mutant colonies were verified to be *B. anthracis* Sterne by colony morphology and species-specific PCR. Heat soak lysates were used as PCR templates. Heat soaks were prepared by transferring a single mutant colony into 200 μL filter-sterilized 1X TE and incubating at 95°C for 20 minutes. The sample was cooled to room temperature and centrifuged at 10,000*xg* for 1 minute. The supernatant was transferred to a new tube and stored at -20°C. A summary of the selection process is shown in Figure 1.

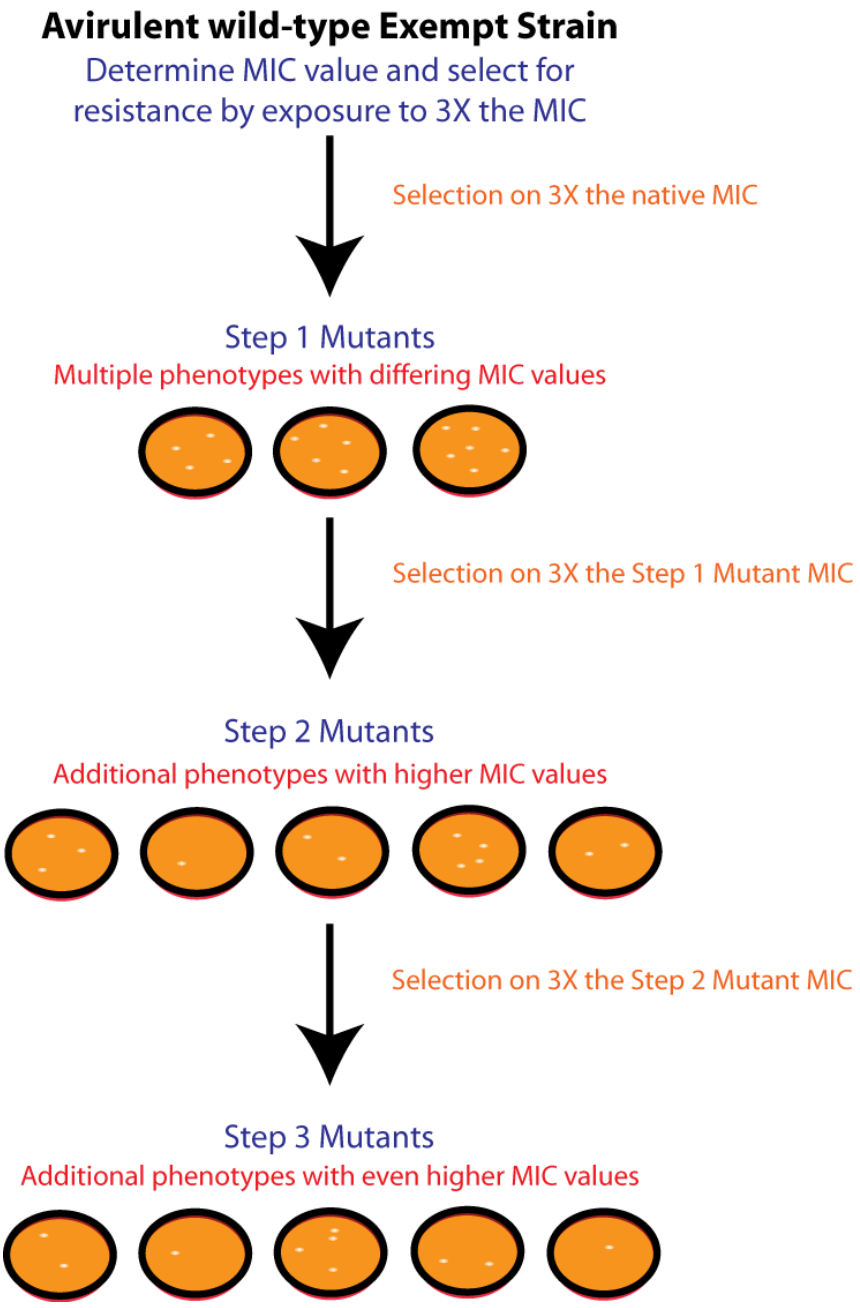



Figure 1. *Bacillus anthracis* ciprofloxacin resistance selection process

*Genomic DNA preparations of Ciprofloxacin resistant Bacillus anthracis isolates, PCR and sequence verification*

Genomic DNA from different isolates was isolated using the Epicentre Masterpure Gram Positive DNA kit. PCR oligonucleotide primers that amplified genome-specific sequences from the different microbial species were designed using Primer3™. PCR used Promega reagents. Sequence verification of a targeted

PCR region, when required, was performed by primer walking off the PCR products using 1/8 of a Big Dye V3.1 sequencing kit. Sanger™ sequencing was performed using ABI3730 and ABI3100 DNA analyzers at the DOE Joint Genome Institute in Walnut Creek, CA and at Lawrence Livermore National Laboratory in Livermore, CA.

*TaqMan Canonical SNP Assay*

**Canonical SNP TaqMan assay design** .We designed 13 *B. anthracis* canonical SNP TaqMan assays to detect various strains of *B. anthracis*. The canonical SNPs were designed based on the phylogenetic analysis conducted by Van Ert *et al*. (1) (Figure 2). A total of 13 branch points were identified for strain level detection of *B. anthracis* strains. Two probes were designed for each SNP; one allele labeled with FAM dye and the other allele was labeled with VIC dye. The expected results for the SNP TaqMan assay are also shown in this table. The strain A0382 is not a fully sequenced strain. We were able to make some predictions based on a *B. anthracis* SNP microarray we have run on this strain. The TaqMan primer and probes were ordered through Applied Biosystems Inc (ABI). The canonical SNP assay ID, their corresponding branch points and the Ames and Sterne genome locations are listed in Table 1. The primer and probe sequences are listed in Table 2.

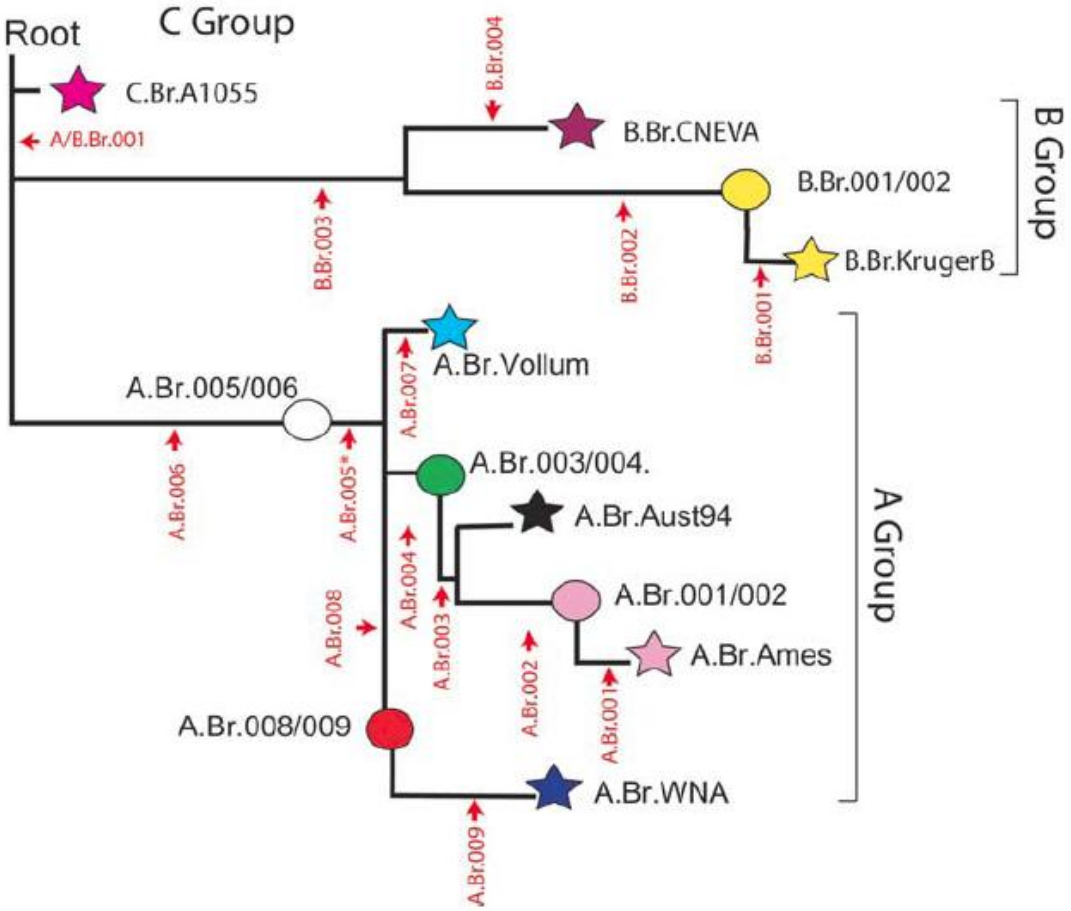



Figure 2. The relationship between canSNPs, sub-lineages and/or sub-groups: The stars in this dendrogram represent specific lineages that are defined by one of the seven sequenced genomes of *B. anthracis*. The circles represent branch points along the lineages that contain specific subgroups of isolates. These sub-groups are named after the canSNPs that flank these positions. Indicated in red are the positions and names for each of the canSNPs. (From Van Ert *et al*. 2007, PLoS ONE).

Table 1. Canonical SNP assay ID, corresponding branch points and the genomics locations on B*. anthracis* Ames and *B. anthracis* Sterne

| Assay ID | Branch point | Ames genome position | Sterne genome position | Ames Refseq | Ames expected | Sterne Refseq | Sterne expected | A0382 Array | A0382 expected |
|---|---|---|---|---|---|---|---|---|---|
| canSNP1 | A.Br.001 | 182106 | 182107 | C | VIC | T | FAM | T | FAM |
| canSNP2 | A.Br.002 | 947759 | 947654 | T | VIC | T | VIC | C | FAM |
| canSNP3 | A.Br.003 | 1493157 | 1493231 | C | VIC | C | VIC | T | FAM |
| canSNP4 | A.Br.004 | 3600659 | 3601360 | C | VIC | C | VIC | NA | |
| canSNP5 | A.Br.006 | 162509 | 162510 | A | VIC | A | VIC | C | FAM |
| canSNP6 | A.Br.007 | 266439 | 266452 | A | FAM | A | FAM | NA | |
| canSNP7 | A.Br.008 | 3947248 | 3947747 | A | FAM | A | FAM | NA | |
| canSNP8 | A.Br.009 | 2589823 | 2590283 | A | FAM | A | FAM | A | FAM |
| canSNP9 | B.Br.001 | 1455279 | 1455347 | A | FAM | A | FAM | A | FAM |
| canSNP10 | B.Br.002 | 1056740 | 1056633 | C | FAM | C | FAM | A | VIC |
| canSNP11 | B.Br.003 | 1494269 | 1494343 | G | FAM | G | FAM | A | VIC |
| canSNP12 | B.Br.004 | 69952 | 69953 | T | FAM | T | FAM | T | FAM |
| canSNP13 | A/B.Br.001 | 3697886 | 3698581 | A | FAM | A | FAM | A | FAM |

Table 2. Canonical SNP primer and probe sequences

| Assay# | Forward Primer | Reverse Primer | Reporter 1 (VIC) | Reporter 2 (FAM) |
|---|---|---|---|---|
| canSNP1 | GGCAAGCGGAACCAAATTTAATCTT | TACGTCATTGTATAATACGGTTTCCCTTT | ATCGACTTCAA**G**TTTCGGT | TCGACTTCAAA**T**TTCGGT |
| canSNP2 | GAGGCAGAAGGAGCAAGTAATGTTA | ACCATAACTGATCCAACGATACCTAAAATC | CCGCCCA**A**CCTAAA | CGCCCA**G**CCTAA |
| canSNP3 | GCTTGCCAAGCTTTTTTTCTATTAT ATATAAAAGGAA | GTAGCTACTGTCATTGTATAAAAACCTCCTT | TTTCTACCTCAA**G**CTTAATT | TTTCTACCTCAA**A**CTTAATT |
| canSNP4 | CCGATACCAGTAAACGACGACATC | CTGGAATTGGTGGAGCTATGGAA | TTGGAATG**C**CCCTAATC | TTTGGAATG**T**CCCTAATC |
| canSNP5 | TTCAAAAAATTCTTTGATCAATATG TTGTTGATCATTC | CTTCCTCATCCCAATCTAGCGTTTT | CATCGCCT**A**GTGCATG | ATCGCCT**C**GTGCATG |
| canSNP6 | GGCGATTGCGAAAAGTATTGTTGAA | TTGGTAACGAGACGATAAACTGAATAATACC | CGAGCTGAAT**G**TAAGGAT | TCGAGCTGAAT**A**TAAGGAT |
| canSNP7 | GGATGCAAATAAACCAAACGGTGAA AA | CATTCGCAACTACGCTATACGTTTT | AATTCTTC**G**CCGCTTGT | AGATAATTCTTC**T**CCGCTTGT |
| canSNP8 | GGCAATCGGCCACTGTTTT | CTACTGTGTATGTTGTTAATAAAAAGTATGAATTTTTAGGT | ACGGCTTT**G**CTTGCAT | AACGGCTTT**A**CTTGCAT |
| canSNP9 | GGGAGAAGTTATTTGCACGGTCATA | TTCAAAAGGTTCGGATATGATACCGATAC | CGGTACAATAGAAGA**G**GATAA | CGGTACAATAGAAGA**A**GATAA |
| canSNP10 | CCGAATGGAGGAGAAGTTGCA | TGCACCTTCTGTGTTCGTTGTTAA | AAAGGAACAG**A**AGTAACG | AGGAACAG**C**AGTAACG |
| canSNP11 | TCGCATAGAAGCAGATGAGCTTAC | TGTGCCATCAAATAACTCTTTCTCAAGT | CATAACGTGAAG**T**GGATAT | AACGTGAAG**C**GGATAT |
| canSNP12 | ACAAGTGCTTGGGTAACCTTCTTT | GCCTTGAGCTTGGTTTAATAAGAAGAAGAA | AACGGGATG**G**TAGAAGT | ACGGGATG**A**TAGAAGT |
| canSNP13 | ACCAGTTATTCCAATCGCTGCA | ACCTTTCGGTAAATAGTCCCCGATA | CTCTTTTATTTAG**G**AGATAGC | CTCTTTTATTTAG**A**AGATAGC |

**SNP assay protocol.** The SNP TaqMan assays were carried out in 10 µL reactions. Each reaction consisted of 2X TaqMan ABI Universal PCR Master Mix (#4304437), 40X Assay (ABI Custom TaqMan SNP Genotyping Assay), and PCR grade water. All reactions were run on an ABI 7900 HT Fast Real-Time PCR System with the following parameters: 1 cycle of 50°C for 2 min, 1 cycle of 95°C for 10 min, 40 cycles of 95°C for 15 sec and 60°C* for 1 min (*Assay #1 only has an annealing temperature of 63°C).

*TaqMan B. thuringiensis SNP Assay*

**TaqMan signature design.** Signature candidates were designed with KPATH (2). Amplicons were 80-250 bp, primer $T_m$'s approximately 60-65◦C, primer lengths 18-26 bases, probes 18-36 bases, and probe $T_m$'s approximately 68-73◦C. The TaqMan signatures specific to *B. thuringiensis* kurstaki or israelensis are shown in Tables 3 and 4. TaqSim software (http://staff.vbi.vt.edu/dyermd/publications/files/TaqSim_Help.pdf) was used to predict which TaqMan triplets (primers with a probe) signatures should detect which targets, assuming no mismatches of primers and probes to targets. The TaqSim results are shown in table 5.

Table 3. *Bacillus thuringiensis* kurstaki assays

| Signature | Forward Primer | Reverse Primer | Probe |
|---|---|---|---|
| 2254619 | CGGTTATATTCTTCTGGGTGTCG | CTCCCAACCTTGGTTTCTGC | CCTAACTTTGACGAGATAAATGGGCCAGCAT |
| 2254620 | TGACATCCTCCCAGAATGTTATAGA | CCCCTGAAGGAGGACTGATG | ATGGGAGGTCTTATTCCATCCACACTGCATA |
| 2254621 | TGCGACATCTGTAAAGTTTAGATCG | CGCTTAGGAAACAACTGCCC | CCCCTTCTCCAAACATAAATCCTCCATCCTA |
| 2254622 | ACTCACCGATGATTCGAACG | ATGTACGCTGCGAGCAAGTC | TGTTATTTCCCCTATTCCTGGAAATGACCGA |
| 2254623 | AGCGTATGCTCGTCTCAAGTAAAA | CCTGCCTTGTGGATCTCTAGC | TGCATCGAACTCAATAAAATATTTGTTTTGGAGGG |

Table 4. *Bacillus thuringiensis* israelensis assays

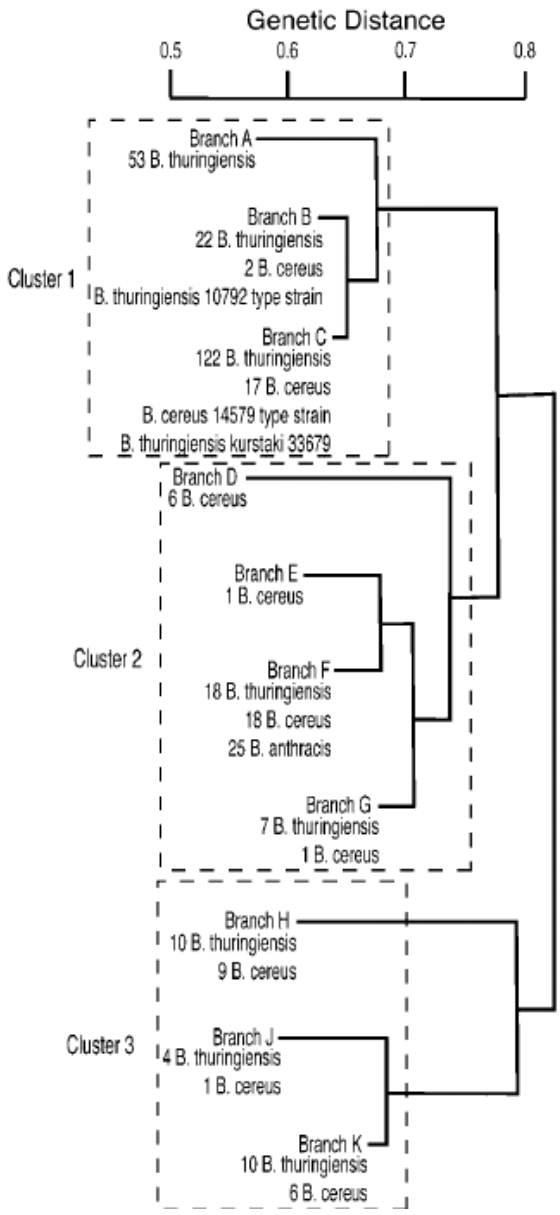
| Signature | Forward Primer | Reverse Primer | Probe |
|---|---|---|---|
| 2253467 | CCCTTCTAAACTACGCGGTGA | TCGTCCGTACATATCTATTCGTCG | TCATCGTCATAACCCGTGTATTGACCAACAC |
| 2253469 | GTTGGAATCGAACCACCCTT | GCTTATTGGGCCCTATGTATTCTG | TGGACTATGACCTCCTGCTCTGATACAGTGG |
| 2253470 | CGCTCACGATATGTTCTAATACCAA | CTTGCGTACATGCTCGCTG | TCTCTCTAACGCCTTCATAGCGCGCC |

Table 5: TaqSim prediction results of *B. thuringiensis* kurstaki and israelensis signatures

| Sig ID/Name (F/IO/R) | Verification/Cross Rxn | Start | Stop | Amp Size |
|---|---|---|---|---|
| sig_candidate_2254619 | kp\|1741242\|704947\|Glued fragments of sequence 704732 (Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, unfinished sequence, whole genome shotgun sequencing project fr...) - 337 fragments\|kpath_id\|1741242 Glued fragments of sequence 704732 (Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, unfinished sequence, whole genome shotgun sequencing project fr...) - 337 fragments | 309878 | 310050 | 173 |
| sig_candidate_2254619 | kp\|1575409\|692450\|Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments\|kpath_id\|1575409 Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments | 21062 | 21234 | 173 |
| 2 total amplicons | | | | |
| sig_candidate_2254620 | kp\|1741242\|704947\|Glued fragments of sequence 704732 (Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, unfinished sequence, whole genome shotgun sequencing project fr...) - 337 fragments\|kpath_id\|1741242 Glued fragments of sequence 704732 (Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, unfinished sequence, whole genome shotgun sequencing project fr...) - 337 fragments | 310794 | 310982 | 189 |
| sig_candidate_2254620 | kp\|1575409\|692450\|Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments\|kpath_id\|1575409 Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments | 21978 | 22166 | 189 |
| 2 total amplicons | | | | |
| sig_candidate_2254621 | kp\|1575409\|692450\|Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments\|kpath_id\|1575409 Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments | 211400 | 211581 | 182 |
| sig_candidate_2254621 | Bacillus thuringiensis serovar kurstaki str. T03a001 - gi\|238801503\|ref\|NZ_CM000751.1\|gnl\|REF_WGS:ACND\|Chr1 - draft sequence\|Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, whole genome shotgun sequencing project from NCBI on Jun 17 2010 6:15PM\|kpath_id\|2540154 | 5513479 | 5513660 | 182 |
| 2 total amplicons | | | | |
| sig_candidate_2254622 | kp\|1575409\|692450\|Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments\|kpath_id\|1575409 Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments | 615057 | 615230 | 174 |
| sig_candidate_2254622 | Bacillus thuringiensis serovar kurstaki str. T03a001 - gi\|238801503\|ref\|NZ_CM000751.1\|gnl\|REF_WGS:ACND\|Chr1 - draft sequence\|Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, whole genome shotgun sequencing project from NCBI on Jun 17 2010 6:15PM\|kpath_id\|2540154 | 5002369 | 5002542 | 174 |
| 2 total amplicons | | | | |
| sig_candidate_2254623 | kp\|1575409\|692450\|Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments\|kpath_id\|1575409 Glued fragments of sequence 692218 (Bacillus thuringiensis serovar kurstaki draft sequence from Microgen on Jan 13 2009 10:42AM) - 729 fragments | 5630308 | 5630456 | 149 |
| sig_candidate_2254623 | Bacillus thuringiensis serovar kurstaki str. T03a001 - gi\|238801503\|ref\|NZ_CM000751.1\|gnl\|REF_WGS:ACND\|Chr1 - draft sequence\|Bacillus thuringiensis serovar kurstaki str. T03a001 Bacillus thuringiensis serovar kurstaki str. T03a001, whole genome shotgun sequencing project from NCBI on Jun 17 2010 6:15PM\|kpath_id\|2540154 | 5222465 | 5222613 | 149 |
| 2 total amplicons | | | | |
| sig_candidate_2253467 | kp\|996745\|632731\|Glued fragments of sequence 630766 (Bacillus thuringiensis serovar israelensis ATCC 35646 Bacillus thuringiensis serovar israelensis ATCC 35646, unfinished sequence, whole genome shotgun sequencing project ...) - 866 fragments\|kpath_id\|996745 Glued fragments of sequence 630766 (Bacillus thuringiensis serovar israelensis ATCC 35646 Bacillus thuringiensis serovar israelensis ATCC 35646, unfinished sequence, whole genome shotgun sequencing project ...) - 866 fragments | 4285750 | 4285917 | 168 |
| sig_candidate_2253467 | kp\|1746320\|705178\|Glued fragments of sequence 705159 (Bacillus thuringiensis IBL4222 Bacillus thuringiensis IBL 4222, unfinished sequence, whole genome shotgun sequencing project from NCBI on May 31 2009 09:46AM) - 383 fragments\|kpath_id\|1746320 Glued fragments of sequence 705159 (Bacillus thuringiensis IBL4222 Bacillus thuringiensis IBL 4222, unfinished sequence, whole genome shotgun sequencing project from NCBI on May 31 2009 09:46AM) - 383 fragments | 3723087 | 3723254 | 168 |
| 2 total amplicons | | | | |
| sig_candidate_2253469 | kp\|996745\|632731\|Glued fragments of sequence 630766 (Bacillus thuringiensis serovar israelensis ATCC 35646 Bacillus thuringiensis serovar israelensis ATCC 35646, unfinished sequence, whole genome shotgun sequencing project ...) - 866 fragments\|kpath_id\|996745 Glued fragments of sequence 630766 (Bacillus thuringiensis serovar israelensis ATCC 35646 Bacillus thuringiensis serovar israelensis ATCC 35646, unfinished sequence, whole genome shotgun sequencing project ...) - 866 fragments | 5016674 | 5016809 | 136 |
| sig_candidate_2253469 | kp\|1746320\|705178\|Glued fragments of sequence 705159 (Bacillus thuringiensis IBL4222 Bacillus thuringiensis IBL 4222, unfinished sequence, whole genome shotgun sequencing project from NCBI on May 31 2009 09:46AM) - 383 fragments\|kpath_id\|1746320 Glued fragments of sequence 705159 (Bacillus thuringiensis IBL4222 Bacillus thuringiensis IBL 4222, unfinished sequence, whole genome shotgun sequencing project from NCBI on May 31 2009 09:46AM) - 383 fragments | 3757890 | 3758025 | 136 |
| 2 total amplicons | | | | |

| | | | | |
|---|---|---|---|---|
| sig_candidate_2253470 | kp\|996745\|632731\|Glued fragments of sequence 630766 (Bacillus thuringiensis serovar israelensis ATCC 35646 Bacillus thuringiensis serovar israelensis ATCC 35646, unfinished sequence, whole genome shotgun sequencing project ...) - 866 fragments\|kpath_id\|996745 Glued fragments of sequence 630766 (Bacillus thuringiensis serovar israelensis ATCC 35646 Bacillus thuringiensis serovar israelensis ATCC 35646, unfinished sequence, whole genome shotgun sequencing project ...) - 866 fragments | 5270719 | 5270846 | 128 |
| sig_candidate_2253470 | kp\|1746320\|705178\|Glued fragments of sequence 705159 (Bacillus thuringiensis IBL4222 Bacillus thuringiensis IBL 4222, unfinished sequence, whole genome shotgun sequencing project from NCBI on May 31 2009 09:46AM) - 383 fragments\|kpath_id\|1746320 Glued fragments of sequence 705159 (Bacillus thuringiensis IBL4222 Bacillus thuringiensis IBL 4222, unfinished sequence, whole genome shotgun sequencing project from NCBI on May 31 2009 09:46AM) - 383 fragments | 3921804 | 3921931 | 128 |
| 2 total amplicons | | | | |

**Genomic DNA isolation from various Bacillus strains.** Various *B. thuringiensis* (Bt), *B. anthracis* (Ba)and *B. cereus* (Bc) strains were selected to cover a wide breadth of genetic diversity of the *Bacillus* genus and with reference to the *Bacillus* phylogenetic tree published by Hill *et al*. (3) (Figure 3). The tree was generated using Amplified Fragment Length Polymorphism (AFLP). The list of *Bacillus* strains used for TaqMan signature testing includes: Bt kurstaki ATCC 33679, Bt kurstaki HD-1, Bt AH547, Bt finitimus HD527, Bt israelensis HD500, Bt sotto. HD774, Bt AH535, Bt AH575, Ba Sterne, Ba Ames, Bt HD95, Bt AH592, Bc ATCC 4342, Bt pakistani HD462, Bt AH678, Bc D21, Bt konkukian 97-27, Bt HD18, Bt pondic HD1101 and Ba A0382.




Hill et al., 2004

Figure 3. AFLP generated phylogenetic tree for *Bacillus thuringiensis* and *Bacillus cereus*

**TaqMan PCR assay protocol.** Each 20µL reaction consisted of 2X TaqMan ABI Universal PCR Master Mix (#4304437), 0.5µM each primer, 0.25µM probe, and PCR grade water. All reactions were run on an ABI 7500 Fast Real-Time PCR System with the following parameters: 1 cycle of 95°C for 20 sec, 40 cycles of 95°C for 3 sec and 60°C* for 30 sec. Each template was tested in duplicate at 1ng per reaction.

*The MLVA Protocol*

**Primers.** The eight primer pairs used to amplify DNA from the different MLVA loci were described by Keim *et al*., 2000. Reverse primers pX01 and pXO2 were modified by adding the sequence GTGTCTT to their 5' ends. This addition encourages complete adenylation of the PCR product (4) to help prevent prominent stutter peaks in the fragment analysis data. Such peaks can be problematic when analyzing MLVA amplicons that may differ in length by as little as two nucleotides. The set of the eight forward MLVA primers were synthesized by Applied Biosystems in Palo Alto, California. Each forward primer contained a 5' fluorescently labeled 6-FAM dye. The reverse primers were unlabeled. Following PCR amplification, the resulting MLVA amplicons contained a fluorescent label that could be analyzed on a DNA sequencer to produce a more precise measurement of their length than can be produced by analysis by agarose gel electrophoresis. Table 6 shows the sequences of the eight different MLVA primers used.

**MLVA PCR.** The 8 MLVA PCR amplification primer sets were separated into different reactions based on product size and compatibility of reaction mixes – different reactions amplify best under slightly different conditions. The reaction 1 group (*vrrA, vrrB2, vrrC2* and CG3) each contained PCR buffer supplied with Clontech Advantage 2 polymerase (40 mM Tricine-KOH, 15 mM KOAc, 3.5 mM $Mg(OAc)_2$, 0.75 µg/mL BSA, 0.0005% (v/v) Tween, 0.0005% (v/v) Nonidet-P400); 0.2 mM of each dNTP (Clontech, Mountain View, CA); sample DNA (see below) ; 0.05 µM each of a 5' labeled forward and an unlabeled reverse primer; and 0.02 µl/µl of Advantage 2 polymerase (Clontech, cat# 639201) per PCR reaction. The reactions were heated for 5 minutes at 95ºC. This was followed by incubation through 35 cycles of 95ºC for 1 minute, 50ºC for 30 seconds, and 68ºC for 1 minute. The reactions were then incubated at 68ºC for an additional 12 minutes to facilitate final extension of the PCR products.

The reaction 2 group (*vrrB1* and *vrrC1*) each contained PCR buffer supplied with Clontech Advantage 2 polymerase (see above); 0.2 mM of each dNTP (Clontech); sample DNA; 0.08 µM each of a 5' labeled forward and an unlabeled reverse primer; and 0.02 µl/µl of Advantage 2 polymerase (Clontech) per reaction. Each reaction was heated for five minutes at 95°C. This was followed by incubation through 35 cycles of 95ºC for 20 seconds, 60ºC for 20 seconds, and 68ºC for 20 seconds. Final extension of the PCR products was facilitated by incubation at 68ºC for 12 minutes.

The reaction 3 group (pXO1-att and pXO2-at) each contained PCR buffer supplied with Clontech Advantage 2 polymerase; 0.2 mM of each dNTP (Clontech); sample DNA; 0.1 µM each of a 5' labeled forward and an unlabeled reverse primer;

and 0.02 µl per µl of Advantage 2 polymerase (Clontech) per reaction. Reactions were heated for 5 minutes at 95ºC. This was followed by incubation through 35 cycles of 95ºC for 30 seconds, 50ºC for 30 seconds, and 68ºC for 30 seconds. Final extension of the PCR products was facilitated by incubation at 68ºC for 12 minutes.

Table 6. MLVA Primer Sequences

| Locus | Forward Primer | Reverse Primer |
|---|---|---|
| *vrrA* | CAC AAC TAC CAC CGA TGG CAC A | GCG CGT TTC GTT TGA TTC ATA C |
| *vrrB1* | ATA GGT GGT TTT CCG CAA GTT ATT C | GAT GAG TTT GAT AAA GAA TAG CCT GTG |
| *vrrB2* | CAC AGG CTA TTC TTT ATC AAA CTC ATC | CCC AAG GTG AAG ATT GTT GTT GA |
| *vrrC1* | GAA GCA AGA AAG TGA TGT AGT GGA C | CAT TTC CTC AAG TGC TAC AGG TTC |
| *vrrC2* | CCA GAA GAA GTG GAA CCT GTA GCA C | GTC TTT CCA TTA ATC GCG CTC TAT C |
| CG3 | TGT CGT TTT ACT TCT CTC TCC AAT AC | AGT CAT TGT TCT GTA TAA AGG GCA T |
| pXO1-att | CAA TTT ATT AAC GAT CAG ATT AAG TTC A | GTG TCT TTC TAG AAT TAG TTG CTT CAT AAT G GC |
| pXO2-at | TCA TCC TCT TTT AAG TCT TGG GT | GTG TCT TGT CTG ATG AAC TCC GAC GAC A |

Table 7. VNTR marker attributes. Note that VNTR loci found within open reading frames are shown in *italics*

| | | Array size (No. of repeats) | | |
|---|---|---|---|---|
| Locus | Repeat size (bp) | Smallest | Largest | No. of alleles |
| *vrrA* | 12 | 2 | 6 | 5 |
| *vrrB1* | 9 | 15 | 23 | 5 |
| *vrrB2* | 9 | 11 | 15 | 3 |
| *vrrC1* | 36 | 4 | 12 | 6 |
| *vrrC2* | 18 | 17 | 19 | 3 |
| CG3 | 5 | 1 | 2 | 2 |
| pXO1-att | 3 | 4 | 11 | 8 |
| pXO2-at | 2 | 6 | 15 | 9 |

**MLVA fragment analysis.** For each reaction, 11 µl of Genetic Analysis grade Hi Di™ formamide (Applied Biosystems, Foster City, CA) along with 0.5 µl of GeneScan™ 600 LIZ size standard (Applied Biosystems) were added to 1 µl of each MLVA reaction. The fragment/formamide/size standard mix was heated at 95ºC for 5 minutes to denature the DNA and immediately cooled on ice for 2-5 minutes before loading onto an ABI 3100 DNA sequencer for fragment analysis. The products were analyzed using a 36 cm gel capillary (Applied Biosystems) with POP-4™ polymer (Applied Biosystems). Data Collection Software v1.1 was used to collect

the fragment profiles. The run module was set at a voltage of 15, an injection time of 22, and a run time of 2200. The files were analyzed with Gene Scan v3.7 software with Large Fragment Analysis enabled.

*Microarrays*

**Sequence analysis and microarray probe design to develop the Virulence Array.** Probe design for virulence and antibiotic resistance gene families was performed by selecting target sequences from the genomes and searching for virulence-related proteins using 712 sets of profile hidden Markov models (HMMs). HMM sets were selected to recognize a collection of several hundred virulence-associated protein families identified from the literature and public databases. There are a total of 574 virulence families present in the 8 bacterial and 8 viral agents (Table 8), totaling 41,535 gene sequences. We selected probes so that each target gene sequence would be covered by at least 13 probes, favoring probes that were conserved among the sequences within that gene family. Our original HMMs did not represent genes in VEE or West Nile viruses. For these, we downloaded the 27 profile HMMs in the PFAM database for Togaviridae (VEE) and Flaviviridae (West Nile virus) and searched all available complete and partial Togaviridae and Flaviviridae sequences, respectively, resulting in an additional 34,082 gene sequences, for which we designed probes that provided coverage of at least 13 probes per target sequence. The algorithms used for probe design were described as in (5). The total number of probes designed for each of the categories is listed in Table 9.

Table 8. Bacterial and viral agents included on the Virulence Array

| **Bacterial Agents** | **Viral Agents** |
|---|---|
| *Bacillus anthracis* | Marburg virus |
| *Yersinia pestis* | Ebola virus (Reston, Zaire, Sudan) |
| *Francisella tularensis* | Variola virus |
| *Burkholderia mallei* | Foot-and-Mouth Disease (FMDV) virus |
| *Burkholderia pseudomallei* | Venezuelan Equine Encephalitis (VEEV) virus |
| *Brucella abortus* | Crimean Congo Hemorrhagic Fever (CCHF) virus |
| *Brucella melintensis* | West Nile virus |
| *Brucella suis* | Rift Valley Fever (RVF) virus |

Table 9. Types of microarray probes for the Virulence Array

| Probe Type | # of Probes |
|---|---|
| Bacterial virulence and A/R gene probes | 83,372 |
| Bacterial species level probes | 842 |
| Bacterial forensic level probes | 1,260 |
| Viral virulence and A/R gene probes | 40,230 |
| Viral forensic level probes | 1,677 |
| BW/LRN Amplicon probes | 502 |
| Vector probes | 35,791 |
| FT/FP discriminating gene probes | 21,890 |
| Random control probes | 2,898 |
| **Total # of probes** | **188,462** |

**Probe design for BioWatch/LRN amplicons.** We designed microarray probes that span the PCR amplicons from the current BioWatch and LRN signatures. These probes will serve as a secondary confirmation when there are any near positive BioWatch events.

**Probe design for bacterial vectors.** We have developed a database with 3,800 complete and partially sequenced vectors and designed microarray probes from unique regions of each vector sequences, using a target goal of 12 probes per vector. Probes are chosen in decreasing order of conservation across the vector sequence database. The initial candidate probe set is screened *in silico* against all sequenced viral and bacterial genomes including naturally occurring plasmids. Candidate vector probes with a similarity above a fixed threshold were removed from the probe set. A cross validation procedure was used to select the threshold to limit the *in silico* predicted false positive rate to 0 while maintaining a predicted high true positive detection rate of 98%, additional details are given in (6).

**Sequence analysis and microarray probe design to develop Census Array.** We included two types of probes on the Census Array: detection probes and census probes. Detection probes are conserved across multiple sequences from within a family or family-unclassified viral group, but not conserved across families or kingdoms (i.e. they are unique to a family). Such probes aim to detect known organisms or discover novel organisms that have not been sequenced but which possess some sequence homology to organisms that have been sequenced, particularly in those regions found to be conserved among previously sequenced members of that family. We have previously design a Lawrence Livermore Microbial Detection Array using this approach (8). These conserved probes may identify an organism to the level of genus or species, for example, but may lack the specificity to pin the identification down to strain or isolate. Census probes, in contrast, represent the least conserved regions, that is, the most strain or isolate specific probes. Such census probes aim to fill the goal of providing higher level discrimination and identification of known species and strains to facilitate forensic resolution, but may fail to detect novel organisms with limited homology to sequenced organisms. We included both types of probes on the Census Array to maximize the capability to

detect both well-characterized and novel microbes and to facilitate high confidence classification at both higher (family) and lower (species and strain) taxonomic levels.

The array design process is diagrammed in Figure 4. We downloaded all sequences, including complete genomes and sequence fragments (genes, noncoding regions, etc.), organized by family, for all bacteria and viruses, from NCBI GenBank, Integrated Microbial Genomics (IMG) project at the Joint Genome Institute, The Comprehensive Microbial Resource (CMR) at the JC Venter Institute, and The Sanger Institute in the United Kingdom, with some additional proprietary whole-genome data from collaborators. Bacteria were those under the superkingdom Bacteria (eubacteria) taxonomy node at NCBI, and did not include the Archaea. Sequence data for complete genomes, viral segments, and plasmids were current as of August 2009, and for sequence fragments as of January 2009. Table 10 summarizes the number of families, species, genomes, and sequence fragments represented on the array. The process of downloading the sequence data into curated groups required more than a week, with automated scripts running 24x7.

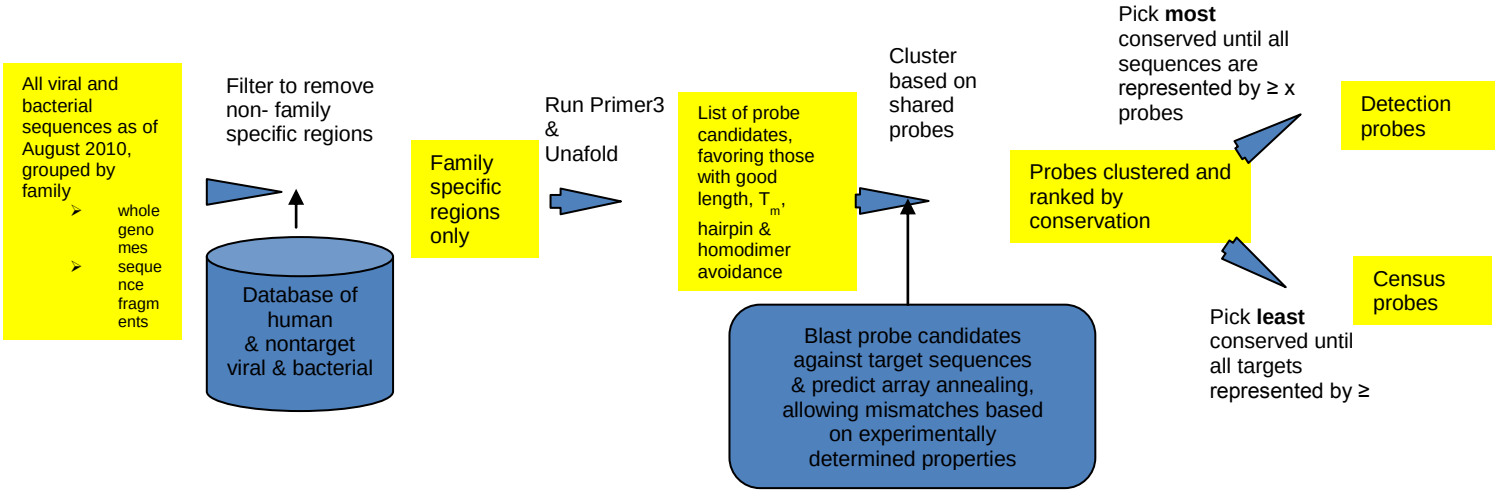



Figure 4: Diagram of the Census Array design process

We began the array design process by identifying the family specific regions in each family. In prior work, we have found that the length of longest perfect match (PM) is a strong predictor of hybridization intensity, and that for probes at least 50 nucleotide (nt) long, PM≤20 bp have signal less than 20% of that with a PM over the entire length of the probe (5). This is similar to results from a systematic study of viral probe hybridization characteristics by (9). Therefore, for each target family we eliminated regions with perfect matches to sequences outside the target family. Using the suffix array software vmatch (10), PM subsequences of at least 17 nt long present in non-target viral families or 25 nt long present in the human genome or non-target bacterial families were eliminated from consideration as possible probe

subsequences. Sequence similarity of probes to non-target sequences below this threshold was allowed, but could be accounted for using the statistical algorithm described below.

From these family-specific regions, we designed probes 50-66 bases long for one family at a time using the methods described in (5). Briefly, we generated candidate probes using MIT's Primer3 software (11), followed by $T_m$ and homodimer, hairpin, and probe-target free energy (ΔG) prediction using Unafold (6). Candidate probes with unsuitable ΔG's or $T_m$'s were excluded as described in (2). Desirable range for these parameters was 50≤length≤66, $T_m$≥80°C, 25%≤GC%≤75%, $\Delta G_{homodimer}$= ΔG of homodimer formation > 15 kcal/mol, $\Delta G_{hairpin}$=ΔG of hairpin formation > -11 kcal/mol, and $\Delta G_{adjusted} = \Delta G_{complement} - 1.45\ \Delta G_{hairpin} - 0.33\ \Delta G_{homodimer} \leq -52$ kcal/mol. An additional minimum sequence complexity constraint was enforced, requiring a trimer frequency entropy of at least 4.5 (calculation described below). If fewer than a minimum number of candidate probes per target sequence passed all the criteria, then those criteria were relaxed to allow a sufficient number of probes per target. To relax the criteria, first candidates that passed the Primer3 criteria but failed the Unafold filters were allowed. If no candidates passed the Primer3 criteria, then regions passing the target-specificity (e.g. family specific) and minimum length constraints were allowed.

Table 10: Summary of sequences represented on the Census Array

| ***Number of Targets*** | ***Virus*** | ***Bacteria*** |
|---|---|---|
| Families | 80 | 274 |
| Groups without family classification | 48 | 65 |
| Species with complete genome, plasmid, or segment data | 2530 | 1290 |
| Species with any sequence data, including sequence fragments | 5719 | 14765 |
| Sequences classified as to Family | 171264 | 728467 |
| Sequences unclassified as to Family | 6996 | 56251 |
| Complete genomes, segments, or plasmids | 55803 | 4122 |

Next, we BLASTed these candidates against the family of target sequences from which they were designed to predict the targets that should be represented by each candidate. A target was considered to be represented if a probe matched it with at least 85% sequence similarity over the total probe length, and a perfectly

matching subsequence of at least 29 contiguous bases spanned the central base of the probe (it could be off center, so long as it spanned the middle position). We ranked the probe candidates by their level of conservation, that is, how many targets they were predicted to represent. From here, we followed two contrasting strategies to pick 1) detection probes and 2) census probes.

For detection probes, we selected probes in *decreasing* order of the number of targets represented by that probe (i.e. probes detecting more targets in the family were chosen preferentially over those that detected fewer targets in the family). For probes that tied in the number of targets represented, a secondary ranking was used to favor probes most dispersed across the target from those probes which had already been selected to represent that target. The probe with the same conservation rank that occurs at the farthest distance from any probe already selected from the target sequence is the next probe to be chosen to represent that target.

For census probes, the process was similar except that we selected probes in *increasing* order of the number of targets represented by that probe (i.e. probes detecting fewer targets in the family were chosen preferentially over those that detected more targets in the family). The same criterion as described above was imposed to maximize positional dispersion of multiple probes across each target sequence. A minimum of 5 probes per target sequence were included. For sequences that diverged from other members of the family or that clustered as a highly conserved subgroup (e.g. multiple sequences from the same outbreak), the detection and census probes could be the same. Duplicate probes were removed in the final array design.

We included 5-30 detection probes per target and 1-10 census probes per target depending on the array density. Several versions of the Census Arrays were designed that differ in density, and thus cost. The standard array fits on the 388K NimbleGen design (Table 11). The NimbleGen 2.1M and Agilent 1M formats allowed more probes per target sequence. These higher density format arrays were not used for the NGFA-5 sample analysis. Detection probes were designed for all targets, both complete genomes and plasmids and sequence fragments. Census probes were design for all viruses, both complete and partial sequences, and for all complete bacterial genomes and plasmids. Census probes were also designed for sequence fragments for the ~240 bacterial families with less available sequence data (<~150MB), although array density limitations did not allow us to include census probes for the sequence fragment data for the 32 families with the most available sequence data (~200 MB-2.5 GB and with thousands of sequences), since those families were already so well-represented by the copious detection probes as well

as census probes for the numerous complete genomes. Moreover, these partial sequences included many extremely highly conserved rRNA genes which are inappropriate for strain discrimination. Additional probes representing the partial sequences for these already heavily represented families was thought to be unnecessary for the goal of strain discrimination.

Table 11: Types of microarray probes for the Standard 388K Census Array

| # of Probes | Probe Type | Comments |
|---|---|---|
| 380088 | Bacteria and virus census and detection probes | **Census probes**: 5 pps for all viral sequences and bacterial whole genomes and plasmids, 1 pps for bacterial sequence fragments from 248 families. **Detection probes**: 5 pps for all sequences (breakdown between census and detection is not relevant due to overlap between these sets) |
| 1821 | Hemorrhagic fever virus BioWatch amplicons | Tiled across amplicons with 50% overlap between probes (2x coverage) |
| 1860 | SFBB sequences | Tiled across amplicons with 0% overlap between probes (1x coverage) |
| 1235 | random controls | |
| 385004 | Total | |

Approximately 1,000- 3,500 random negative control sequences length and GC% matched to the target probes were included. These had no appreciable homology to known sequences based on BLAST similarity, and were used to assess background hybridization intensity. We included probes that tile across the BioWatch amplicons for the viral hemorrhagic fevers, to increase sensitivity in case the array were used to confirm a BioWatch hit for these organisms. Probes were also included to represent unpublished viral sequence fragments provided by our collaborators at the San Francisco Blood Systems Research Institute (abbreviated as SFBB for SF Blood Bank).

**Microarray Processing**

An overview of the microarray process is shown in Figure 5.

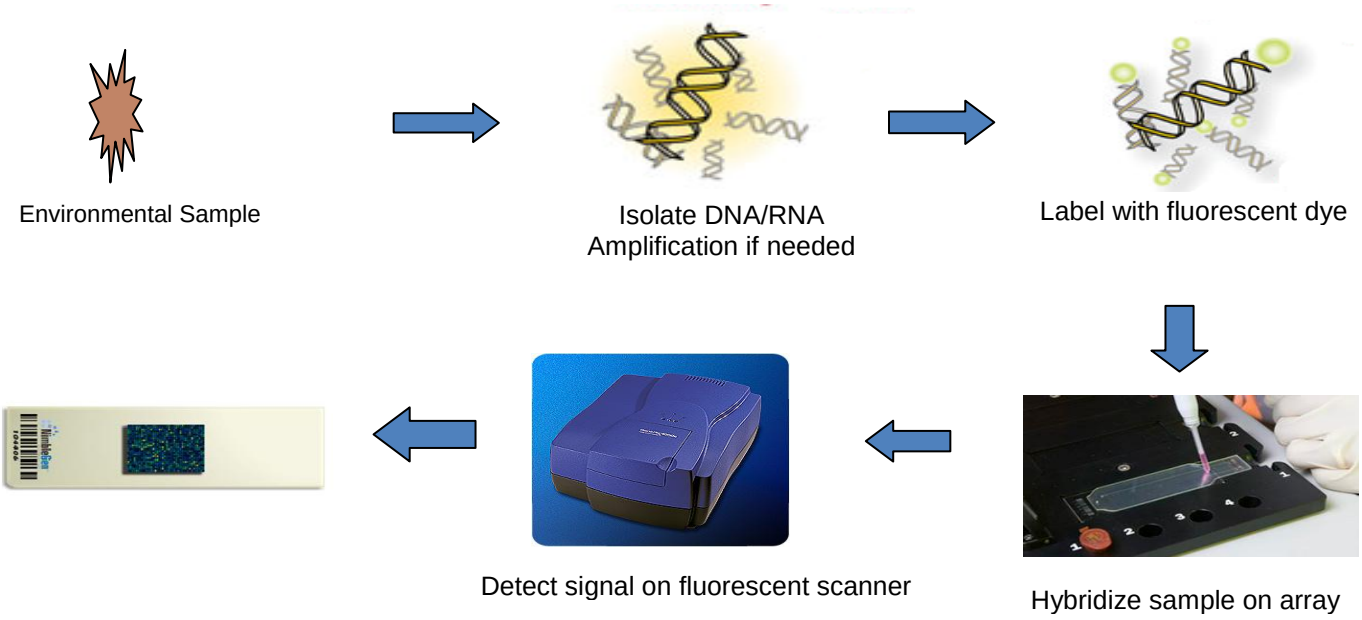



Figure 5. Microarray hybridization process

**Microarray hybridization.** All genomic *Bacillus anthracis* Ciprofloxacin resistant DNA samples were sonicated prior to labeling and hybridization. Sonication was performed using a Branson Digital Sonifier S450D (Branson Ultrasonics, Danbury, CT) at a setting of 100% amplitud**e** for 30 seconds. Sonication was repeated 4 times. The fragment sizes ranged from 500-2000 base pairs. Sample DNA concentrations were measured using a Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE).

The entire 40µL of amplified *B. anthracis* Ames spiked soil and aerosol sample product from the whole genome amplification and purification process as well as the *B. anthracis* ciprofloxacin resistant DNA was fluorescently labeled using the Roche NimbleGen One-Color DNA Labeling Kit #05223555001 (Madison, WI) according to the recommended protocols. The DNA was purified after labeling, and hybridized using the NimbleGen Hybridization Kit (Cat. 05583683001) to the LLNL Virulence Array or Census Array according to manufacturers' instructions. The microarrays were allowed to hybridize for 17 hours and washed using the NimbleGen Wash Buffer Kit #05584507001 according to manufacturer's instructions. Microarrays were scanned on an Axon GenePix 4000B 5 µm scanner from Molecular Devices (Sunnyvale, CA). The scanned tif image files were aligned using the NimbleScan Version 2.4 software and pair text files were exported for data analysis.

**Microarray Data Analysis.** A maximum likelihood analysis method was used to analyze the microbial hits from samples hybridized to the array. The method was recently published in (8). An example of the analysis results is shown in Figure 6 where *B. thuringiensis* israelensis was run on the Virulence Array. The analysis results of *B. thuringiensis* kurstaki run on the Census Array are shown in Figure 7. The right-hand columns of bar graphs show the unconditional and conditional log-odds ratios for each target genome listed at right. The unconditional log-odds is the larger of the two scores; thus the lighter and darker-colored portions represent the unconditional and conditional scores respectively. Targets are color-coded and grouped by taxonomic family, according to the legend at bottom; they are listed within families in decreasing order of conditional log-odds ratio scores. Targets predicted as likely to be present are indicated in red text. The vertical orange dashed line marks 0 on the log-odds ratio scale.

The left-hand columns of the bar graphs show the expectation (mean) values of the numbers of probes expected to be present given the presence of the corresponding target genome. The larger "expected" score is obtained by summing the conditional detection probabilities for all probes; the smaller "detected" score is derived by limiting this sum to probes that were actually detected. Because probes often cross-hybridize to multiple related genome sequences, the numbers of "expected" and "detected" probes often greatly exceed the number of probes that were actually designed for a given target organism.

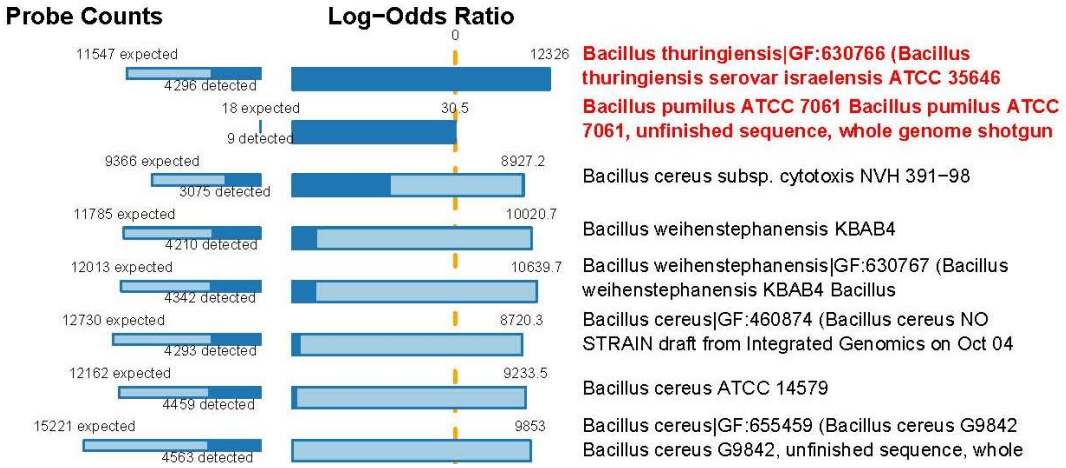



Figure 6. Virulence Array results for *B. thuringiensis* israelensis

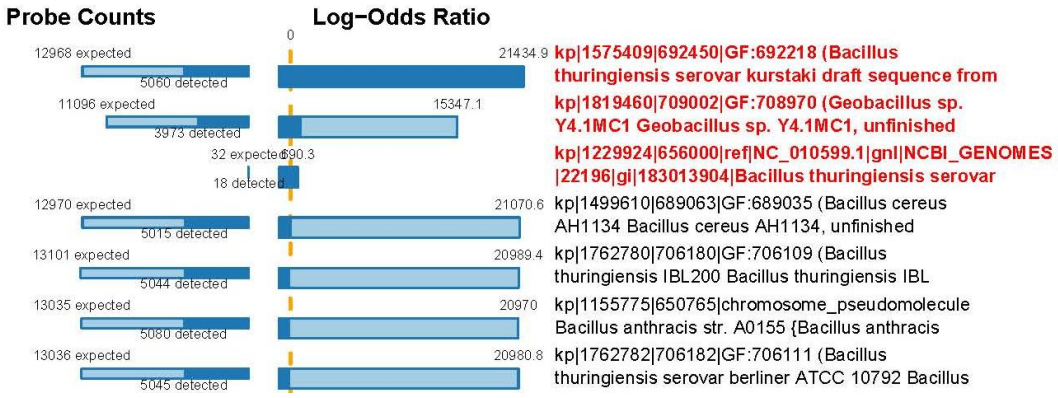



Figure 7. Census Array result for *B. thuringiensis* kurstaki

The probe count bar graphs are designed to provide some additional guidance for interpreting the prediction results. For example, we see that the prediction for presence of *B. pumilus* ATCC 7061 is based on only 9 detected probes; it is therefore given much less weight than the prediction for *B. thuringiensis* serovar *israelensis*, which is based on over 4000 probes.

*Illumina sequencing*

***Bacillus anthracis* spiked soil and aerosol samples.** Extracted DNA from *B. anthracis* spiked soil and aerosol samples were sequenced using Illumina technology. Eureka Genomics prepared a paired end non-indexed standard Illumina library on each DNA sample (even though in some of the samples single end data was generated). Eureka Genomics generated sequencing reads using Illumina GAIIx sequencing technology from 6 aerosol samples and from 6 soil samples with one sample per lane. No information on the composition of the spiked samples (e.g. proportion of *B. anthracis* spiked into each sample) or whether any of the spiked samples was control (no spike) was provided to the Eureka Genomics Bioinformatics Department. In this regard, this was a blind study. Eureka Genomics analyzed these twelve sets of sequencing reads to determine the impact of the environment on the detection of *B. anthracis* by Illumina sequencing.

***Bacillus anthracis* ciprofloxacin resistant mutants.** Illumina paired end libraries were also prepared from 1 µg of genomic DNA (gDNA) from each of the eleven third round Ciprofloxacin resistant *B. anthracis* isolates, for the purpose of single-end sequencing on the Genome Analyzer IIx. Briefly, the gDNA was fragmented, end repaired, A' tagged, ligated to adaptors, size-selected and enriched with 13 cycles of PCR. Each library was assigned one lane of a flow cell to undergo

cluster amplification and sequencing on the Genome Analyzer IIx, and 36 cycles of single-end sequence data was generated. The resulting sequencing reads were filtered using the default parameters of the Illumina QC pipeline (Bustard + Gerald).

As an additional quality control step, all reads were analyzed using the PIQA pipeline, which examines genomic reads produced by Illumina machines and provides tile-by-tile and cycle-by-cycle graphical representations of cluster density, quality scores, and nucleotide frequencies to allow easy identification of defective tiles, mistakes in sample/library preparations and abnormalities in the frequencies of appearance of sequenced genomic reads. All reads were determined to be of sufficient quality to proceed with subsequent analysis.

**Mapping and identifying candidate mutations.** The sequence reads from each of the samples were mapped with up to 1 mismatch to the reference *B. anthracis* Sterne genome (AE017225.1). To avoid uncertainty associated with identifying mutations in repeatable parts of the reference genome, for each position in the reference sequence a *uniqueness score* based on the subsequences covering this nucleotide was determined. Specifically, the copy number of each subsequence of size 36 (the length of reads used in sequencing) present in the reference genome was first calculated; the *uniqueness score* of each position in the reference genome was then defined as the total number of subsequences (factoring in the copy number) which covered this position. For example, in this metric, the score of 36 will appear only if each subsequence covering a given nucleotide is unique in the reference; higher scores indicate that one or more subsequences are present in the reference in several copies. 94.11 % (1,784,242 bases) of the reference genome has a uniqueness score of 36. Mutations in these positions can be detected without the ambiguity caused by the presence of repeatable regions.

A given position is predicted to contain a mutation if: (1) the number of reads confirming the mutation on each strand exceeds the *minimum count threshold* – ensuring that only positions that achieve the minimum required coverage are considered, and (2) the proportion of reads confirming a mutation out of all the reads covering a given position exceeds a *ratio threshold* – ensuring that only mutations that have the minimum required support are identified. As a compromise between mutation detection sensitivity and false discovery rate, the *minimum count threshold* was set at 10% of the median of the nucleotide-by-nucleotide coverage for each sample, and the *ratio threshold* was set at 30% of the total coverage on a per-nucleotide basis. In the present analysis, mutations confirmed on both strands (if the number of reads supporting the mutation exceeds the *minimum count threshold* on each of the strands separately) are distinguished from mutations for which such condition was met on only one strand. In the case of insertions, the mapping process results in the association of both perfect matches (PM) and insertions to the same location on the reference genome. Thus different *ratio threshold* criteria are used to detect different types of mutations at a given genome position. The criterion for detecting a substitution of base *B* for the reference base is:

$$\frac{SubB^{+} + SubB^{-}}{PM^{+} + PM^{-} + Del^{+} + Del^{-} + SubACTG^{+} + SubACTG^{-}} \geq ratio\ threshold$$

The criterion for detecting a deletion is:

$$\frac{Del^{+} + Del^{-}}{PM^{+} + PM^{-} + Del^{+} + Del^{-} + SubACTG^{+} + SubACTG^{-}} \geq ratio\ threshold$$

The criterion for detecting an insertion of base *B* on the plus strand is:

$$\frac{InsB^{+} + InsB^{-}}{Del^{+} + Del^{-} + SubACTG^{+} + SubACTG^{-} + InsACTG^{+} + InsACTG^{-}} \geq ratio\ threshold$$

In the numerators of the above formulas, $SubB^{+/-}$, $Del^{+/-}$, and $InsB^{+/-}$ stand for the numbers of reads confirming a substitution, deletion, or insertion, respectively, mapping to the genome strand indicated by the superscript. For substitutions and insertions, $SubB^{-}$ and $InsB^{-}$ indicate the numbers of reads mapped to the minus strand in which the base complementary to *B* is substituted or inserted. In the denominators, the variables *PM*, *SubACTG*, and *InsACTG* respectively indicate the numbers of reads confirming a perfect match (PM), a substitution of any base, or an insertion of any base, at the genome position of interest.

While paired end data was generated, the reads were decoupled and a single-end read assembly (using in house algorithms) was performed on each of the sequence data sets. These contigs are shorter in length than contigs obtained with paired end data, but in general have fewer errors. Each mutation identified in each sample was confirmed to be present on the contigs assembled for that sample. Mutations (including insertions, deletions, and substitutions) that pass both thresholds and appear on both strands are less likely to be artifacts of sequencing read generation or artifacts of mapping. Mutations that only appear on one strand and cannot be verified on the opposite strand (something that is not common, given sufficient coverage), such as insertions, other than 'G' after 'G', 'C' after 'C', 'A' after 'A', and 'T' after 'T' are either artifacts of sequencing/mapping (false positives) or positions in the genome that did not have sufficient coverage to be verified on both strands.

*454 sequencing*

***Bacillus anthracis* ciprofloxacin resistant mutants.** Genomic DNA samples from the *Bacillus anthracis* Sterne ciprofloxacin mutants were provided to the DNA Sequencing Center at Brigham Young University. The samples were sequenced according to standard Roche 454 procedures. Each mutant was sequenced per half-plate of the run with a total of 2 454 Life sciences Titanium runs on the Genome Sequencer FLX. Bases on these parameters it was approximated that ~166 million bases of sequence data would be generated with a median read length of 242 bases.

***Bacillus anthracis* spiked soil and aerosol samples.** Spiked soil and aerosol *B. anthracis* samples were also analyzed by 454 sequencing. Brigham Young University generated 454 sequencing reads from one 96-well plate with 4 aerosol and 4 soil samples.

**Sequence Generation.** The 454 sequence reads were prepared by Roche 454 and provided to ORNL by LLNL. The quality filtered reads and quality scores were sent in FASTA and SFF formats. Reads were generated for the four ciprofloxacin resistant strains (M1-1, M1-6, M10-8-1, M19-2), as well as the non-resistant parental strain (Dugway).

Since the Sterne strain is the most recent fully annotated ancestor, reads from each of the mutant strains as well as the parental strain were mapped to the Sterne chromosome (GenBank: AE017225.1) and the annotated pX01 plasmid (GenBank: AF065404.1) using the mapping software (gsMapper) provided by the vender (Roche) of the sequencing instrument.

Reads were initially mapped using 100%, 97%, 94%, 91%, and 90% minimum identity thresholds between the reads and reference to obtain an understanding of the data set. One hundred percent, 97%, 94%, and 91% are approximately equivalent to 0, 1, 2, and 3 mismatches in a 35 base pair read (thus these thresholds correspond to the thresholds available in Illumina mapping software used in the complementary experiment). Ninety percent was used in subsequent genomic variation analysis since it is the default recommended for the software and based on a survey of the different thresholds seemed to be reasonable. Mapping statistics for the mappings using 90% minimum identity were parsed from the output and are presented in (Table 12, Script #5).

For each mapping gsMapper provides a high quality difference file representing SNPs and indels. SNPs and indels for each strain were taken from the 'variants' tab in the gsMapper software. Variants from mapping the parental strain (Dugway) to Sterne were assumed to represent variations accumulated prior to antibiotic resistance selection or as potential sequencing errors in the reference genome or parental strain. Variants detected from each resistant strain that were not in the parental strain were found by selectively removing rows corresponding to SNPs in both the parental strain and in the resistant strain (Table 12, Script #3). The lowest quality score associated with each variation was obtained to identify detected variations that may be an error due to sequencing quality (Table 12: Script

#8). All variants common among all the resistant strains were identified manually. In-house scripts were used to annotate (gene locus/protein id, etc.) the positions of all the SNPs and indels (Table 12: #1, #2). There were two SNPs common in all four of the resistant strains that (a) were not present in the parental strain, and (b) were present in regions that encode proteins. The protein variants of these two SNPs were found manually.

Large deletions that are identified as unmapped portions of the reference genome were found using the mappings between the reads and Sterne reference. The 454alignmentinfo.tsv file generated by the gsMapper software was parsed (Table 12: Script #4) to find large regions of the reference genome that had a consensus base of "-" in the mapping (regions indicated in Figure 8 with non-unique reads 'red' and coverage greater than one – the far right side). These regions had a total depth of greater than or equal to 2, but no unique depth. Regions omitted from 454alignmentinfo.tsv were also found using the same script. These regions have no consensus base, i.e. regions with a total depth less than 2, even if it had a unique depth of 1 (the far left and middle regions in Figure 8). To discern between areas with no consensus base that have a unique depth of 1 (indicated by region with single unique read 'blue' in Figure 8) and those that do not, the 454AllContigs.fna file, generated by the gsMapper software, was parsed to find regions not mapped (Table 12: #6).

The 454ReadStatus.txt file, generated by the gsMapper software, was parsed using an in-house script to find the identification (ID) of the reads not mapped for all mappings using 90% minimum identity (Table 12: #7).


Figure 8: Graphical explanation of how non-unique regions and regions with coverage less than two are recorded in the provided files from gsMapper. Regions in the figure are not included in files whose name has a strike through in it. Large deletions were found by looking for regions not represented in the .454alginmentsInfo.tsv file.

Table 12: Scripts used in 454 analysis.

| Script # | Script Name | Function |
|---|---|---|
| 1 | annotate_Ba.py | Provides annotation for variant regions of the chromosome (GenBank: AE017225.1) |
| 2 | annotate_px01.py | Provides annotation for variant regions of the PX01 plasmid (GenBank: AF065404.1) |
| 3 | diff2.py | Finds the variants present in the resistant strains that are not in the parental strain |
| 4 | large_deletesv2.py | Parses 454AlignmentInfo.tsv provided by gsMapper to find large areas mapped as '-' and areas with no consensus base |
| 5 | map_stats.py | Parses out desired mapping statistics from 454MappingQC.xls provided by gsMapper |
| 6 | parse_contigs.py | Parses 454AllContigs.fna provided by gsMapper to find regions that are not mapped |
| 7 | parse_unmapped.py | Parses out names of unmapped reads from 454ReadStatus.txt provided by gsMapper |
| 8 | quality_scores.py | Parses out lowest quality score for a variant region using 454AlignmentInfo.tsv provided by gsMapper |

# RESULTS

## *TaqMan Assays*

**Canonical SNP assays with pure *Bacillus anthracis* DNA.** We isolated genomic DNAs from *B. anthracis* Ames, Sterne and A0382. The DNAs were tested in 2 replicates with the canonical SNP assays. The data is shown in Table 13. Ames is mostly reactive to CanSNPs 1-5 (with signal from the VIC dye) as expected because these five canonical SNPs correspond to the branches with phylogenetic lineages closely related to Ames. Canonical SNP1 is specific to Ames strain. The Sterne strain is mostly reactive to CanSNPs 2-5, also as expected since CanSNP1 is specific to Ames, while CanSNPs 2-5 correspond to the phylogenetic lineages of the Sterne strain. The A0382 strain has not been sequenced. It is reactive to CanSNP 10 and 11, suggesting that it is closely related to the Kruger strain.

Table 13. Canonical SNP results from *B. anthracis* DNAs

| | Ames | | Sterne | | A0382 | |
|---|---|---|---|---|---|---|
| | **Ave Ct** | **Stdev** | **Ave Ct** | **Stdev** | **Ave Ct** | **Stdev** |
| **canSNP1** | 20.11 | 0.36 | 18.59 | 0.1 | 19.08 | 0.24 |
| **canSNP2** | 22.16 | 0.15 | 20.1 | 0.2 | – | – |
| **canSNP3** | 23.07 | 0.19 | 20.51 | 0.15 | 22.16 | 0.09 |
| **canSNP4** | 22.01 | 0.13 | 19.94 | 0.28 | 21.8 | 0.23 |
| **canSNP5** | 21.54 | 0.25 | 20.07 | 0.24 | 21.72 | 0.12 |
| **canSNP6** | 20.3 | 0.14 | 18.72 | 0.17 | 19.22 | 0.13 |
| **canSNP7** | 22.86 | 0.16 | 20.54 | 0.17 | 21.26 | 0.16 |
| **canSNP8** | 22.84 | 0.17 | 19.58 | 0.14 | 20.1 | 0.13 |
| **canSNP9** | 21.82 | 0.1 | 19.58 | 0.1 | 20.02 | 0.09 |
| **canSNP10** | 23.59 | 0.35 | 21.48 | 0.63 | 20.07 | 0.24 |
| **canSNP11** | 21.28 | 0.18 | 19.01 | 0.14 | 19.73 | 0.13 |
| **canSNP12** | 20.72 | 0.21 | 19.09 | 0.11 | 19.76 | 0.13 |
| **canSNP13** | 23.67 | 0.07 | 21.4 | 0.09 | 22.66 | 0.11 |

FAM
VIC
– = Undetermined

**Determination of the limit of detection of the Canonical SNP assays using *Bacillus anthracis* Ames spiked into BioWatch aerosol samples.** We performed limit of detection testing of the canonical SNP assays using serially diluted *Bacillus anthracis* Ames spiked into 100 pg of DNA from BioWatch aerosol filter extracts. Duplicate experiments were run to ensure repeatability and data consistency. 1, 10, 100, 1,000, 10,000, and 100,000 copies of *B. anthracis* Ames were tested. Table 14 below shows results of the TaqMan SNP assays at each of the *Bacillus anthracis* DNA concentrations. When 10 copies of *Bacillus anthracis* DNA were spiked into the aerosol sample, numerous canonical SNP assays were undetermined. This experiment suggested that our detection limit for *Bacillus anthracis* Ames is 100 genome copies when the DNA was spiked into 100 pg of the aerosol DNA sample.

Table 14. Limit of detection of *Bacillus anthracis* Ames DNA spiked in Biowatch aerosol samples

| **Amount *B. anthracis* DNA** | 560 pg | | 56 pg | | 5.6 pg | | 560 fg | | 56 fg | | 5.6 fg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***B. anthracis* DNA Copy #** | 100,000 copies | | 10,000 copies | | 1,000 copies | | 100 copies | | 10 copies | | 1 copy | |
| **% BA DNA in aerosol DNA** | 98.20% | | 35.90% | | 5.30% | | 0.56% | | 0.06% | | 0.01% | |
| | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** |
| **canSNP1** | 21.26 | 0.33 | 25.04 | 0.15 | 28.38 | 0.10 | 31.87 | 0.28 | 36.26 | 1.07 | – | – |
| **canSNP2** | 23.99 | 1.51 | 28.71 | 2.38 | 31.66 | 1.03 | 34.83 | 1.31 | – | – | – | – |
| **canSNP3** | 22.66 | 1.02 | 27.09 | 0.33 | 30.81 | 0.54 | 34.55 | 0.37 | 37.07 | N/A | – | – |
| **canSNP4** | 21.97 | 0.84 | 26.11 | 0.38 | 29.60 | 0.54 | 32.93 | 0.30 | 35.81 | 0.14 | – | – |
| **canSNP5** | 22.21 | 0.78 | 26.39 | 0.75 | 30.18 | 0.70 | 33.63 | 0.85 | 36.13 | N/A | – | – |
| **canSNP6** | 21.34 | 0.54 | 25.50 | 0.32 | 29.12 | 0.45 | 32.24 | 0.43 | 36.17 | 0.10 | – | – |
| **canSNP7** | 23.40 | 0.75 | 27.48 | 0.47 | 30.67 | 0.43 | 33.79 | 0.94 | 36.51 | N/A | – | – |
| **canSNP8** | 22.55 | 0.54 | 27.18 | 0.62 | 30.58 | 0.45 | 33.86 | 0.59 | 37.05 | N/A | – | – |
| **canSNP9** | 22.05 | 0.69 | 26.32 | 0.57 | 29.73 | 0.49 | 33.01 | 0.62 | 36.77 | 0.23 | – | – |
| **canSNP10** | 22.76 | 0.19 | 27.08 | 0.13 | 30.65 | 0.40 | 34.11 | 0.07 | 36.87 | N/A | – | – |
| **canSNP11** | 21.15 | 0.01 | 25.44 | 0.11 | 28.65 | 0.07 | 32.02 | 0.49 | 36.32 | 0.90 | – | – |
| **canSNP12** | 21.85 | 0.13 | 25.77 | 0.01 | 29.33 | 0.02 | 32.83 | 0.15 | 35.97 | 0.00 | – | – |
| **canSNP13** | 21.92 | 0.28 | 26.01 | 0.10 | 29.27 | 0.17 | 32.76 | 0.40 | 36.75 | N/A | – | – |

FAM

VIC

**–** = undetermined

**Determination of the limit of detection of the Canonical SNP assays using *B. anthracis* Ames spiked into soil samples.** We performed a similar limit of detection test of the SNP assays using serially diluted *B. anthracis* Ames spiked into 1 ng of DNA from soil extracts. The soil was a combination of soils collected locally in San Francisco and Oakland. Six different DNA concentration levels of *B. anthracis* Ames were tested, from 1 copy to 100,000 copies. Table 15 below shows results of the SNP assays at each of the *B. anthracis* DNA concentration. When 100 copies of *B. anthracis* DNA were spiked into the soil sample numerous SNP assays were undetermined. This experiment suggested that our detection limit for *B. anthracis* Ames is 1000 copies when the DNA was spiked into 1 ng of soil DNA sample.

Table 15. Limit of detection of *B. anthracis* Ames DNA spiked in soil samples.

| **Amount *B. anthracis* DNA** | 560 pg | | 56 pg | | 5.6 pg | | 560 fg | | 56 fg | | 5.6 fg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***B. anthracis* DNA Copy #** | 100,000 copies | | 10,000 copies | | 1,000 copies | | 100 copies | | 10 copies | | 1 copy | |
| **% BA DNA in soil DNA** | 35.90% | | 5.30% | | 0.56% | | 0.06% | | 0.01% | | 0.00% | |
| | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** | **Avg** | **Stdev** |
| **canSNP1** | 27.18 | 1.19 | 30.25 | 0.28 | 33.33 | 0.32 | 36.62 | 0.51 | – | – | – | – |
| **canSNP2** | 30.96 | 2.53 | 32.34 | N/A | – | – | – | – | – | – | – | – |
| **canSNP3** | 27.54 | 0.23 | 30.50 | 0.53 | 34.46 | 0.15 | – | – | – | – | – | – |
| **canSNP4** | 24.89 | 0.15 | 27.43 | 0.40 | 31.09 | 0.01 | 35.10 | 0.17 | – | – | – | – |
| **canSNP5** | 26.74 | 0.60 | 29.18 | 1.35 | 33.54 | 0.18 | – | – | – | – | – | – |
| **canSNP6** | 24.29 | 0.69 | 27.02 | 0.79 | 31.12 | 0.56 | 35.11 | 0.04 | – | – | – | – |
| **canSNP7** | 26.66 | 1.40 | 29.01 | 1.07 | 33.01 | 0.28 | – | – | – | – | – | – |
| **canSNP8** | 24.51 | 0.59 | 27.98 | 0.68 | 31.91 | 0.10 | 36.18 | 0.27 | – | – | – | – |
| **canSNP9** | 23.73 | 0.98 | 26.71 | 0.64 | 31.10 | 0.30 | 35.05 | 0.17 | – | – | – | – |
| **canSNP10** | 27.04 | N/A | 29.78 | 0.29 | 34.63 | N/A | – | – | – | – | – | – |
| **canSNP11** | 24.78 | 0.22 | 27.21 | 1.02 | 31.34 | 0.38 | – | – | – | – | – | – |
| **canSNP12** | 26.93 | 0.17 | 29.65 | 0.20 | 34.12 | 0.20 | – | – | – | – | – | – |
| **canSNP13** | 27.12 | 0.33 | 28.26 | 0.77 | 32.30 | 0.02 | 36.44 | 0.20 | – | – | – | – |

FAM

VIC

– = undetermined

***Bacillus thuringiensis* kurstaki Real-time PCR assays with pure DNA.** We isolated genomic DNAs from *B. thuringiensis* kurstaki ATCC 33679, HD-1 and 342. The DNAs were tested in duplicate with the 5 different signatures at 1ng per reaction. The data is shown in Table 16. A panel of 18 near neighbor *Bacillus* DNAs was collected to test the exclusivity of the Real-time kurstaki assays. The data is shown in Table 17.

Table 16. Inclusivity Panel: *B. thuringiensis* kurstaki target results

| | Signatures | | | | |
|---|---|---|---|---|---|
| **Template** | **2254619** | **2254620** | **2254621** | **2254622** | **2254623** |
| ATCC 33679 | 21.3 (0.2) | 22.3 (0.2) | 21.3 (0.2) | 20.5 (0.3) | 20.0 (0.4) |
| HD-1 | 20.6 (0.1) | 20.6 (0.0) | 19.4 (0.2) | 17.1 (0.1) | 18.7 (0.1) |
| 342 | 20.6 (0.1) | 20.6 (0.0) | 22.1 (3.4) | 16.5 (0.1) | 18.3 (0.1) |

Data is expressed as Ct (St Dev)

Table17. *B. thuringiensis* kurstaki exclusivity panel results

| **Template** | | | **Signatures** | | | | |
|---|---|---|---|---|---|---|---|
| **Species** | **Subspecies/Strain** | **Branch** | **2254619** | **2254620** | **2254621** | **2254622** | **2254623** |
| *thuringiensis* | israelensis HD500 | A | 37.1 (1.2) | 36.1 (1.2) | ND | ND | ND |
| | finitimus HD527 | G | ND | ND | ND | 37.6 (N/A) | ND |
| | sotto HD774 | A | 37.3 (0.1) | 37.8 (0.1) | ND | ND | ND |
| | pakistani HD462 | C | ND | ND | ND | ND | ND |
| | konkukian 97-27 | C | ND | ND | ND | ND | ND |
| | AH547 | K | 20.7 (0.2) | 21.1 (0.3) | ND | ND | ND |
| | AH535 | J | 37.0 (1.5) | 36.7 (1.0) | ND | 36.5 (N.A) | ND |
| | AH575 | H | 39.0 (1.0) | 38.9 (N/A) | ND | ND | ND |
| | AH592 | H | ND | ND | ND | ND | ND |
| | AH678 | K | ND | ND | ND | ND | ND |
| | HD95 | B | ND | ND | ND | ND | ND |
| | HD1101 | F | ND | ND | ND | ND | ND |
| | HD18 | B | ND | ND | ND | ND | ND |
| *anthracis* | Sterne | F | ND | ND | ND | ND | ND |
| | Ames | F | ND | ND | ND | ND | ND |
| | A0382 | F | ND | ND | ND | ND | ND |
| *cereus* | ATCC 4342 | E | ND | ND | ND | ND | ND |
| | D21 | D | ND | ND | ND | ND | ND |

Data is expressed as Ct (St Dev). ND = Not Detected

Based on the data shown in Tables 16 and 17 it was determined that signatures 2254621 and 2254623 were the best candidates for *Bacillus thuringiensis* kurstaki. These both showed high reactivity with the inclusivity panel and did not

react with any templates in the exclusivity. Ultimately 2254623 was deemed superior due to its slightly better detection of the kurstaki targets.

***Bacillus thuringiensis* israelensis Real-time PCR assays with pure DNA.** DNA was isolated from B. thuringiensis israelensis strains HD500 and ATCC 35646. All 3 signatures were tested against these target templates at 1ng per reaction. The data for this testing is below in Table 18.

Table 18. Inclusivity Panel: *B. thuringiensis* israelensis target results

| | **Signatures** | | |
|---|---|---|---|
| **Template** | **2253467** | **2253469** | **2253470** |
| HD500 | 19.9 (0.4) | 18.8 (0.3) | 17.7 (0.1) |
| ATCC 35646 | 19.7 (0.2) | 18.7 (0.1) | 17.7 (0.1) |

Data is expressed as Ct (St Dev).

An exclusivity panel of 20 near neighbor *Bacillus* DNAs was tested against each signature. Each template DNA was tested in duplicate reactions with 1ng per reaction. The data for this testing can be seen below in Table 19.

Table 19: *B. thuringiensis* israelensis exclusivity panel results

| Template | | | Signatures | | |
|---|---|---|---|---|---|
| **Species** | **Subspecies/Strain** | **Branch** | **2253467** | **2253469** | **2253470** |
| *thuringiensis* | kurstaki ATCC 33679 | C | ND | ND | ND |
| | kurstaki 342 | | ND | ND | ND |
| | kurstaki HD1 | C | ND | ND | ND |
| | finitimus HD527 | G | ND | ND | ND |
| | sotto HD774 | A | 34.8 (1.4) | 32.3 (0.7) | 32.1 (0.5) |
| | pakistani HD462 | C | ND | ND | ND |
| | konkukian 97-27 | F | 38.4 (N/A) | 35.6 (N/A) | 36.4 (0.2) |
| | AH547 | K | 35.6 (0.1) | 34.2 (0.8) | 34.2 (0.6) |
| | AH535 | J | 34.7 (0.2) | 33.6 (0.5) | 32.3 (0.0) |
| | AH575 | H | 35.8 (0.1) | 34.3 (0.1) | 34.8 (1.0) |
| | AH592 | H | ND | ND | ND |
| | AH678 | K | ND | ND | ND |
| | HD95 | B | ND | ND | ND |
| | HD1101 | F | 36.6 (0.2) | 35.1 (0.8) | 33.7 (1.0) |
| | HD18 | B | ND | ND | ND |
| *anthracis* | Sterne | F | ND | ND | ND |
| | Ames | F | ND | ND | ND |
| | A0382 | F | ND | ND | ND |
| *cereus* | ATCC 4342 | E | ND | ND | ND |
| | D21 | D | ND | ND | ND |

Data is expressed as Ct (St Dev). ND = Not Detected

The 3 israelensis signatures reacted very similarly to each other. Each signature showed high reactivity with the inclusivity panel of templates. Signature 2253470 showed the highest sensitivity with the panel and 2253469 the lowest. The results with the exclusivity panel were mixed. Each signature reacted weakly (~32-38 Ct) with 6 different *B. thuringiensis* near neighbors. Although not ideal, the difference in reactivity between the inclusivity and exclusivity panels does point to these signatures being very specific.

***Bacillus thuringiensis* kurstaki Real-time PCR assay 2254623 with environmental samples.** The *B. thuringiensis* kurstaki signature 2254623 was used for further tests involving environmental samples. Aerosol filters collected on 5/3/2007 from a DHS surrogate study were tested. These filters were from an area in which *B. thuringiensis* kurstaki was sprayed in order to control the Gypsy Moth. In addition, a sample extracted from gauge wipes used in an EPA exercise was tested

using the 2254623 signature. This wipe was used to wipe dirty indoor surfaces and was inoculated with *B. thuringiensis* kurstaki spores. Both samples were too low in DNA concentration to be accurately read on the Qubit fluorometer and therefore were in the femto-picogram range. Five microliters of each sample was tested in duplicate and the results can be seen in Table 20.

Table 20: *B. thuringiensis* kurstaki signature 2254623 environmental sample results

| Sample | Avg Ct | St. Dev |
|---|---|---|
| Gypsy Moth 5/3/07 #605 | 22.78 | 0.3 |
| EPA Gauge wipes | 30.7 | 0.3 |

# *MLVA*

*Determining the limit of detection of the MLVA assays using B. anthracis Ames spiked into BioWatch aerosol samples*

We performed limit of detection testing of the MLVA assays using serially diluted *B. anthracis* Ames DNA spiked into 100 pg of DNA extracted from BioWatch aerosol filter extracts. Duplicate experiments were run to ensure repeatability and data consistency. Samples containing 1, 10, 100, 1,000, 10,000, and 100,000 copies of the *B. anthracis* Ames genome were tested. Table 21 below shows the results of the MLVA assays. Only one of the eight MLVA assays detected the Ames DNA in the sample containing the BioWatch aerosol DNA containing only one copy of the Ames genome. Six of the eight MLVA assays detected 10 genome copies of Ames DNA but the fragment size amplified by the CG3 primers was not correct for one of the two reactions. All eight of the MLVA assays amplified the appropriate DNA amplicons when applied to samples containing BioWatch aerosol DNA and 100 copies of Ames genome although the pXO2-at assay produced product in only one of the two reactions. These results suggest that the MLVA assays can detect from 10 to 100 copies of the Ames genome present in the BioWatch aerosol samples although, at the lower concentrations, the amplicon size produced may not be a reliable indicator of which strain is present. Still, the presence of *B. anthracis* DNA in the sample would be unequivocal.

Table 21. Limit of detection of *B. anthracis* Ames DNA spiked in Biowatch aerosol samples. Text in red indicates an amplicon of the wrong length. ND = not detected.

| Added *B. anthracis* DNA | 560 pg | | 56 pg | | 5.6 pg | | 560 fg | | 56 fg | | 5.6 fg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Added # of genomes* | 100,000 | | 10,000 | | 1,000 | | 100 | | 10 | | 1 | |
| % Ames DNA of total DNA in sample | 98.20% | | 35.90% | | 5.30% | | 0.56% | | 0.06% | | 0.01% | |
| | Intensity | | Intensity | | Intensity | | Intensity | | Intensity | | Intensity | |
| | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 |
| *VrrA* | 9400 | 9600 | 9400 | 10000 | 9800 | 9700 | 9500 | 9300 | 1400 | 2100 | ND | ND |
| *VrrB1* | Overloaded | 5800 | 7800 | 6200 | 8200 | 8100 | 9600 | 9500 | 6600 | 4800 | 1200 | ND |
| *VrrB2* | 7500 | 7600 | 9300 | 9900 | 9500 | 9000 | 1700 | 900 | ND | 500 | ND | 60 |
| *VrrC1* | Overloaded | 8000 | 10000 | 8400 | 9000 | 9000 | 9000 | 9000 | 990 | 400 | 300 | 300 |
| *VrrC2* | Overloaded | Overloaded | 8000 | Overloaded | 10000 | 9000 | 9000 | 9000 | 1800 | 2000 | ND | ND |
| CG3 | 5000 | 5400 | 5000 | 4900 | 9600 | 8600 | 400 | 4900 | 400 | 1000 | 200 | 120 |
| pX01 | 5600 | 4800 | 4500 | 4300 | 8600 | 8700 | 5900 | 4600 | 700 | 950 | 72 | 79 |
| pX02 | 7200 | 7200 | 4700 | 3700 | 400 | 300 | 90 | ND | ND | ND | ND | ND |

*Determining the limit of detection of the MLVA assays using B. anthracis Ames spiked into soil samples*

We performed a similar limit of detection test of the MLVA assays using serially diluted *B. anthracis* Ames DNA spiked into 1ng of DNA extracted from soils. The soil was a combination of soils collected locally in San Francisco and Oakland. Duplicate reactions for each MLVA allele were completed. Samples containing 1, 10, 100, 1,000, 10,000, and 100,000 copies of *B. anthracis* Ames were subjected to the MLVA assays. Table 22 shows the results of these MLVA assays for samples containing each of the different *B. anthracis* DNA concentrations. MLVA assays were not able to detect the Ames DNA in samples containing one copy of the Ames genome. However, seven of the eight MLVA assays detected the Ames DNA when 10 copies of the Ames genome were present in at least one of the two reactions. Only the pXO2-at assay was unable to detect the Ames DNA at this concentration in either of the two reactions. When 100 copies of *B. anthracis* Ames DNA was spiked into the soil sample, all of the MLVA assays were able to detect this DNA although the pXO2-at primers detected the Ames DNA in only one of the two reactions. Generally, the MLVA assays that were most sensitive when tested against samples containing BioWatch filter DNA were also the most sensitive at detecting the Ames DNA in samples containing the soil DNA. However, all but one of the MLVA assays was able to reliably detect 100 copies of the *B. anthracis* Ames DNA in the samples containing the soil DNA. Five of the eight MLVA assays were able to detect 10 copies of Ames DNA.

Table 22. Limit of detection of *B. anthracis* Ames DNA spiked in soil DNA samples. ND = not detected.

| Added *B. anthracis* DNA | 560 pg | | 56 pg | | 5.6 pg | | 560 fg | | 56 fg | | 5.6 fg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Added # of genomes | 100,000 | | 10,000 | | 1,000 | | 100 | | 10 | | 1 | |
| % Ames DNA of total DNA | 35.90% | | 5.30% | | 0.56% | | 0.06% | | 0.01% | | 0.00% | |
| | Intensity | | Intensity | | Intensity | | Intensity | | Intensity | | Intensity | |
| | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 | Rep 1 | Rep 2 |
| *VrrA* | 9000 | 9000 | 8600 | 8900 | 8000 | 8900 | 2000 | 8700 | 1400 | 1800 | ND | ND |
| *VrrB1* | 8800 | 8600 | 8800 | 8400 | 8500 | 8400 | 8500 | 8500 | 2400 | 1100 | ND | 900 |
| *VrrB2* | 9000 | 9000 | 8800 | 8500 | 1500 | 8500 | 50 | 3000 | ND | 1100 | ND | 3000 |
| *VrrC1* | Overloaded | Overloaded | Overloaded | 8600 | 8700 | 8500 | 8500 | 8600 | 2600 | 1500 | ND | 300 |
| *VrrC2* | Overloaded | 8700 | Overloaded | 9000 | Overloaded | 8700 | 8000 | 8700 | ND | 2000 | ND | 700 |
| cG3 | 8600 | 8600 | 8600 | 8400 | 8700 | 8700 | 6500 | 5800 | 1000 | 900 | ND | ND |
| pX01 | 8800 | Overloaded | 8600 | Overloaded | 8200 | Overloaded | 4500 | 8600 | 1000 | 1100 | ND | 68 |
| pX02 | 8700 | 8900 | 8700 | 5000 | 8000 | 800 | 1800 | ND | ND | ND | ND | ND |

# *Microarray*

**Hybridization of *B. anthracis* and *B. thuringiensis* DNAs on the Virulence Array.** We isolated genomic DNAs from *B. anthracis* Ames, Sterne and A0382 and DNAs from *B. thuringiensis* israelensis HD500, kurstaki ATCC 33679. The DNAs were hybridized to the Virulence Array. The data is shown in Table 23 The Virulence Array corrected identified each of the DNA to at least the species level. For *B. thuringiensis*, the identification was at the strain level. The probes for this array were designed for species level identification based the detection of probes designed on known virulence genes and antibiotic resistance genes. So it is not unexpected that the *B. anthracis* strains were not correctly detected to the strain level.

Table 23. Virulence Array results from *B. anthracis* and *B. thuringiensis* DNAs

| DNA hybridized on array | Virulence Array top hit |
|---|---|
| *B. anthracis* Ames | *B. anthracis* USA6153 |
| *B. anthracis* Sterne | *B. anthracis* A0174 |
| *B. anthracis* A0382 | *B. anthracis* USA6153 |
| *B. thuringiensis* israelensis HD500 | *B. cereus* G9842<br>*B. thuringiensis* israelensis plasmid pBtoxis |
| *B. thuringiensis* kurstaki ATCC 33679 | *B. cereus* ATCC14579<br>*B. thuringiensis* kurstaki<br>*B. thuringiensis* plasmid pBMB67 |

**Determination of the limit of detection of the Virulence Array using *B. anthracis* Ames spiked into BioWatch aerosol samples.** We performed limit of detection testing of the Virulence Array using serially diluted *B. anthracis* Ames spiked into BioWatch aerosol filter extracts that have been subjected to whole genome amplification. Duplicate experiments were run to ensure repeatability and data consistency. 1, 10, 100, 1,000, 10,000, and 100,000 copies of *B. anthracis* Ames were tested. Table 24 below shows results of the Virulence Array at each of the *B. anthracis* DNA concentration. When 10 copies of *B. anthracis* DNA were spiked into the aerosol sample, only one of the two replicate experiments detected *Bacillus cereus*, a very close near neighbor to *B. anthracis*, suggesting that there were not enough probes specific to *B. anthracis* detected at this concentration. This experiment suggested that our detection limit for *B. anthracis* Ames is at 100 genome copies per 100 pg of aerosol DNA sample.

Table 24. Limit of detection of B. anthracis Ames DNA spiked in Biowatch aerosol samples.

| **Amount aerosol filter DNA** | 100 pg | 100 pg | 100 pg | 100 pg | 100 pg | 100 pg |
|---|---|---|---|---|---|---|
| **Amount *B. anthracis* DNA** | 560 pg | 56 pg | 5.6 pg | 560 fg | 56 fg | 5.6 fg |
| ***B. anthracis* DNA Copy #** | 100,000 copies | 10,000 copies | 1000 copies | 100 copies | 10 copies | 1 copy |
| **% BA DNA in aerosol DNA** | 98.2% | 35.9% | 5.3% | 0.56% | 0.06% | 0.006% |
| **Virulence Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. cereus* | Not detected |

**Determination of the limit of detection of the Virulence Array using *B. anthracis* Ames spiked into soil samples.** We performed as similar limit of detection testing of the Virulence Array using serially diluted *B. anthracis* Ames spiked into soil extracts that have been subjected to whole genome amplification. The soil was a combination of soils collected locally in San Francisco and Oakland. Duplicate experiments were run to ensure repeatability and data consistency. 1, 10, 100, 1,000, 10,000, and 100,000 copies of *B. anthracis* Ames were tested. Table 25 below shows results of the Virulence Array at each of the *B. anthracis* DNA concentration. When 100 copies of *B. anthracis* DNA were spiked into soil sample, only one of the two replicate experiments detected *Bacillus cereus*, a very close near neighbor to *B. anthracis*, suggesting that there were not enough probes specific to *B. anthracis* detected at this concentration. This experiment suggested that our detection limit for *B. anthracis* Ames is 1000 genome copies per 1 ng of soil DNA sample.

Table 25. Limit of detection of B. anthracis Ames DNA spiked in soil samples.

| **Amount soil DNA** | 1 ng | 1 ng | 1 ng | 1 ng | 1 ng | 1 ng |
|---|---|---|---|---|---|---|
| **Amount *B. anthracis* DNA** | 560 pg | 56 pg | 5.6 pg | 560 fg | 56 fg | 5.6 fg |
| ***B. anthracis* DNA Copy #** | 100,000 copies | 10,000 copies | 1000 copies | 100 copies | 10 copies | 1 copy |
| **% BA DNA in soil DNA** | 35.9% | 5.3% | 0.56% | 0.06% | 0.006% | 0.0006% |
| **Virulence Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. cereus** | Not detected | Not detected |

* Only one of two replicated detected *B. cereus*.

**Analysis of DNA samples from previous DHS or EPA exercises to release and detect *B. thuringiensis* spores.** Aerosol filters were collected during the days when *B. thuringiensis* kurstaki was sprayed to control Gypsy Moth in a DHS surrogate study. Gauge wipes were used to wipe dirty indoor surfaces and inoculated with *B. thuringiensis* kurstaki spores in an exercise conducted with EPA. Genomics DNAs were extracted from filters or wipe samples and run on the Virulence Array. The results are shown in Table 26. We were able to positively identify *B. thuringiensis* kurstaki using the Virulence Array from both the air filter samples and the gauge wipe samples.

Table 26. Detection of *B. thuringiensis* from environmental air or wipe samples

| **Sample** | **Air filters collected around Gypsy Moth control study** | **Gauge wipes collected during an EPA exercise** |
|---|---|---|
| *B. thuringensis* kurstaki specific TaqMan assay | Average Ct = 22.77 ± 0.27 | Average Ct = 30.70 ± 0.31 |
| Virulence Array top hits | *B. cereus* ATCC 14579<br>*B. thuringiensis* kurstaki<br>*Burkholderia phymatum* STM815<br>*Ralstonia pickettii* 12J<br>*P. aeruginosa* LES | *B. cereus* ATCC 14579<br>*B. thuringiensis* kurstaki<br>*Delftia acidovorans*<br>*S. aureus* str. JKD6009 |

**Hybridization of *B. anthracis* and *B. thuringiensis* DNAs on the Census Array.** The primary goal of array analysis was to identify, for each sample, the organism(s) with known genomic sequence that best explains the pattern of bright (detected) and dark (undetected) probes on the array. Some of the organisms we tested have not been sequenced; for these, our measure of success is whether the analysis identifies the correct species (when other strains of the same species have been sequenced). We isolated genomic DNAs from *B. anthracis* Ames, Sterne and A0382 and DNAs from *B. thuringiensis* israelensis HD500, kurstaki ATCC 33679. The

DNAs were hybridized to the Census Array. The data is shown in Table 27. The Census Array corrected identified each of the DNA to at least the species level. For *B. thuringiensis*, the identification was at the strain level.

*B. anthracis* isolates were not correctly identified to the strain level on the Census Array. Our analysis algorithm scores targets by adding contributions from probe hits to target genomic sequences. Although our target database maintains separate entries for finished chromosome and plasmid sequences, a typical draft genome sequence (such as the one for *B. anthracis* A0155) consists of "glued" contigs from both the chromosome and plasmids. In a species with very little genetic diversity such as *B. anthracis*, the score contributions from probes that hit the plasmid sequences in glued genomes outweigh the score penalties due to mismatches against the chromosome sequences. In this case, the algorithm assigns higher scores to the draft genome targets. Our team will address this issue by splitting glued genome sequences in the target database into chromosome and plasmid components, so that they can be compared correctly against finished sequences by our analysis algorithm.

Table 27. Census Array results from *B. anthracis*, *B. thuringiensis* and *F. tularensis* DNAs

| **DNA hybridized on array** | **Census Array top hit** |
|---|---|
| *B. anthracis* Ames | *B. anthracis* A0155,<br>*B. anthracis* Sterne |
| *B. anthracis* Sterne | *B. anthracis* A0155 |
| *B. anthracis* A0382 | *B. anthracis* A0155 |
| *B. thuringiensis* Israelensis HD500 | *B. thuringiensis* IBL4222 |
| *B. thuringiensis* Kurstaki ATCC 33679 | *B. thuringiensis* kurstaki |
| *F. tularensis* holarctica LVS | *F. tularensis* LVS |

**Evaluation of mixtures of biothreat bacterial samples.** Genomic DNAs from *B. anthracis* Ames, *Y. pestis* CO92, *F. tularensis* LVS, *Brucella abortis*, *B. pseudomallei* PHLS9 and *B. mallei* 23344 were mixed together in one single sample and hybridized on the 388K Census Array. The results are shown in Table 28.

All six bacterial species were correctly identified. *B. mallei was* detected as a secondary hit to *B. pseudomallei* in the Burkholderiaceae family using our microarray analysis. *F. tularensis* LVS was correctly identified at the strain level.

Table 28. Census Array results from a mixture of bacterial DNAs

| Actual species | Actual strain | Predicted species | Predicted strain |
|---|---|---|---|
| *B. anthracis* | Ames | *B. anthracis* | A0193 |
| *Y. pestis* | CO92 | *Y. pestis* | PEXU2 |
| *F. tularensis* | holarctica LVS | *F. tularensis* | LVS |
| *B. pseudomallei* | PHLS9 | *B. pseudomallei* | E208 |
| *B. mallei* | ATCC 23344 | *B. mallei* | GB8 |
| *Brucella* | abortis | *Brucella* | melitensis |

**Determination of the limit of detection of the Census Array using *B. anthracis* Ames spiked into BioWatch aerosol samples.** We performed limit of detection testing of the Census Array using serially diluted *B. anthracis* Ames spiked into BioWatch aerosol filter extracts that have been subjected to whole genome amplification. Duplicate experiments were run to ensure repeatability and data consistency. 1, 10, 100, 1,000, 10,000, and 100,000 copies of *B. anthracis* Ames were tested. Table 29 below shows results of the Census Array at each of the *B. anthracis* DNA concentration. When 10 copies of *B. anthracis* DNA were spiked into aerosol sample, only one of the two replicate experiments detected *B. anthracis*, suggesting that there were not enough probes specific to *B. anthracis* detected at this concentration. This experiment suggested that our detection limit for *B. anthracis* Ames is 100 genome copies when the DNA was spiked into 100 pg of aerosol DNA sample. Table 30 outlines a comparison of the aerosol spiked samples tested with MLVA, Real Time PCR, and Microarray. All methods detected B. anthracis when >100 copies were spiked into the aerosol samples, only the MLVA and Real Time PCR assays were able to detect the *B. anthracis* at fewer than 100 copies.

Table 29. Limit of detection of *B. anthracis* Ames DNA spiked in BioWatch aerosol samples

| **Amount aerosol filter DNA** | 100 pg | 100 pg | 100 pg | 100 pg | 100 pg | 100 pg |
|---|---|---|---|---|---|---|
| **Amount *B. anthracis* DNA** | 560 pg | 56 pg | 5.6 pg | 560 fg | 56 fg | 5.6 fg |
| ***B. anthracis* DNA Copy #** | 100,000 copies | 10,000 copies | 1000 copies | 100 copies | 10 copies | 1 copy |
| **% BA DNA in aerosol DNA** | 98.2% | 35.9% | 5.3% | 0.56% | 0.06% | 0.006% |
| **Census Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* | Not detected |

Table 30. Comparison of detection limits *B. anthracis* Ames DNA aerosol samples

| | | | | | | |
|---|---|---|---|---|---|---|
| **Amount aerosol filter DNA** | 100 pg | 100 pg | 100 pg | 100 pg | 100 pg | 100 pg |
| **Amount *B. anthracis* DNA** | 560 pg | 56 pg | 5.6 pg | 560 fg | 56 fg | 5.6 fg |
| ***B. anthracis* DNA Copy #** | 100,000 copies | 10,000 copies | 1000 copies | 100 copies | 10 copies | 1 copy |
| **% BA DNA in aerosol DNA** | 98.2% | 35.9% | 5.3% | 0.56% | 0.06% | 0.006% |
| **MLVA results*** | 8/8 pos | 8/8 pos | 8/8 pos | 8/8 pos | 5/8 pos | 3/8 pos |
| **Canonical SNP results**** | Ct=21.27 | Ct=25.05 | Ct=28.38 | Ct=31.87 | Ct=36.26 | neg |
| **Virulence Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. cereus**** | Not detected |
| **Census Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* **** | Not detected |

* Eight MLVA (Multi-locus Variable Number Tandem Repeat Analysis) markers were tested and the samples that have the correct sized markers were called positive. (Keim, *et al*. J. Bacteriol. **182**, 2928-2936.)

** Taqman assays that discriminate thirteen canonical SNPs on the major branches of *B. anthracis* phylogenetic tree were designed and tested. (Van Ert *et al*. 2007, PLoS ONE). The average Ct from two replicate samples on SNP A. Br004 was listed on the table.

****B. cereus* was detected in one of two replicate arrays.

*****B. anthracis* was detected in one of two replicate arrays.

**Determination of the limit of detection of the Census Array using *B. anthracis* Ames spiked into soil samples.** We performed as similar limit of detection testing of the Census Array using serially diluted *B. anthracis* Ames spiked into soil extracts that have been subjected to whole genome amplification. The soil was a combination of soils collected locally in San Francisco and Oakland. Duplicate experiments were run to ensure repeatability and data consistency. 1, 10, 100, 1,000, 10,000, and 100,000 copies of *B. anthracis* Ames were tested. Table 31 below shows results of the Census Array at each of the *B. anthracis* DNA concentration. When 100 copies of *B. anthracis* DNA were spiked into soil sample, only one of the two replicate experiments detected *B. anthracis*, suggesting that there were not enough probes specific to *B. anthracis* detected at this concentration. This experiment suggested that our detection limit for *B. anthracis* Ames is 1000 genome

copies when the DNA was spiked into 1 ng of soil DNA sample. Table 32 shows all methods identifying B. anthracis at 1000 copies and more, while MLVA was able to detect the DNA at 10 copies.

Table 31. Limit of detection of *B. anthracis* Ames DNA spiked in soil samples

| **Amount soil DNA** | 1 ng | 1 ng | 1 ng | 1 ng | 1 ng | 1 ng |
|---|---|---|---|---|---|---|
| **Amount *B. anthracis* DNA** | 560 pg | 56 pg | 5.6 pg | 560 fg | 56 fg | 5.6 fg |
| ***B. anthracis* DNA Copy #** | 100,000 copies | 10,000 copies | 1000 copies | 100 copies | 10 copies | 1 copy |
| **% BA DNA in soil DNA** | 35.9% | 5.3% | 0.56% | 0.06% | 0.006% | 0.0006% |
| **Census Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis** | Not detected | Not detected |

*only 1 of 2 replicates detected

Table 32. Comparison of detection limits *B. anthracis* Ames DNA aerosol samples

| **Amount soil DNA** | 1 ng | 1 ng | 1 ng | 1 ng | 1 ng | 1 ng |
|---|---|---|---|---|---|---|
| **Amount *B. anthracis* DNA** | 560 pg | 56 pg | 5.6 pg | 560 fg | 56 fg | 5.6 fg |
| ***B. anthracis* DNA Copy #** | 100,000 copies | 10,000 copies | 1000 copies | 100 copies | 10 copies | 1 copy |
| **% BA DNA in soil DNA** | 35.9% | 5.3% | 0.56% | 0.06% | 0.006% | 0.0006% |
| **MLVA results*** | 8/8 pos | 8/8 pos | 8/8 pos | 8/8 pos | 5/8 pos | neg |
| **Canonical SNP results**** | Ct=24.90 | Ct=27.43 | Ct=31.09 | Ct=35.11 | neg | neg |
| **Virulence Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. cereus**** | Not detected | Not detected |
| **Census Array top hit** | *B. anthracis* | *B. anthracis* | *B. anthracis* | *B. anthracis* **** | Not detected | Not detected |

* Eight MLVA markers were tested and the samples that have the correct sized markers were called positive.
** Taqman assays that discriminate thirteen canonical SNPs on the major branches of *B. anthracis* phylogenetic tree were designed and tested. The average Ct from two replicate samples on SNP A. Br004 was listed on the table.
***B. cereus* was detected in one of two replicate arrays.
****B. anthracis* was detected in one of two replicate arrays.

**Analysis of DNA samples from previous DHS or EPA exercises to release and detect *B. thuringiensis* spores.** Aerosol filters were collected during the days when *B. thuringiensis* kurstaki was sprayed to control Gypsy Moth in a DHS surrogate study. Gauge wipes were used to wipe dirty indoor surfaces and inoculated with *B. thuringiensis* kurstaki spores in an exercise conducted with EPA. Genomics DNAs were extracted from filters or wipe samples and run on the Census Array. The results are shown in Table 33. We were able to positively identify *B. thuringiensis* kurstaki using the Census Array from both the air filter samples and the gauge wipe samples.

Table 33. Detection of *B. thuringiensis* from environmental air or wipe samples

| **Sample** | **Air filters collected around Gypsy Moth control study** | **Gauge wipes collected during an EPA exercise** |
|---|---|---|
| *B. thuringiensis* kurstaki specific TaqMan assay | Average Ct = 22.77 ± 0.27 | Average Ct = 30.70 ± 0.31 |
| Census Array top hits | *B. thuringiensis* kurstaki<br>*Bacillus megaterium QM B1551*<br>*Magnetospirillum Magnetotacticum MS−1*<br>*Thioalkalivibrio sp.*<br>*Ralstonia pickettii 12D* | *B. thuringiensis* kurstaki<br>*Magnetospirillum magnetotacticum MS−1*<br>*Alkalilimnicola ehrlichei*<br>*Tolumonas auensis DSM* |

# *Illumina and 454 Sequencing*

**Soil and Aerosol Spiked Samples.**

For Illumina sequencing, the *B. anthracis* soil spiked samples were sequenced with 51 cycles of paired end reads, while aerosol spiked samples were sequenced with 51 cycles of single end reads. For the purposes of analysis, the paired end reads were decoupled (pairing information ignored) and used as if single reads were generated. As a result, soil spikes samples have roughly double the amount of

sequence reads (decoupled singletons) generated. It should be noted that while the number of sequencing reads for soil samples is doubled, the number of independent samples (shots) from each sample is not. Since the paired end read is typically generated from a single sequence fragment (chimeric reads issues none withstanding) the sensitivity of the soil samples, in terms of the ability to detect rare environmental organisms, is not significantly increased when compared to aerosol samples.

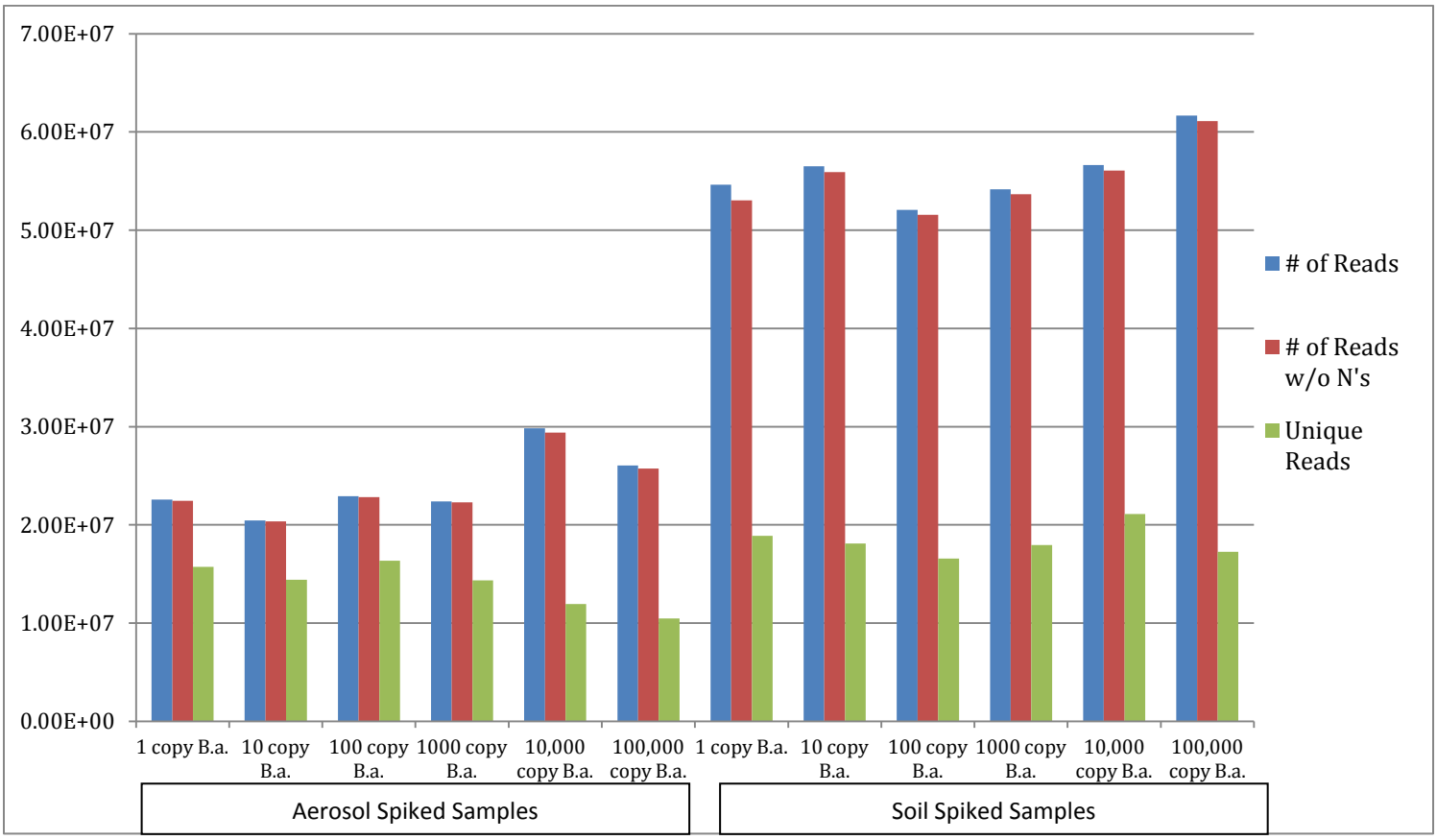



Figure 9. The total number of **Illumina sequencing** reads obtained for each spiked sample. The data is available in Appendix 1.

We chose two approaches to examine the sequence content of the 454 sequence data derived from spiked soil and aerosol samples. The first approach utilized the vendor's software. The vendor, Roche Scientific, provides software (gsMapper) that will map 454 reads to a reference sequence(s). The reference sequence set can be a combination of different sequences in standard Fasta format. The mappings were done using a 98% minimum identity as a cutoff so that reads mapping with less that 98% identity were not considered. The second approach utilized standard database search methods. We used the NCBI taxonomy database, NCBI blast software (megablast) and the NCBI nucleotide database. Because the nucleotide database can be cross referenced to the taxonomy database we could use

a best hit approach to compile a list of organisms from the taxonomy database to understand the DNA composition of the samples.

In each approach we had to consider the variance in the number of reads generated by the sequencing runs on each of the eight samples. The number of reads obtained for each sample is presented in Figure 10. In general, normalization was performed by using the percent of the reads of interest in a set to the total number of reads in that set. When this normalization strategy was not used, the strategy used is described in that section.

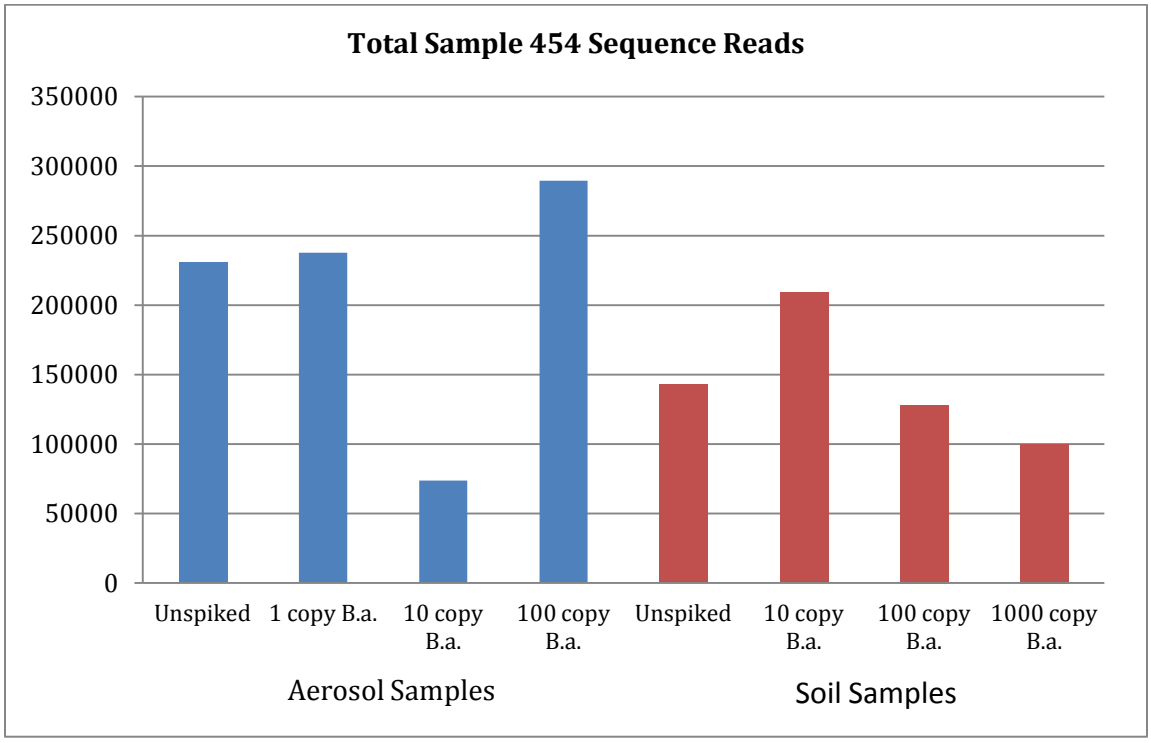



Figure 10. The total number of **454 sequencing** reads obtained for each environmental sample. The data is available in Appendix 4.

**Results of the Read Mapping Analysis.** Illumina reads for 6 aerosol samples (Each 100pg aerosol sample spiked with 1, 10, 100, 1000, 10,000, or 100,000 copies *Bacillus anthracis* Ames) and 6 soil samples (Each 1ng soil sample spiked with 1, 10, 100, 1000, 10,000, or 100,000 copies *Bacillus anthracis* Ames) along with 454 reads in 4 aerosol samples (Each 100pg aerosol sample was unspiked or spiked with 1, 10, 100 copies *Bacillus anthracis* Ames) and 4 soil samples (Each 1ng soil sample was unspiked or spiked with 10, 100, 1000 copies *Bacillus anthracis* Ames) were mapped to the GenBank entries listed in Table 34.

Table 34. List of GenBank Organisms used during the analysis

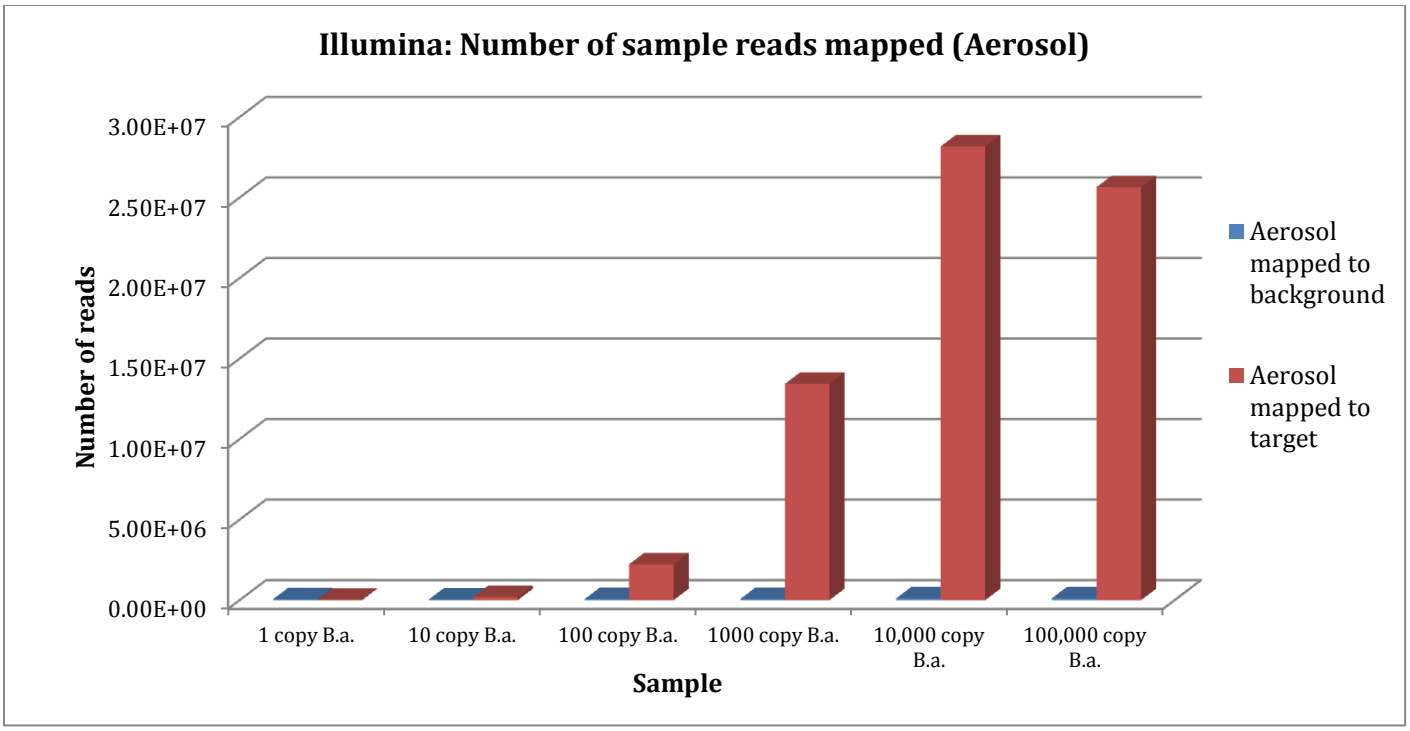
| | GenBank Definition | GenBank Accession |
|---|---|---|
| | **Target Reference Genomes** | |
| 1 | Bacillus thuringiensis str. Al Hakam, complete genome | CP000485.1 |
| 2 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence | CP000486.1 |
| 3 | Bacillus cereus biovar anthracis str. Cl, complete genome | CP001746.1 |
| 4 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-XO1, complete sequence | CP001747.1 |
| 5 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-XO2, complete sequence | CP001748.1 |
| 6 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14, complete sequence | CP001749.1 |
| 7 | Bacillus anthracis str. Ames, complete genome | AE016879 |
| 8 | Bacillus anthracis virulence plasmid PX01, complete sequence | AF065404.1 |
| 9 | Bacillus anthracis plasmid pXO2, complete sequence | AF188935.1 |
| | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence | |
| | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence | |
| | Bacillus anthracis str. Sterne, complete genome | |
| | **Background Reference Genomes** | |
| 10 | Burkholderia pseudomallei strain K96243, chromosome 1, complete sequence | BX571965.1 |
| 11 | Escherichia coli O157:H7 EDL933, complete genome | AE005174.2 |
| 12 | Francisella tularensis subsp. Tularensis SCHU S4 complete genome | AJ749949.2 |
| 13 | Pseudomonas aeruginosa PAO1, complete genome | AE004091.2 |
| 14 | Rhodopseudomonas palustris CGA009 complete genome | BX571963.1 |
| 15 | Sinorhizobium meliloti 1021 complete chromosome | AL591688.1 |
| 16 | Staphylococcus aureus subsp. Aureus N315 DNA, complete genome | BA000018.3 |
| 17 | Streptomyces coelicolor A3 (2) complete genome | AL645882.2 |
| 18 | Yersinia pestis CO92 complete genome | AL590842.1 |
| 19 | Bacillus subtilis subsp. Subtilis str. 168 complete genome | AL009126.3 |
| 20 | Clostridium botulinum A str. Hall, complete genome | CP000727.1 |

Target reference genomes are meant to be used to identify *B. anthracis* reads in each sample or to understand specificity of the analysis strategy. Background reference genomes were used to get a relative understanding of sample to sample variation. Red- 454 Sequencing only, Green- Illumina Sequencing only.

**Comparison of BWA mapping results between *B. anthracis* Sterne and a background set of organisms using the BWA read mapping software.** The first analysis was performed by using the mapping software BWA for Illumina and Roche software for 454 Sequencing to map the reads to two reference sets. The first reference set was the *B. anthracis Ames* chromosome, and the pXO1 and pXO2 plasmids (Target Reference Genome, Table 34). The second set was eleven finished genomes in GenBank arbitrarily chosen to represent our background set (Background Reference Genomes in Table 34).

The differences between the relative number of Illumina reads mapping to the reference genome sets from the *B. anthracis* spiked aerosol and soil samples are shown in Figure 11a. We see the relative number of reads in each soil sample that map to the *Bacillus anthracis* chromosome and plasmids increases from the 1 copy *B. anthracis* to the 100,000 copy sample in Illumina (in the unspiked to the 100 copy *B. anthracis* spiked sample for 454), and there is no accompanying increase in the number of corresponding hits to a select set of background species. Similar ordering is observed in the Aerosol samples. In particular, when considering total number of Illumina reads mapped (Figure 11a), the number of reads mapped to the target increases from the 1 copy to the 10,000 copy sample and slightly reduces in 100,000 copy sample. This ordering is mostly likely an artifact of 10,000 *B. anthracis* copy aerosol spiked sample having higher number of total reads. This behavior is not observed when mapped reads are normalized by totals (Figure 11b).

The average fold increase in total number of Illumina reads mapped to BA is 5.5 for Soil and 7.5 for Aerosol samples, BUT the fold increase varies from 0.9 to 11.4 in soil and 3.3 to 16.6 in Aerosol. Coverage levels ranged from 0.2X to 245X for Aerosol and 0.02X to 430X for soil samples.



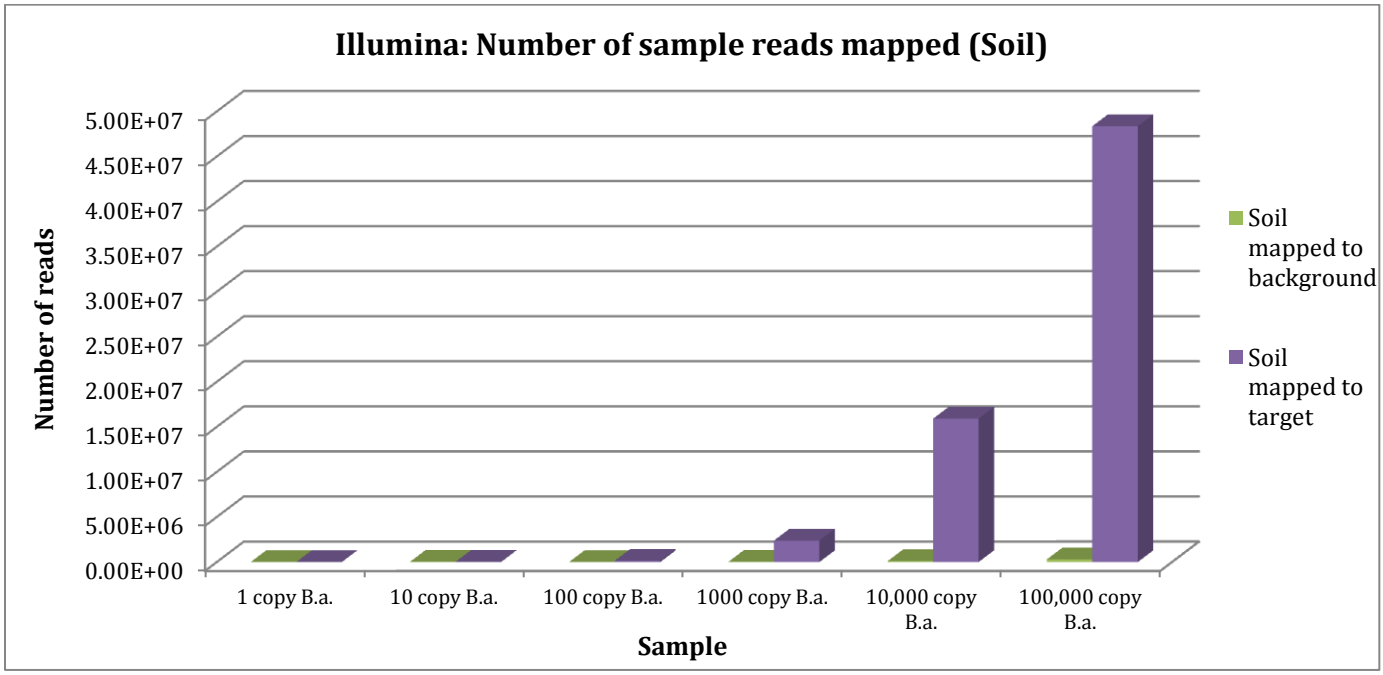
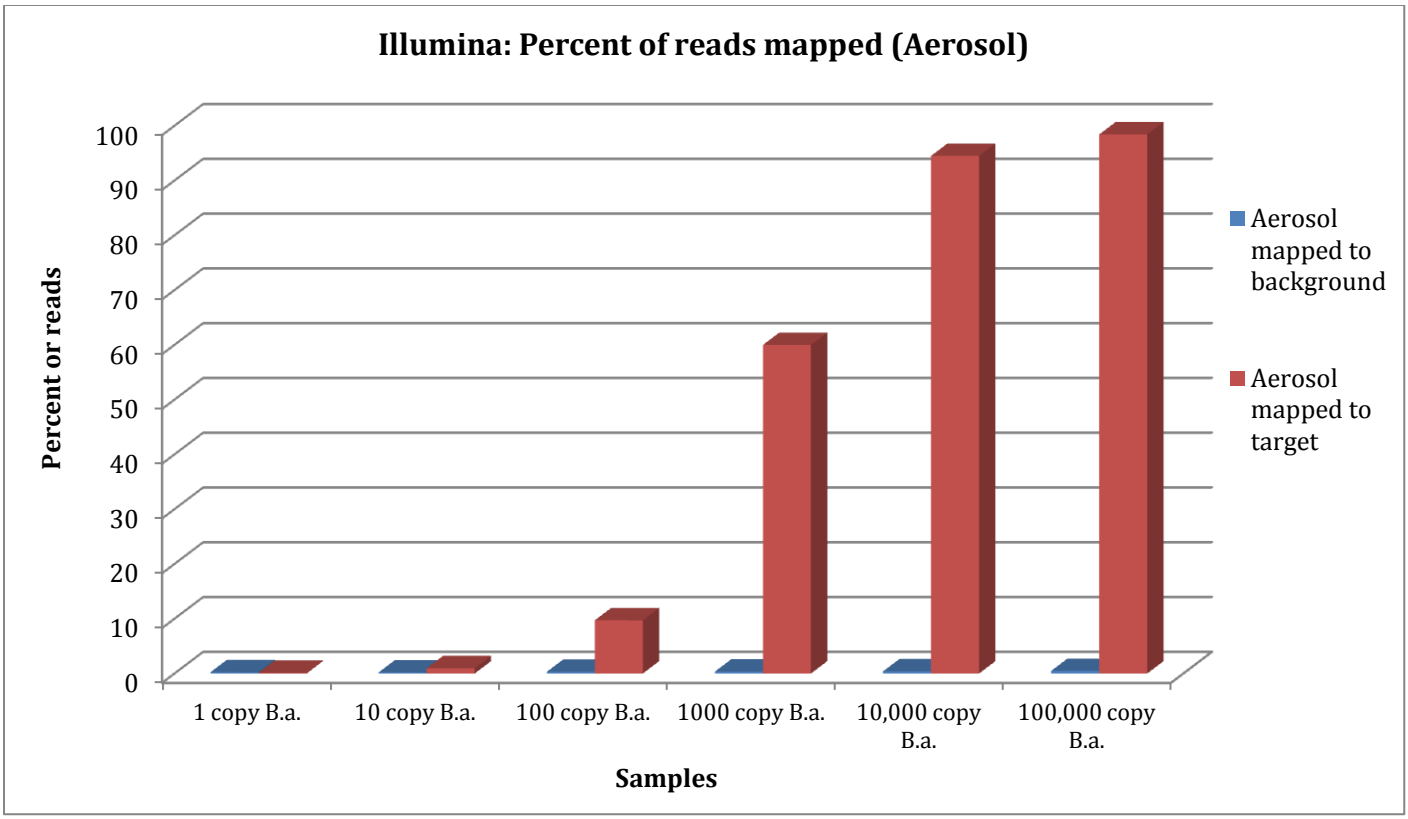




Figure 11a. The number of **Illumina** reads that mapped (up to 3 MM) to either *B. anthracis* or a set of non *B. anthracis* (background) is shown for each soil and aerosol read set over the 6 spiked conditions.



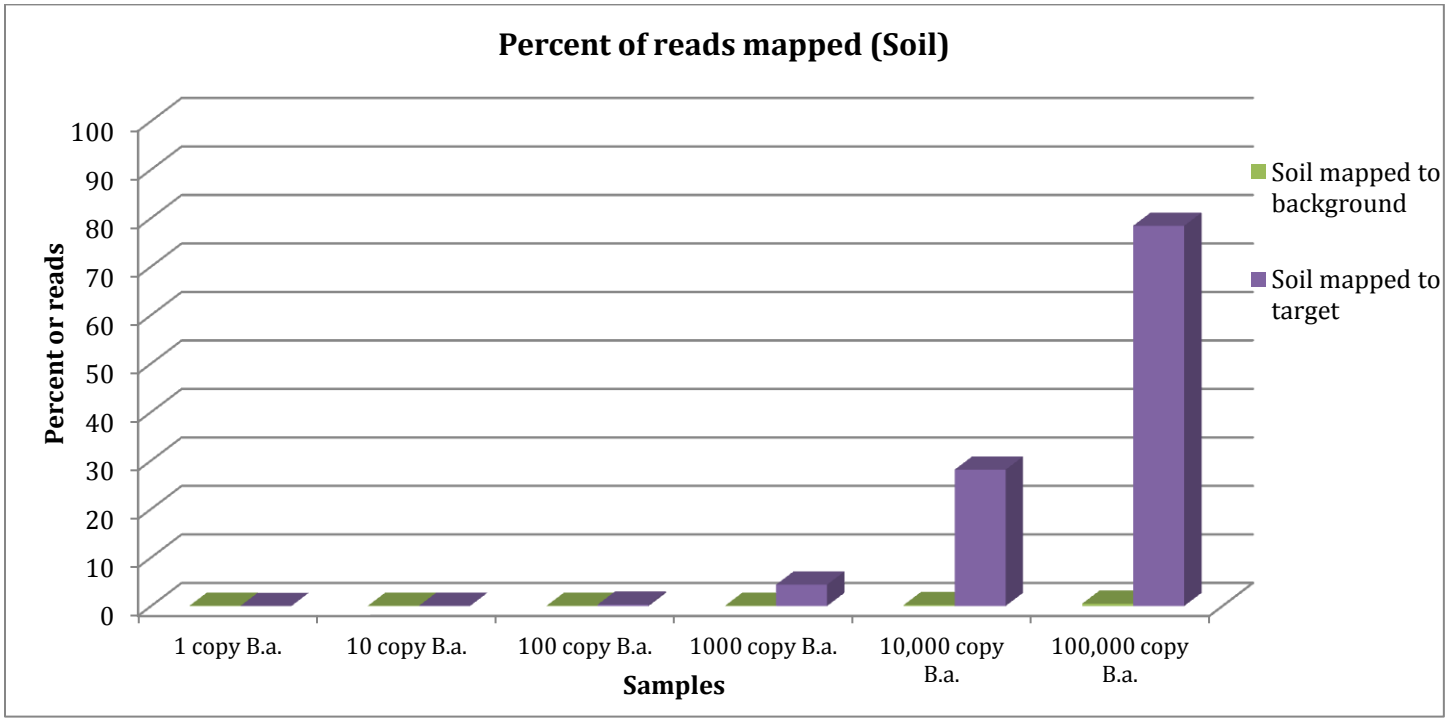
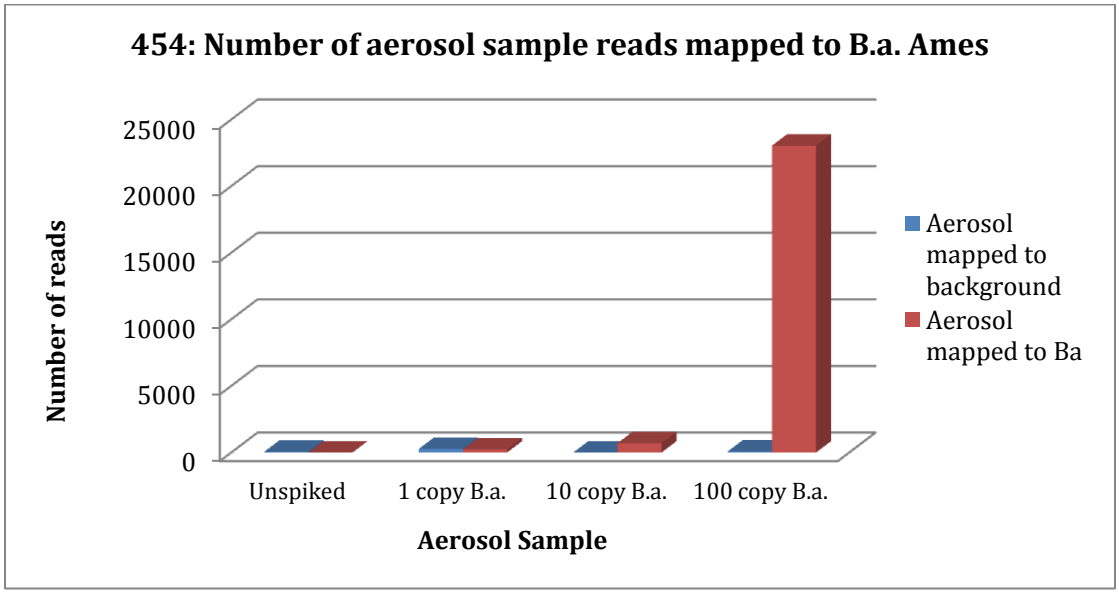
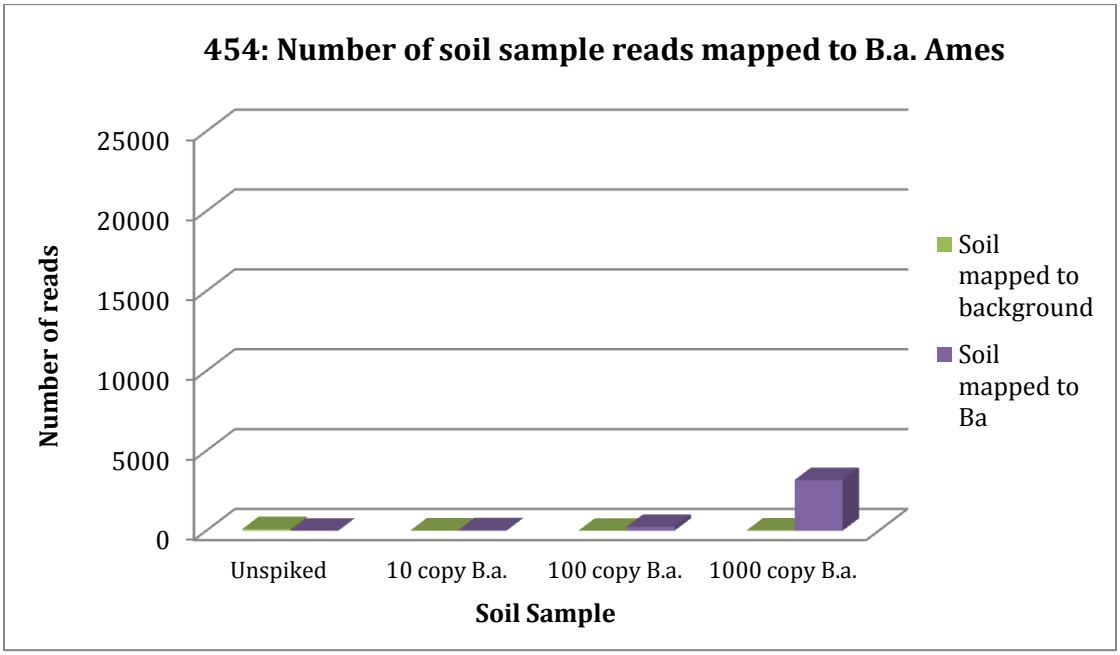
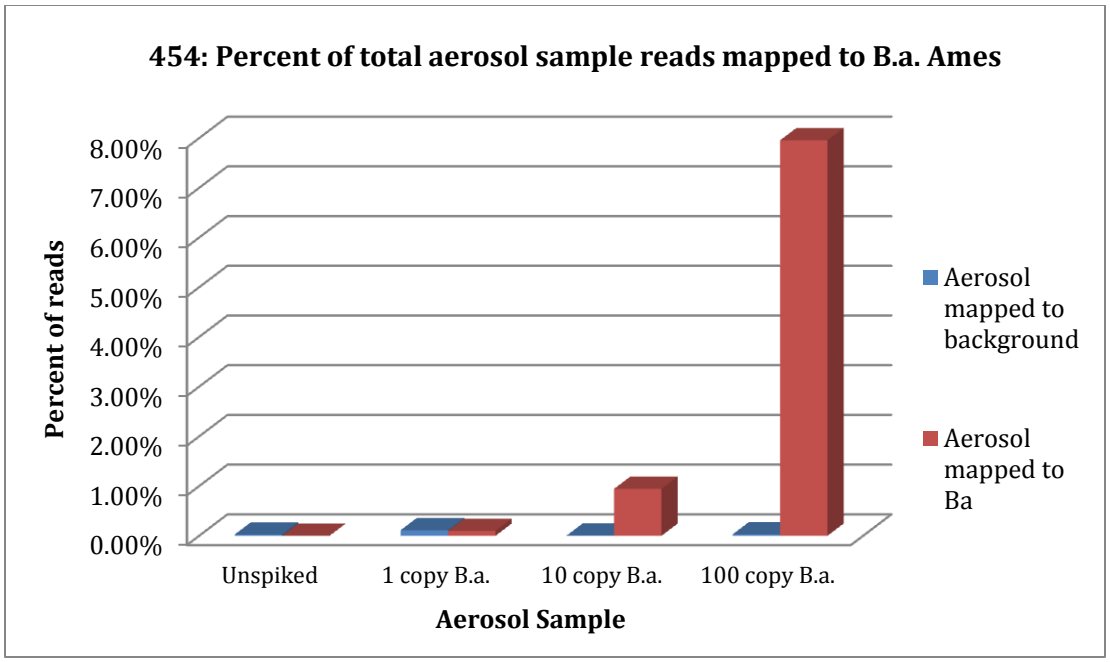




Figure 11b. The proportion of **Illumina** reads (mapped normalized by totals) that mapped (up to 3 MM) to either *B. anthracis* or a set of non *B. anthracis*.(background) is shown for each soil and aerosol read set over the 6 spiked conditions.

In 454 Sequencing, significant increase in the number of reads mapped to *B. anthracis* was observed in the sample spiked with 100 copies of Ba Ames in 100 pg of aerosol DNA and in the sample where 1000 copies Ba Ames DNA was spiked in 1 ng of soil DNA. When plotted using the percent of reads mapped to Ba Ames, about 1% of reads from 10 copy aerosol sample was mapped to Ba Ames, and 8% in the 100 copy aerosol sample. In soil samples, about 0.2% of the 100 copy sample was mapped to Ba Ames and 3% mapped in 1000 copy sample. The effect of normalizing reads mapped as a percent of the total reads can be seen by comparing the two graphs in Figure 12.





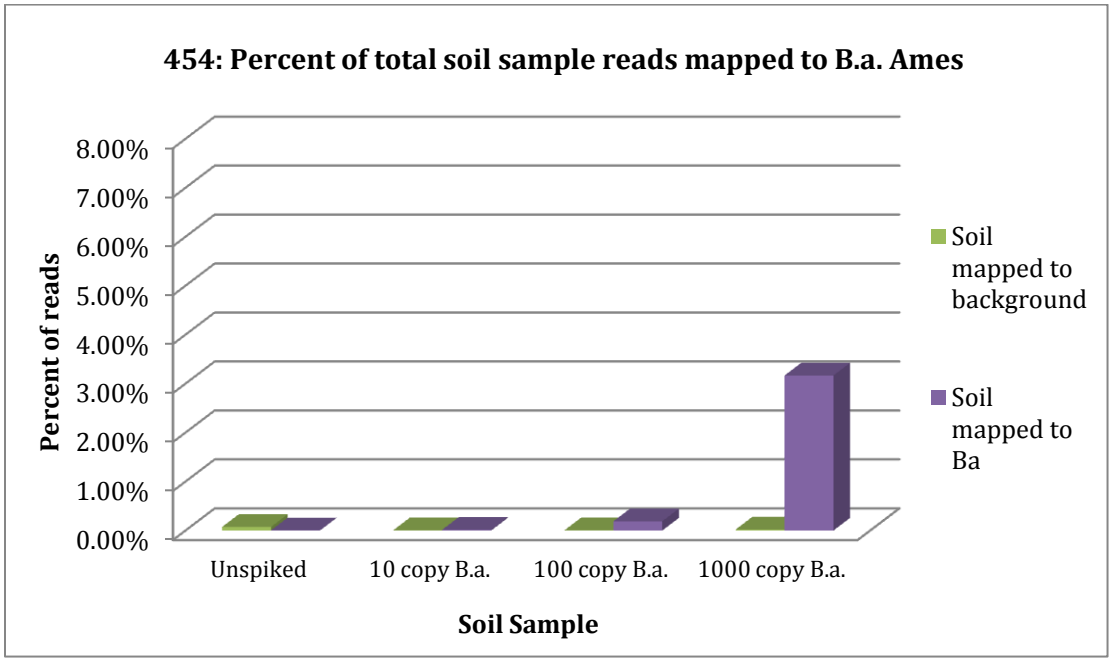




Figure 12. In the top 2 charts, the number of 454 reads that mapped to either B. a. or a set of non B. a. (background) is shown for each soil and aerosol read set over the 4 spiked conditions. In the bottom 2 charts, the number of 454 reads that mapped to either B. a. or a set of non B. a. (background) was divided by the total number to normalize for the variation in the number of reads in each sequence set. The data is available in Appendix 5.

**Mapping specificity using the BWA read mapping software.** The specificity of the mapping was examined in two studies. In each study all the reads were mapped to both the *B. anthracis* str. Ames genome supplemented with a pXO2 plasmid, and a close relative. In the first study, the *B. thuringiensis* str. Al Hakam genome was used as the close relative, and in the second study the *B. cer*eus biovar anthracis str. CI genome was used as the close relative.

For Illumina reads, the top-hit only approach was attempted. However, due to significant sequence similarity between target organism and its close relatives, the ability of the reads to discriminate between true positive and false positive identifications was limited. We have employed a modified approach where the reads that mapped to only one of the organisms (target or its close relative) were considered, the reads that mapped to both references or to neither reference, were not considered to calculate specificity. This approach showed greater discriminatory power and we believe is more appropriate given the short length of Illumina GAIIx reads.

In each 454 sequencing study, the sequence reads that mapped to multiple sequences are not considered; only uniquely mapped reads were counted.

**Mapping specificity using a reference sequence set of *B. anthracis* (Sterne chromosome and pXO1 and pXO2 plasmids) and the near-neighbor *B. thuringiensis* Al Hakam chromosome and plasmid.** The decoupled Illumina reads from each sample (6 aerosol and 6 soil samples) were mapped to the reference sequence set and the resulting data is shown in Figures 13 and 14. While in both soil and aerosol samples, the proportion of reads (out of total) that could be mapped to *B. anthracis Str Ames* was consistently higher than proportion of reads mapped to *B thuringiensis str Al Hakam*, that the difference in proportion of mapped reads was not large (up to 17% difference, maximum). Discriminatory power of using Illumina GAIIx reads is limited (Figure 15a). This is especially true for samples with low amounts of target organism spiked-in (i.e. 1, 10, and 100 copy spiked samples).

It is important to note that due to high degree of sequence similarity between *B. anthracis Str Ames* and *B. thuringiensis str Al Hakam*, a large portion of reads will map to both genomes simultaneously. These are the reads that make it difficult to distinguish between the two genomes. One solution is to only consider the reads mapping onto the portions of the reference genomes that are different from each other (i.e. differentiate between the genomes using only the portions of the genomes that matter). This approach, as can be seen from Figure 15b, is significantly better at discriminating between close relatives than the original approach.

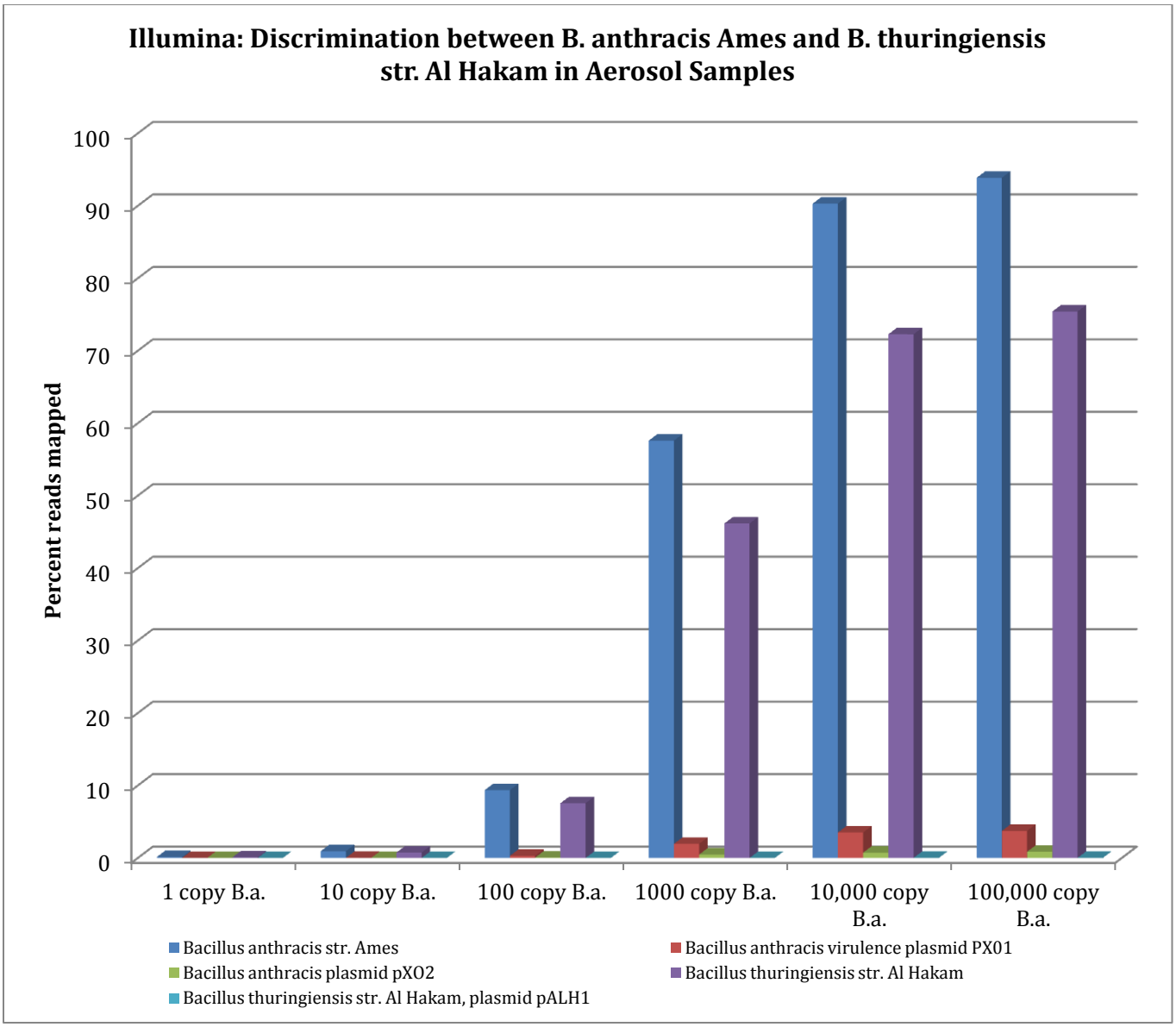



Figure 13. The percent of Illumina reads from the aerosol samples that mapped to the reference set. The data is available in the Appendix 3.

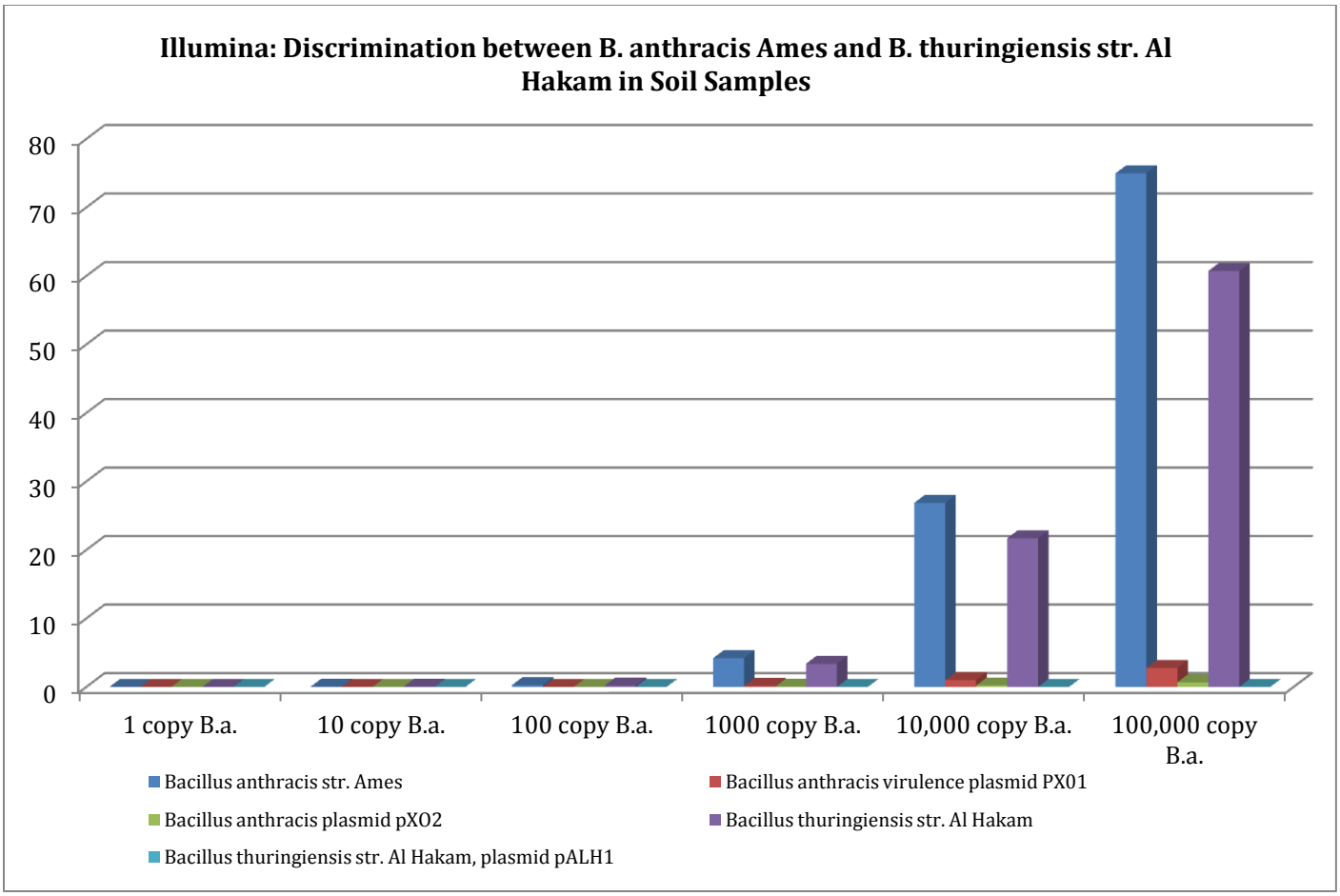
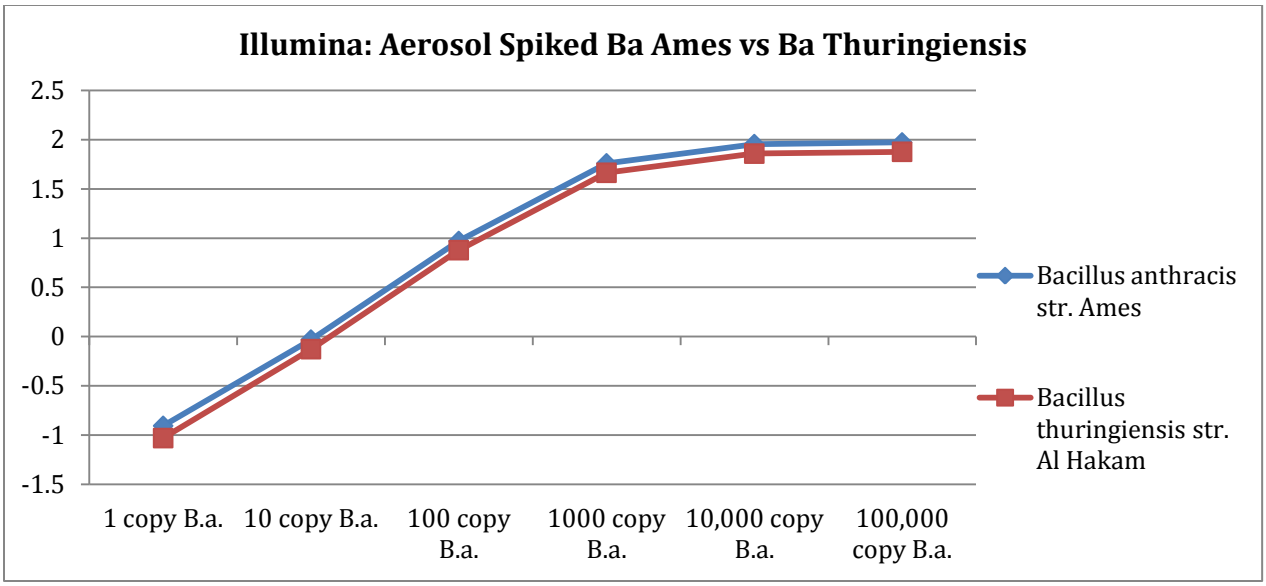



Figure 14. The percent of Illumina reads from the soil samples that map to the sequences in the reference set. The data is available in the Appendix 3.



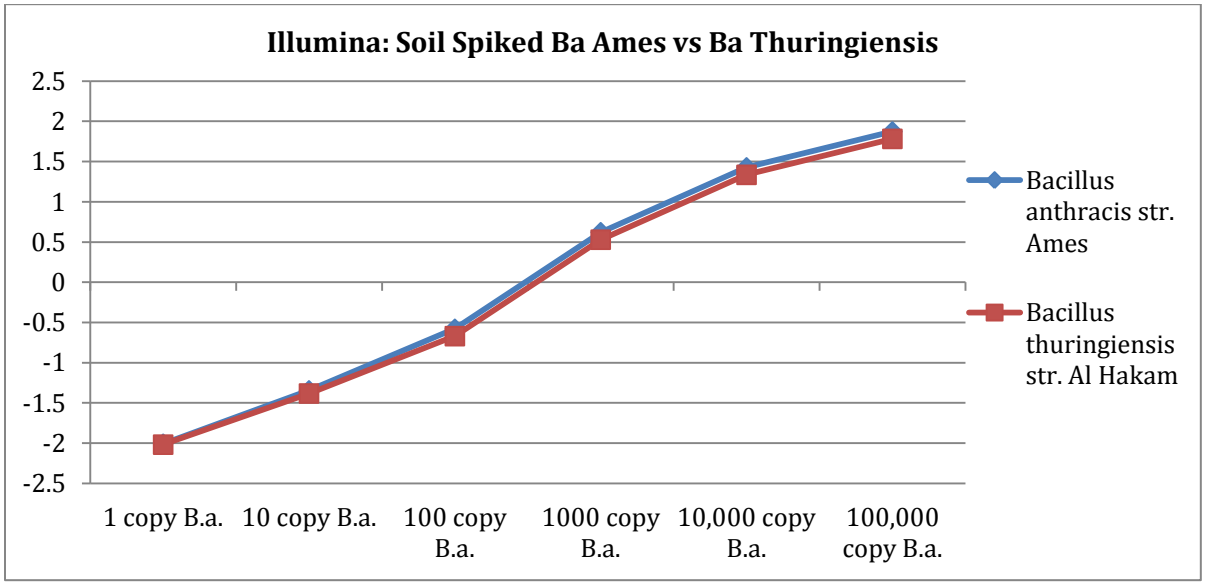
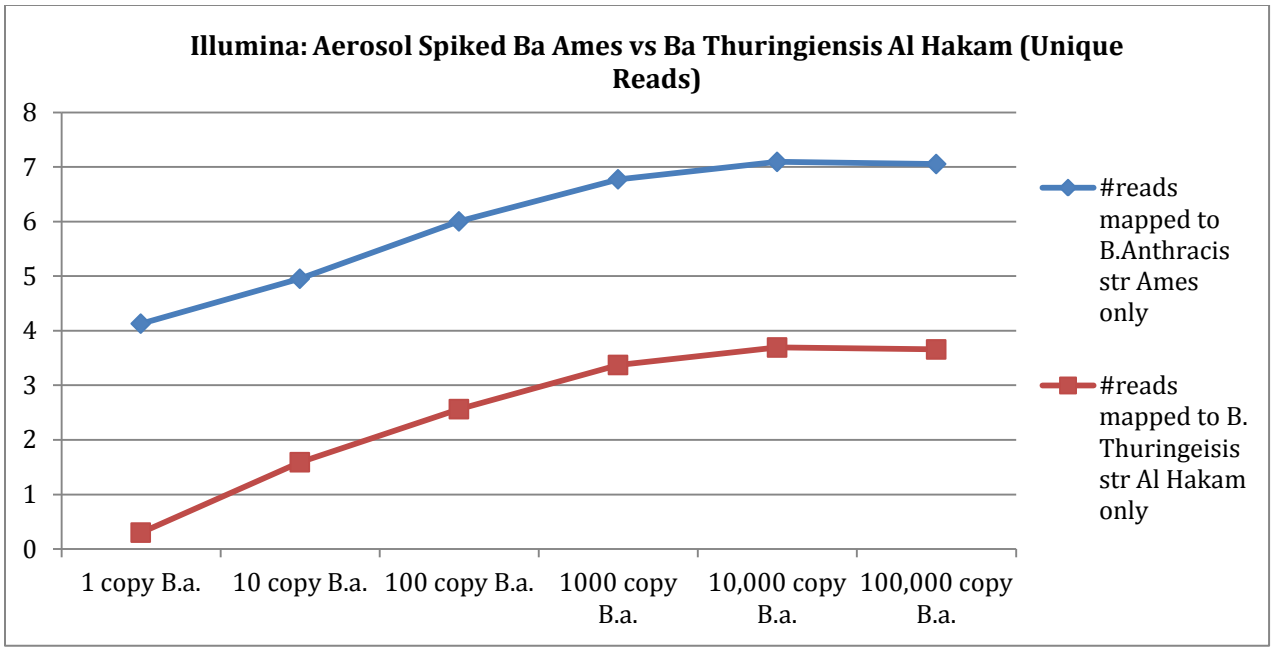




Figure 15a. Log plots of the ratio of mapped **Illumina reads** to total reads for 6 aerosol samples (top panel) and 6 soil samples (bottom panel). Mapping data is for the *B. anthracis* Ames and *B. thuringiensis* Al Hakam and is available in Appendix 3.



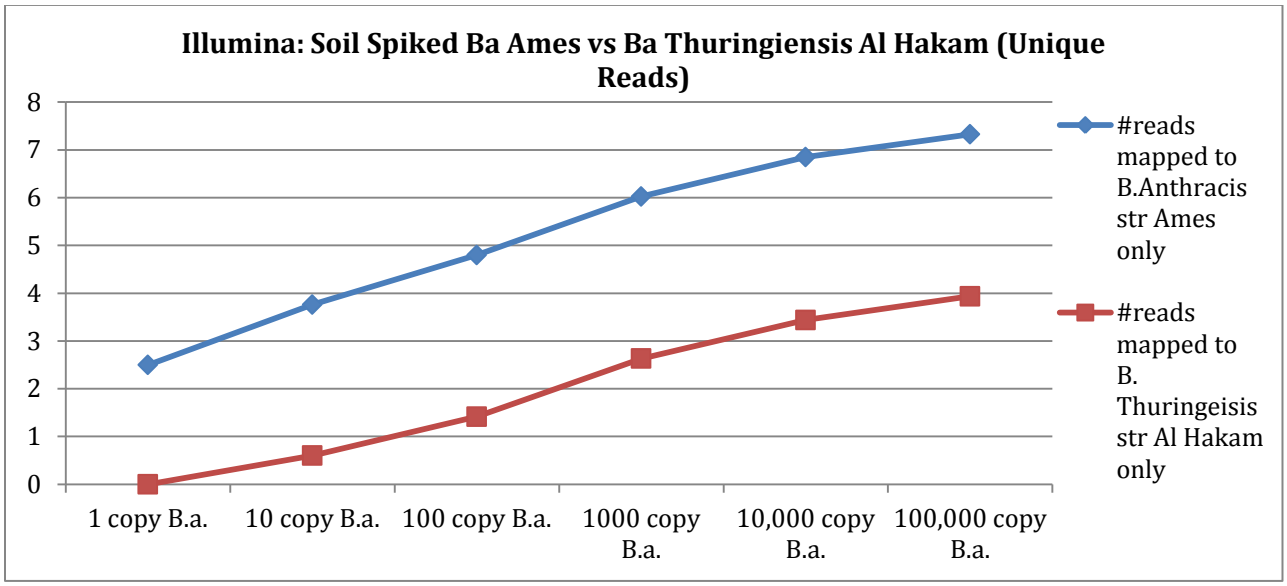




Figure 15b. Log plots of the ratio of mapped **Illumina reads** to total reads for 6 aerosol samples (top panel) and 6 soil samples (bottom panel). Only the reads mapping uniquely to either *B. anthracis Ames* or *B. thruringiensis* were considered. Reads mapping to both genomes were discarded. Mapping data is for the *B. anthracis* Sterne and *B. thuringiensis* Al Hakam and is available in Appendix 3.

The 454 reads from each sample were mapped to the reference sequence set and the resulting data is shown in Figures 16 and 17. In the aerosol samples there is an approximately 10 fold increase in the number of mapped reads as the sample number increases (unspiked < 1 copy B.a. < 10 copy B.a. < 100 copy B.a.). In all but

the unspiked control sample, the vast majority of the mapped reads map to the *B. anthracis* Sterne chromosome and pXO1 plasmid. Interestingly, there is an approximate 10x increase in the number of reads mapping to the Al Hakam strain, albeit at a 100-1000 fold lower level than to the Sterne strain (Figure 18). In the soil samples the results are similar, however the total number of mapped reads is considerably lower, probably due to a significantly more complex sample. Subsequently, the 10 fold increase does not start until the 100 copy B.a. spiked sample (unspiked = 10 copy B.a. < 100 copy B.a. < 1000 copy B.a.).

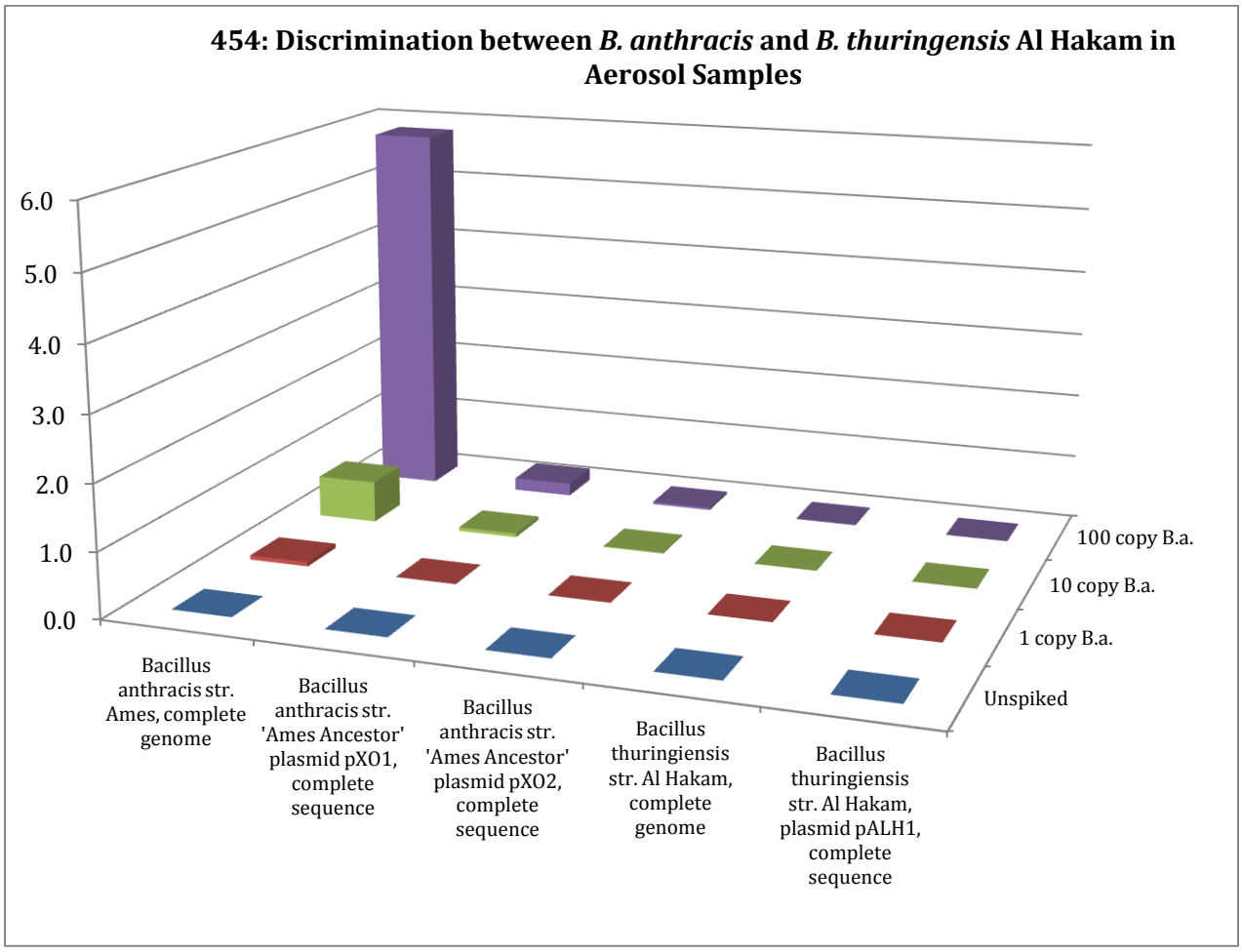



Figure 16. The percent of 454 mapped reads from the aerosol samples to each of the sequences in the reference set. The data is available in the Appendix 6.

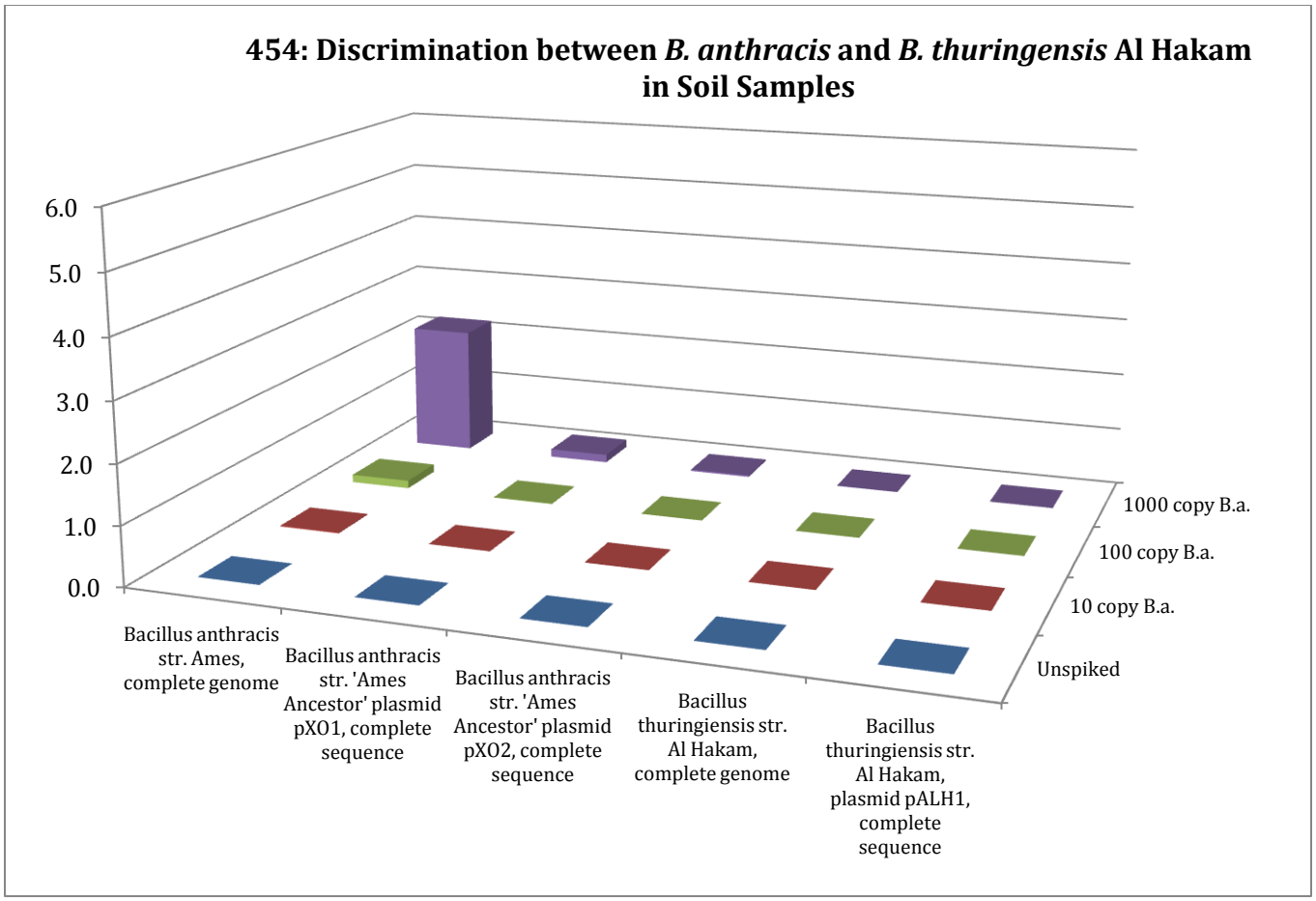



Figure 17. The percent of 454 mapped reads from the soil samples to each of the sequences in the reference set. The data is available in the Appendix 6.

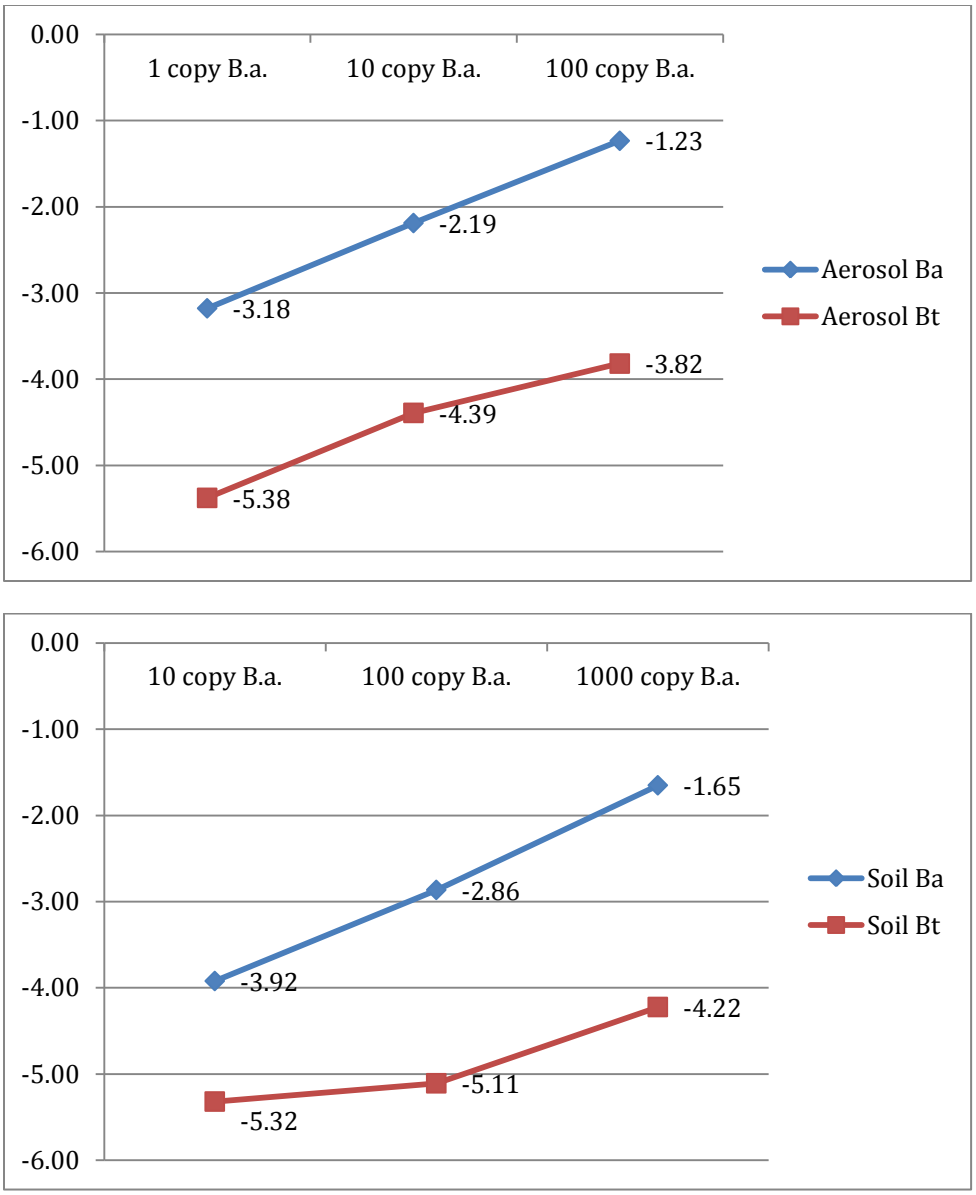



Figure 18. Log plots of the ratio of mapped **454 reads** to total reads for the aerosol sample spiked with 1-100 copies *B. anthracis* Ames and the soil sample spiked with 10-1000 copies *B. anthracis* Ames. Mapping data is for the *B. anthracis* Ames (Ba) or *B. thuringiensis* Al Hakam (Bt) and is available in Appendix 6.

**Mapping specificity using reference sequence set of *B. anthracis* (Sterne chromosome and pXO1 and pXO2 plasmids) and the near-neighbor *B. cereus* biovar anthracis (chromosome and pCI-XO1, pCI-XO2 and pBAsICI14 plasmids).** The reads from each Illumina sequenced sample were mapped to the reference sequence set and the resulting data is shown in Figures 19, 20 and 21. In both soil and aerosol samples, the proportion of reads (out of total) that could be mapped to *B. anthracis Str Ames* was consistently higher than proportion of reads

mapped to *B. cereus biovar anthracis.* As in the previous example, the discrimination ability of Illumina GAIIx reads is limited (particularly at low spike levels) until the additional correction of using only reads that uniquely map to each genome is considered.

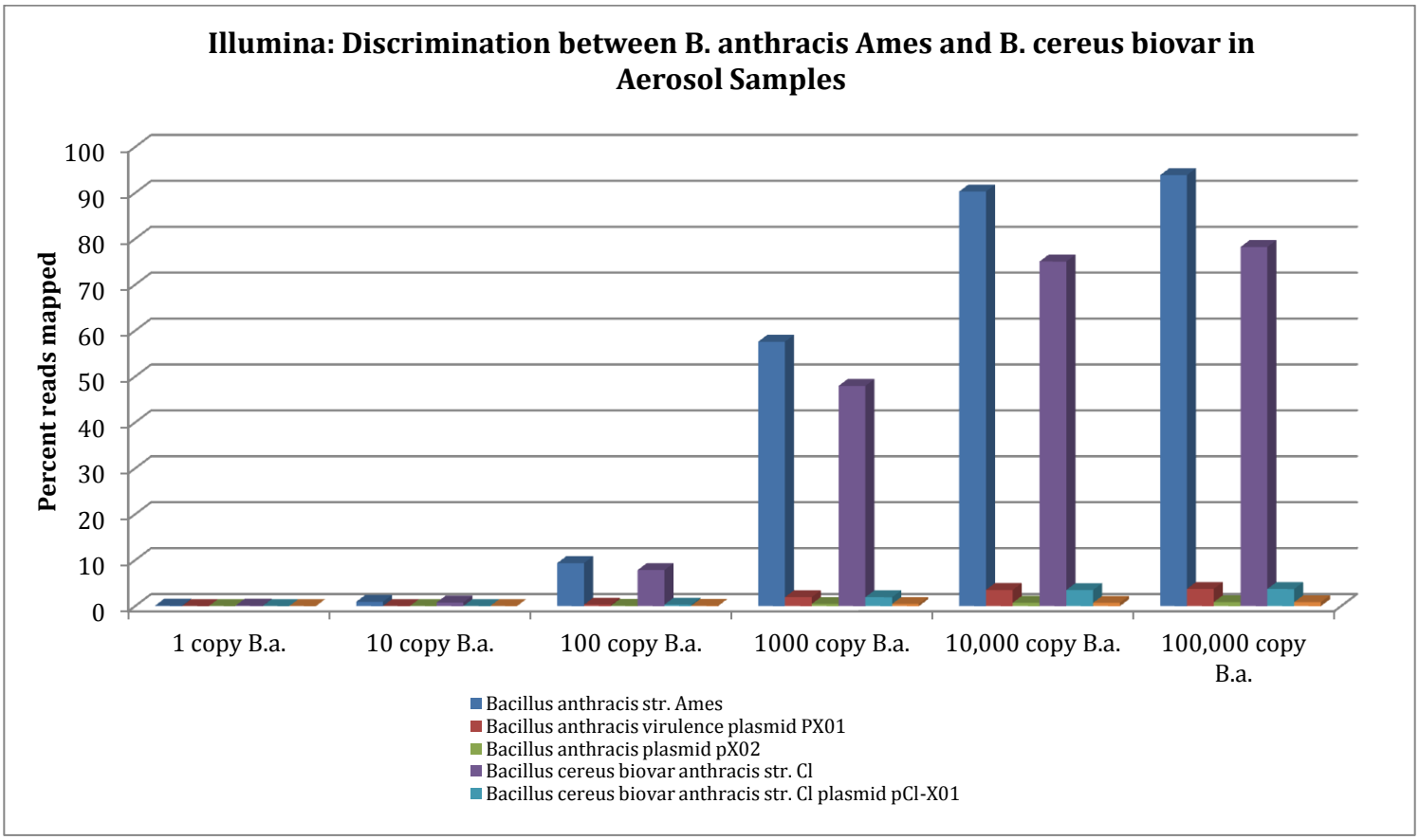



Figure 19. The percent of **Illumina reads** from the aerosol samples that mapped to only one member of the reference set. The data is available in the Appendix 3.

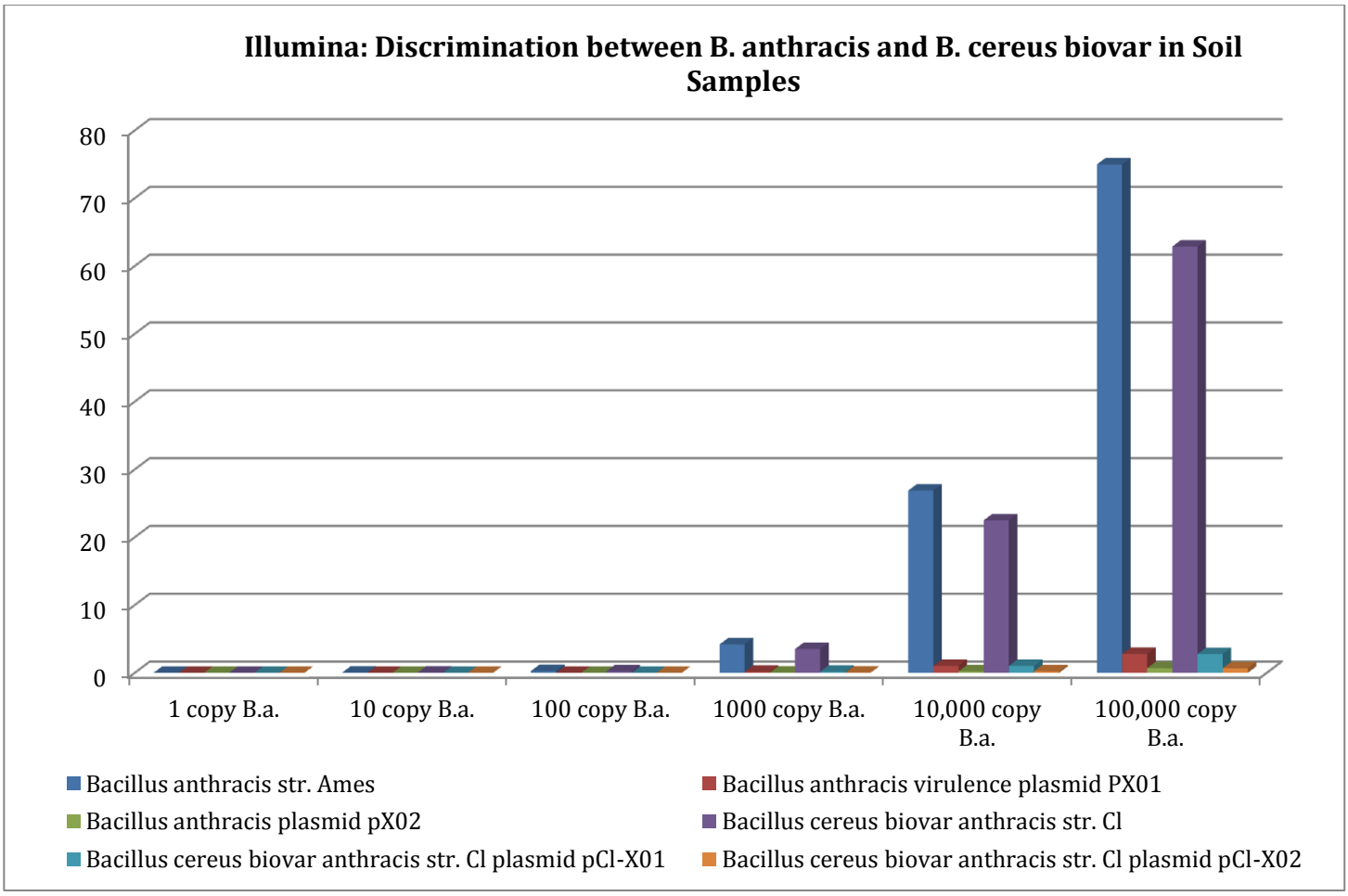
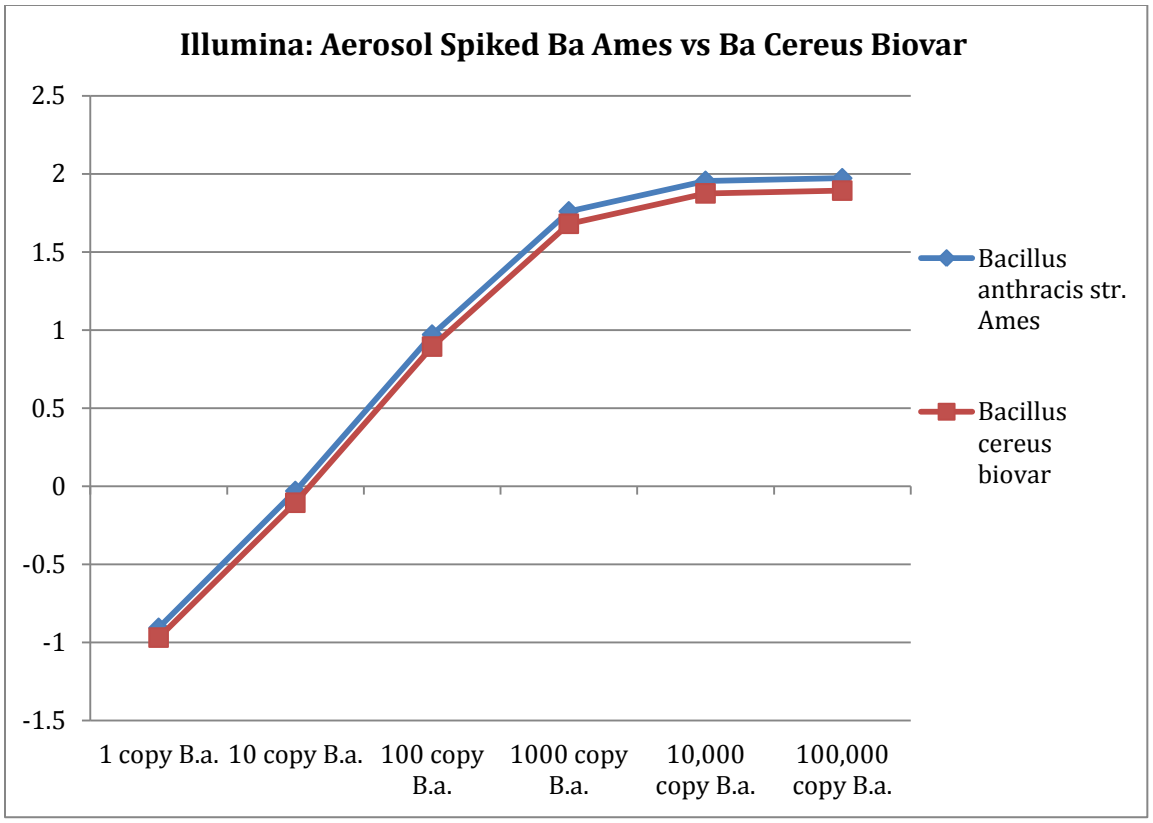



Figure 20. The percent of Illumina reads from the soil samples that only mapped to one of the reference set. The data is available in the Appendix 3



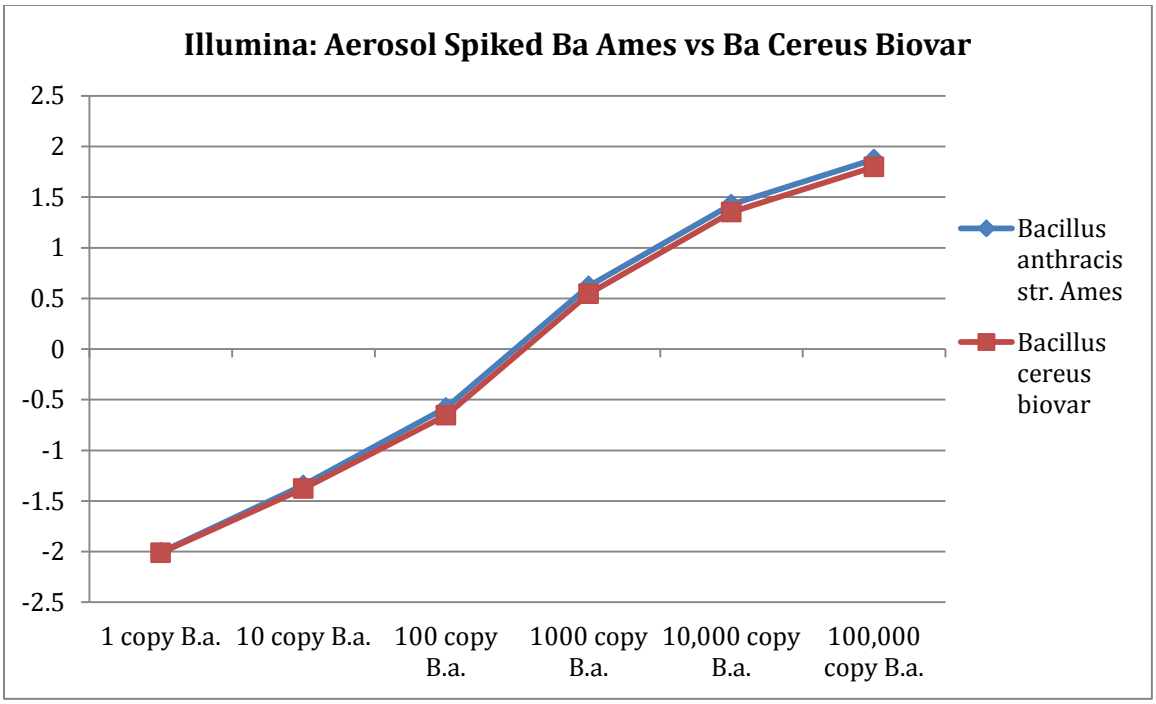




Figure 21a. Log plots of the ratio of mapped Illumina reads to total reads for aerosol spiked samples (top panel) and soil spiked samples (bottom panel). Mapping data is for the *B. anthracis* Sterne and B. *cereus* biovar is available in Appendix 3.

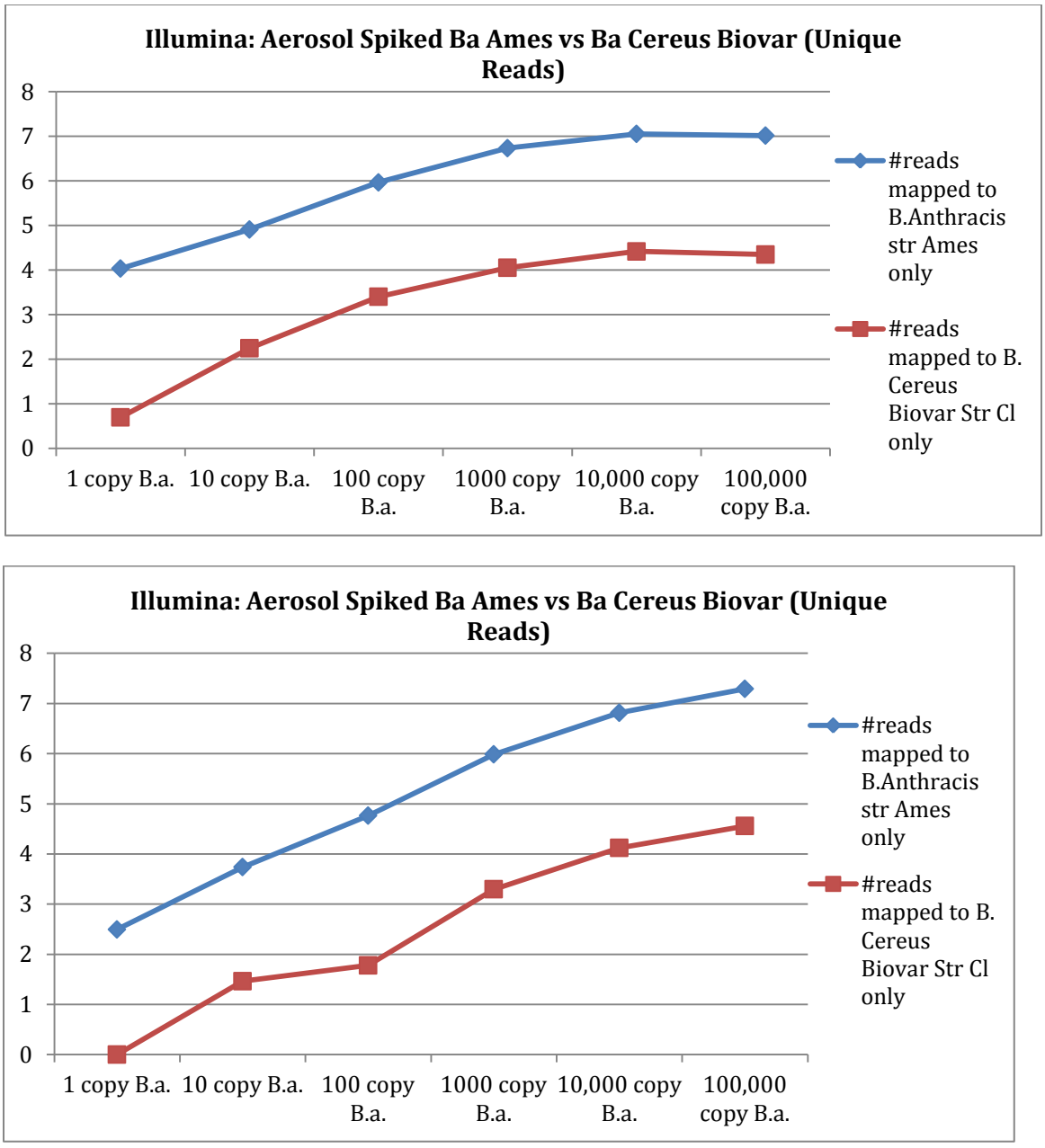



Figure 21b. Log plots of the ratio of mapped Illumina reads to total reads for aerosol spiked samples (top panel) and soil spiked samples (bottom panel). Only the reads mapping uniquely to either *B. anthracis Ames* or *B. cereus biovar anthracis* were considered. Reads mapping to both genomes were discarded. Mapping data is for the *B. anthracis Ames* and *B. cereus biovar anthracis* is available in Appendix 3.

The reads from each 454 sample were mapped to the reference sequence set and the resulting data is shown in Figures 22, 23 and 24. The results and conclusions are almost identical to the previous mapping study except for one point. Interestingly, reads mapping to the pXO1 and pXO2 plasmid sequence showed no increase in any of the samples. This is because in order for a read to be counted as mapped, it can map to only one sequence in the reference set. Because the pCI-XO1 and pCI-XO2 have a very high degree of similarity with the pXO1 and pXO2 plasmid, reads that may have mapped to the pXO1 and pXO2 plasmid in the first specificity study using the *B. thuringiensis* Al Hakam strain did not map uniquely in this study.

This study shows the vendor (Roche Scientific) supplied gsMapper mapping software can readily distinguish between 2 very closely related species or 2 strains of the same species of the Bacillus cereus group.

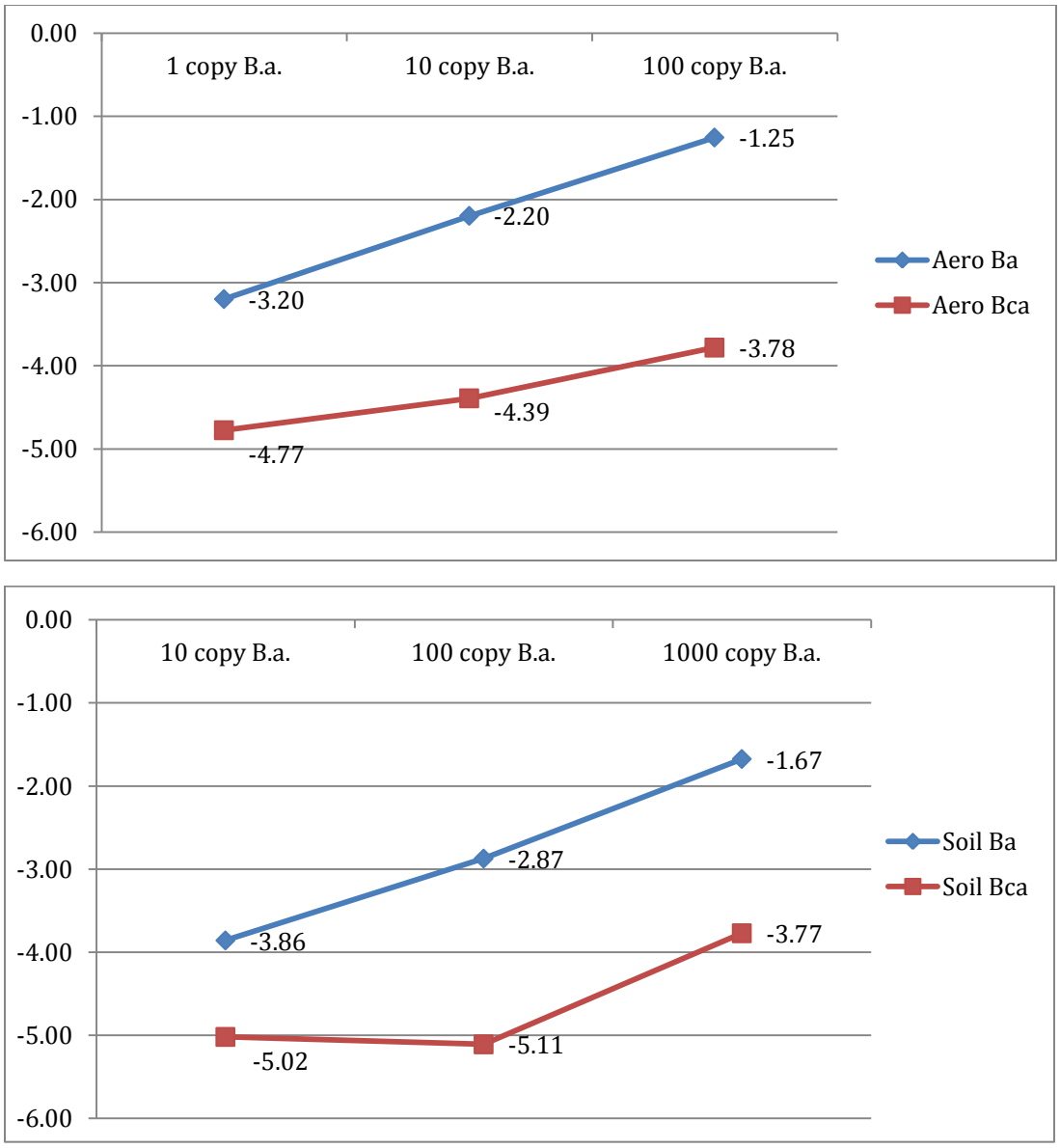



Figure 22. Log plots of the ratio of mapped **454 reads** to total reads for the aerosol sample spiked with 1-100 copies *B. anthracis* Ames and the soil sample spiked with 10-1000 copies *B. anthracis* Ames. Mapping data is for the *B. anthracis* (Ba) or *B. cereus* (Bca) chromosomes and is available in Appendix 7.

**454:Discrimination between *B. anthracis* and *B. cereus* biovar anthracis CI in Aerosol Samples**

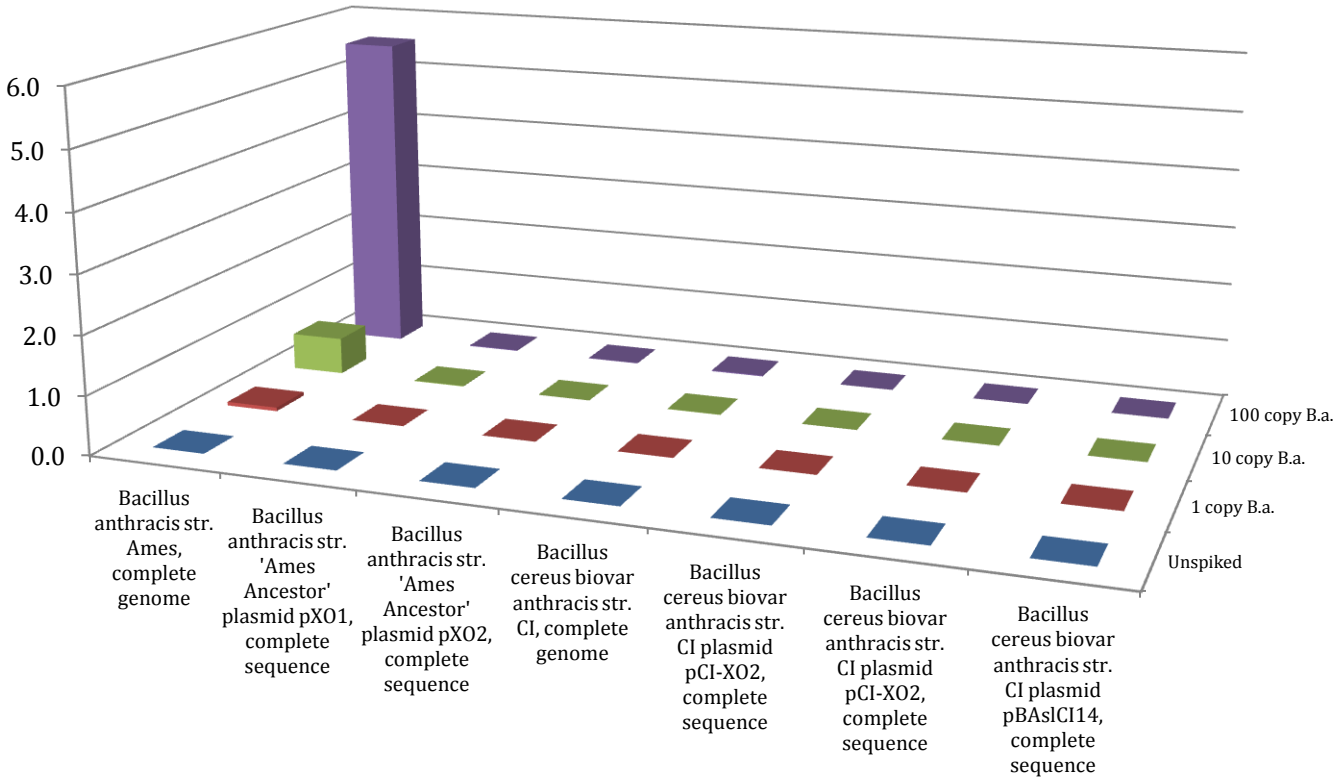



Figure 23. The percent of mapped **454 reads** from the aerosol samples to each of the sequences in the reference set. The data is available in the Appendix 7.

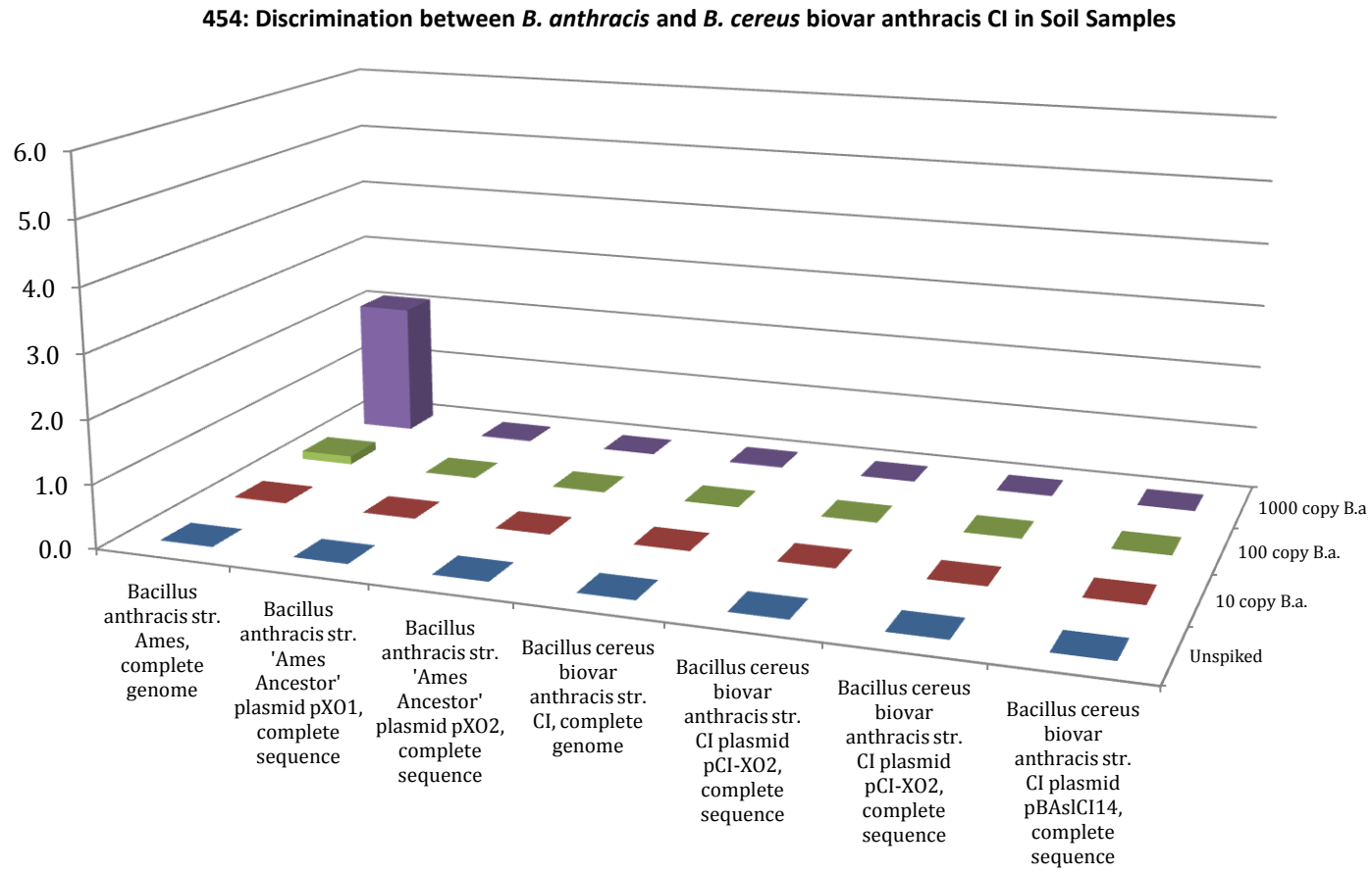



Figure 24. The percent of 454 mapped reads from the soil samples to each of the sequences in the reference set. The data is available in the Appendix 7.

**Results of the NCBI Blast Analysis.** A taxonomic distribution of the organisms in each sample was created by searching each set of sequencing reads against the NCBI GenBank database.

For Illumina reads the local version of the database was used. Only GenBank sequences classified as bacteria (or Archaea) have been considered for this analysis (viral and eukaryote analysis is currently underway and will be available upon completion).

The length of Illumina GAIIx reads (36 and 51 in this case) makes it likely that a read has multiple optimal alignments in different genomes (regions of sequence similarity between organisms longer than n-mer size are particularly prone to this). As such, a top-hit only approach carries a risk of producing a false positive hit (read maps to relative of the organism present in the sample, rather than the actual organism). The necessity to produce all-hits for all reads, coupled with large numbers of unique reads within the sequenced samples (roughly 63 and 89 million unique reads for aerosol and soil samples, respectively), makes

Megablast approach infeasible in a reasonable timeframe (estimated 4,800+ CPU hours). Instead of Megablast, a publically available short-reads mapping software (Bowtie), was used to map sequencing reads from each sample to all GenBank sequences classified as bacteria. All hits (up to 3 mismatches) were kept.

The resulting output from each Bowtie run was parsed to obtain the taxonomy IDs and names of the organisms matched by each read. All possible hits for each read were recorded and classified on the basis of their taxonomic classification (using NCBI's taxID). To avoid bias resulting from over representation of certain species within GenBank (species with multiple reference genomes of sub-strains artificially inflate the number of mapped reads) and bias associated with reads present in multiple copies within a genome (e.g. sRNA reads), each read was counted as matching to a given taxID (species or substrain) only once.

In order to improve the specificity of the microbiome characterization approach, additional analysis was performed to identify reads uniquely identifying a species. This is particularly helpful when trying to distinguish between reads mapping to closely related species that share significant sequence similarity. It should be noted that when multiple sub-strains of a given bacterial species are present (different taxIDs but nearly identical reference sequence), this statistic will exhibit low values, erroneously suggesting that the bacteria is not present in the sample. To correct for this issue, we have "collapsed" sub-strains of various bacteria, such that reads mapping to any of the sub-strains will instead be counted as mapping to parent species (e.g. *Chlamydia trachomatis* strains D/UW-3/CX, B/TZ1A828/OT, 434/Bu, L2b/UCH-1/proctitis, B/Jali20/OT, E/11023, E/150, G/9768, G/11222, G/11074, G/9301, D-EC, D-LC, Sweden2, will be counted as simply *Chlamydia trachomatis*). The full listing of sub-strains collapsed in this manner is available upon request.

The list of organism scientific names was sorted and the number of reads that uniquely map to a species and the number of occurrences of each organism name on the basis of **total hits** was tallied. No additional normalization was necessary due to the filtering options (see above) accounting for the major sources of expected bias. We have identified the top 15 species in each sample and created a union of top organisms in all samples (22 unique species in Aerosol and 24 unique species in Soil samples). See tables 35 and 36 for further details.

Table 35. Top 15 species in 6 **soil** samples spiked with *Bacillus anthracis*, sorted on the basis of number of **Illumina** reads matching only one taxID (i.e. hits that can uniquely identify a species)

| **1 copy B.a** | **10 copy B.a** | **100 copy B.a.** | **1000 copy B.a.** | **10,000 copy B.a** | **100,000 copy B.a.** |
|---|---|---|---|---|---|
| Ralstonia pickettii | Ralstonia pickettii | Ralstonia pickettii | Ralstonia pickettii | **Bacillus anthracis** | **Bacillus anthracis** |
| Nitrosospira multiformis | Nitrosospira multiformis | Nitrosospira multiformis | **Bacillus anthracis** | Ralstonia pickettii | Ralstonia pickettii |
| Cupriavidus metallidurans | Cupriavidus metallidurans | Cupriavidus metallidurans | Nitrosospira multiformis | Nitrosospira multiformis | Nitrosospira multiformis |
| Ralstonia solanacearum | Ralstonia solanacearum | Ralstonia solanacearum | Cupriavidus metallidurans | Cupriavidus metallidurans | Cupriavidus metallidurans |
| Delftia acidovorans | Delftia acidovorans | Delftia acidovorans | Ralstonia solanacearum | Ralstonia solanacearum | Bacillus cereus |
| Cupriavidus necator | Cupriavidus necator | Cupriavidus necator | Delftia acidovorans | Cupriavidus necator | Ralstonia solanacearum |
| Cupriavidus taiwanensis | Cupriavidus taiwanensis | **Bacillus anthracis** | Cupriavidus necator | Delftia acidovorans | Delftia acidovorans |
| Cupriavidus pinatubonensis | Hyphomicrobium denitrificans | Cupriavidus taiwanensis | Rhodococcus erythropolis | Cupriavidus taiwanensis | Cupriavidus necator |
| Stenotrophomonas maltophilia | Cupriavidus pinatubonensis | Bacillus megaterium | Cupriavidus taiwanensis | Propionibacterium acnes | Cupriavidus taiwanensis |
| Hyphomicrobium denitrificans | Arthrobacter sp. | Rhodococcus erythropolis | Hyphomicrobium denitrificans | Cupriavidus pinatubonensis | Bacillus thuringiensis |
| uncultured bacterium | Bacillus megaterium | Cupriavidus pinatubonensis | Cupriavidus pinatubonensis | Hyphomicrobium denitrificans | Cupriavidus pinatubonensis |
| Pseudomonas aeruginosa | uncultured bacterium | Hyphomicrobium denitrificans | uncultured bacterium | Pseudomonas fluorescens | Hyphomicrobium denitrificans |
| Magnetospirillum gryphiswaldense | Pseudomonas aeruginosa | uncultured bacterium | Bradyrhizobium sp. BTAi1 | uncultured bacterium | Arthrobacter sp. |
| Pseudomonas fluorescens | Stenotrophomonas maltophilia | Acidovorax sp. JS42 | Bacillus megaterium | Bacillus cereus | Bacillus weihenstephanensis |
| Acidovorax sp. JS42 | Bradyrhizobium sp. BTAi1 | Pseudomonas aeruginosa | Magnetospirillum gryphiswaldense | Acidovorax sp. JS42 | Stenotrophomonas maltophilia |

Table 36. Top 15 species in 6 samples **aerosol** spiked with *Bacillus anthracis*, sorted on the basis of number of **Illumina** reads matching only one taxID (i.e. hits that can uniquely identify a species)

| **1 copy B.a** | **10 copy B.a** | **100 copy B.a** | **1000 copy B.a** | **10,000 copy B.a** | **10,0000 copy B.a** |
|---|---|---|---|---|---|
| Ralstonia pickettii | Ralstonia pickettii | Ralstonia pickettii | **Bacillus anthracis** | **Bacillus anthracis** | **Bacillus anthracis** |
| Cupriavidus metallidurans | Cupriavidus metallidurans | **Bacillus anthracis** | Ralstonia pickettii | Ralstonia pickettii | Ralstonia pickettii |
| Ralstonia solanacearum | Ralstonia solanacearum | Cupriavidus metallidurans | Cupriavidus metallidurans | Cupriavidus metallidurans | Bacillus cereus |
| Bradyrhizobium sp. BTAi1 | Delftia acidovorans | Ralstonia solanacearum | Bradyrhizobium sp. BTAi1 | Bacillus cereus | Delftia acidovorans |
| Bradyrhizobium japonicum | Cupriavidus necator | Bradyrhizobium sp. BTAi1 | Ralstonia solanacearum | Ralstonia solanacearum | Cupriavidus metallidurans |
| Delftia acidovorans | Bradyrhizobium sp. BTAi1 | Bradyrhizobium japonicum | Bradyrhizobium japonicum | Delftia acidovorans | Bacillus thuringiensis |
| Rhodopseudomonas palustris | Bradyrhizobium japonicum | Delftia acidovorans | Delftia acidovorans | Bradyrhizobium sp. BTAi1 | Propionibacterium acnes |
| Cupriavidus necator | Cupriavidus taiwanensis | Rhodopseudomonas palustris | Rhodopseudomonas palustris | Bradyrhizobium japonicum | Ralstonia solanacearum |
| Cupriavidus taiwanensis | **Bacillus anthracis** | Cupriavidus necator | Hyphomicrobium denitrificans | Hyphomicrobium denitrificans | Bradyrhizobium sp. BTAi1 |
| Cupriavidus pinatubonensis | Cupriavidus pinatubonensis | Cupriavidus taiwanensis | Cupriavidus necator | Rhodopseudomonas palustris | Bacillus atrophaeus |
| Hyphomicrobium denitrificans | Rhodopseudomonas palustris | Cupriavidus pinatubonensis | Cupriavidus taiwanensis | Bacillus thuringiensis | Bacillus weihenstephanensis |
| Bradyrhizobium sp. ORS278 | Pseudomonas aeruginosa | Bradyrhizobium sp. ORS278 | Bradyrhizobium sp. ORS278 | Cupriavidus taiwanensis | Bradyrhizobium japonicum |
| Pantoea vagans | Hyphomicrobium denitrificans | Acidovorax sp. JS42 | Cupriavidus pinatubonensis | Cupriavidus necator | Rhodopseudomonas palustris |
| Pseudomonas aeruginosa | Stenotrophomonas maltophilia | Hyphomicrobium denitrificans | Stenotrophomonas maltophilia | Cupriavidus pinatubonensis | Cupriavidus taiwanensis |
| Stenotrophomonas maltophilia | Bradyrhizobium sp. ORS278 | Pantoea vagans | Pseudomonas aeruginosa | Bacillus weihenstephanensis | Cupriavidus necator |

These results show the same basic pattern seen in the studies using BWA. The proportion of reads attributed to *B. anthracis* increase for as the quantity of spiked *B. anthracis* increases (for both soil and aerosol samples). It should be noted that for soil samples, *B. anthracis* is not in the top 15 organisms for the 1 and 10 copy spiked samples (it is #68 for sample 1 and #45 for sample 2). Similarly, for aerosol samples, *B. anthracis* is not in the top 15 for the 1 copy spiked sample (it is #68). The relative concentrations of the top organisms in soil and aerosol samples can be seen in Figures 25 and 26.

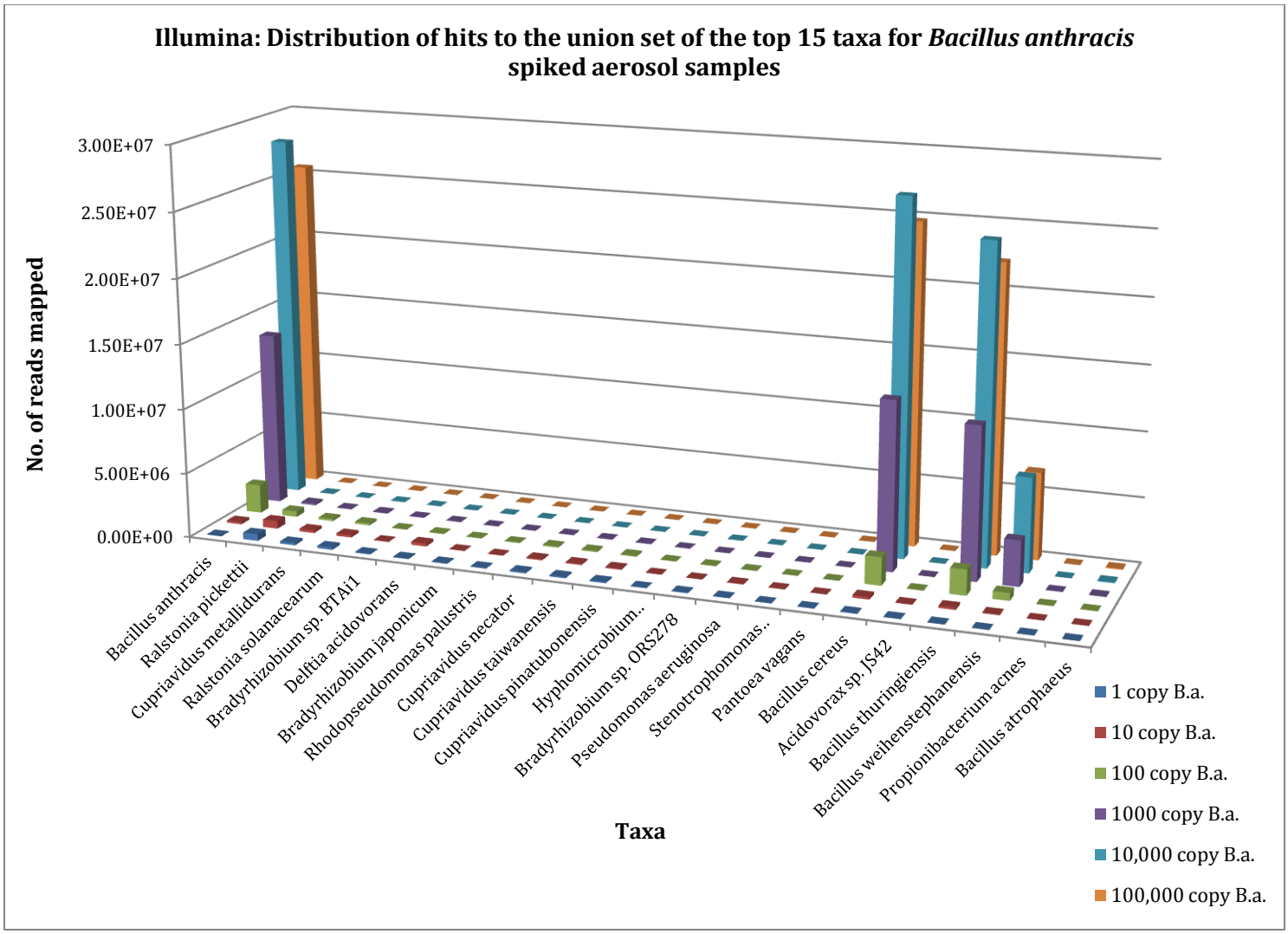



Figure 25. Total number of Illumina reads mapping with up to 3MM to the top 15 (sorted by number of reads mapping only one taxID) species in each sample (shown is union top 15 species in Aerosol spiked B.a. smaples).

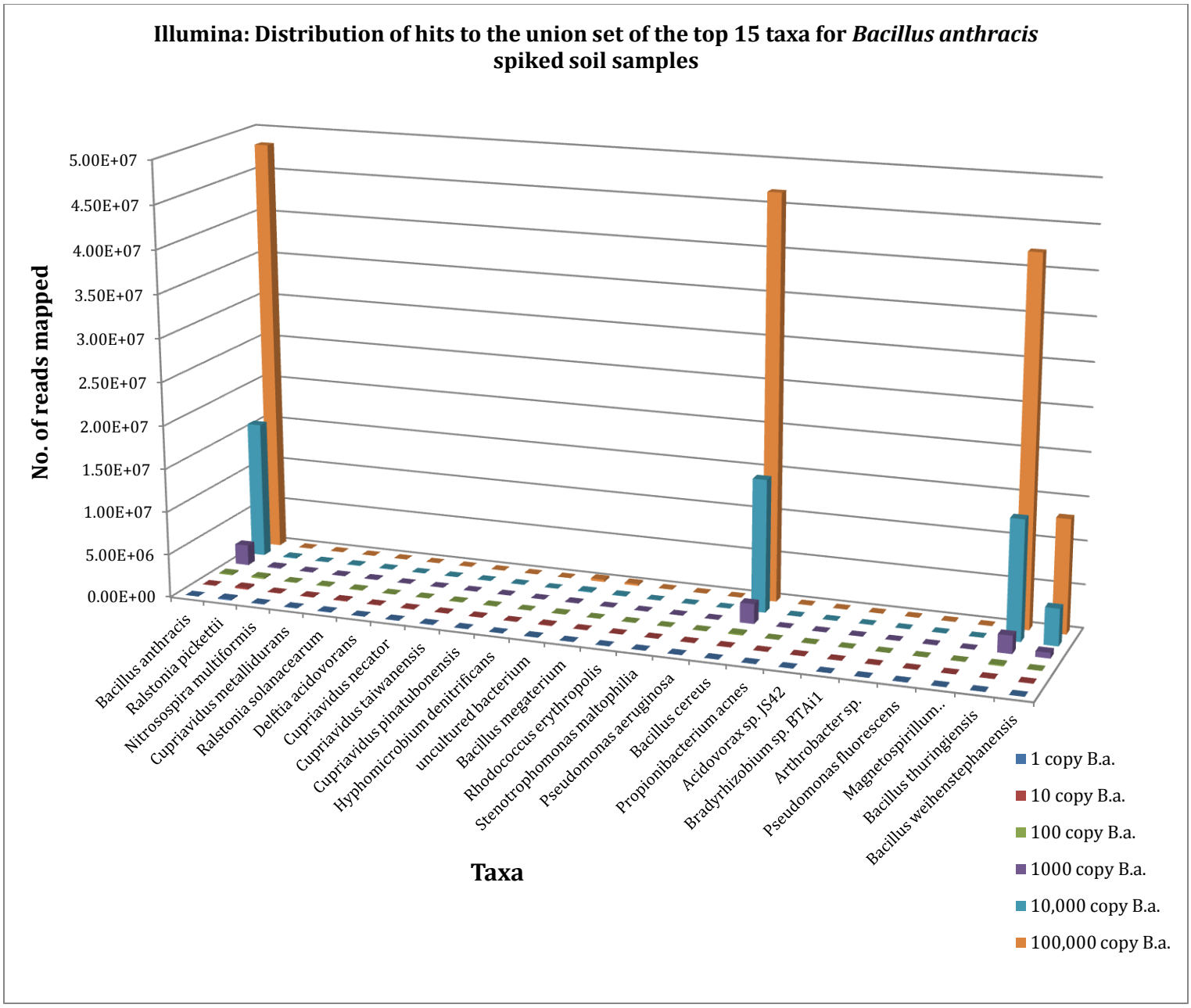



Figure 26. Total number of Illumina reads mapping with up to 3MM to the top 15 (sorted by number of reads mapping only one taxID) species in each sample (shown is union top 15 species in Soil spiked B.a. samples).

Table 37. Illumina aerosol and soil sample mapping

| Sample concentration | *B. anthracis* | | Background set | |
|---|---|---|---|---|
| | Reads Mapped | Percent Mapped | Reads Mapped | Percent Mapped |
| 1 copy BA/100 pg of aerosol DNA | 28,142 | 0.12% | 54,317 | 0.24% |
| 10 copies BA/100 pg of aerosol DNA | 196,147 | 0.96% | 40,622 | 0.20% |
| 100 copies BA/100 pg of aerosol DNA | 2,228,084 | 9.72% | 68,825 | 0.30% |
| 1000 copies BA/100 pg of aerosol DNA | 13,433,595 | 59.98% | 77,827 | 0.35% |
| 10,000 copies BA/100 pg of aerosol DNA | 28,183,408 | 94.50% | 115,857 | 0.39% |
| 100,000 copies BA/100 pg of aerosol DNA | 25,650,528 | 98.41% | 123,711 | 0.47% |
| 1 copy of BA/1 ng of soil DNA | 5424 | 0.01% | 17,633 | 0.03% |
| 10 copies of BA/1 ng of soil DNA | 27,129 | 0.05% | 31,490 | 0.06% |
| 100 copies of BA/1 ng of soil DNA | 142,562 | 0.27% | 18,134 | 0.03% |
| 1000 copies of BA/1 ng of soil DNA | 2,383,829 | 4.40% | 29,355 | 0.05% |
| 10,000 copies of BA/1 ng of soil DNA | 15,914,018 | 28.11% | 106,003 | 0.19% |
| 100,000 copies of BA/1 ng of soil DNA | 48,366,091 | 78.41% | 293,616 | 0.48% |

In the 454 sequencing analysis, the megablast program from NCBI version 2.2.18 was used with the following parameters:-m 8 -F F -I T -v 1 -b 1 -e 1e-20. The NCBI nucleotide and taxonomy databases were downloaded in Dec. 2010. Each input file was split into 256 smaller files and each of these smaller files were used as input to the megablast program running on a local Linux cluster.

The resulting output from each megablast run was parsed to obtain the NCBI sequence id (gi number) of the best hit for each query sequence. The gi number was

then used in a query against a local copy of the NCBI taxonomy database to obtain the scientific name of the organism from which the best hit sequence came from.

The list of organism scientific names was sorted and the number of occurrences of each organism name was tallied. Because some organism names go beyond the species name, only the first two parts of the name was used, which in most cases was the genus and species name.

The normalized occurrences of each organism were computed for each environmental sample. The normalization was required because of the variation in the number of sequencing reads generated for each sample. There were two normalization schemes considered. In the first normalization, the number of times a particular genus/species name occurs is divided by the total number of genus/species occurrences (total number of reads that had a megablast hit satisfying the cutoff parameter). Because this has the potential to be influenced by the introduction of *Bacillus anthracis* DNA, we also normalized over the total number of reads in the sample. This second normalization strategy was selected for further use because the relative impact of the introduction of *B. anthracis* DNA into the sample would be less if we used the total number of reads in the sample. Since the number of different taxa for a given sample (aerosol or soil) could exceed 5000, we created a union set of the top 25 taxa (taxa with the most hits) of each sample and show those results in Figure 27 and Figure 28. The data for these unions are available in the Appendix 8, and the full data sets of all hits are available in the supplemental data files described in Appendix 8.

These results show the same basic patterns seen in the studies using GSmapper. There is also an approximately 10 fold increase in the number of Blast hits to *B. anthracis* and the related *B. cereus* genomes as the copy number increases (Unspiked < 1 copy < 10 copy < 100 copy). The specificity is not as high when compared to the gsMapper results; 2-10 fold depending on the sample using megablast whereas with the gsMapper software the specificity was approximately 30-400 fold. This is probably caused by the fact that if the score of the megablast hit to 2 genomes is the same, the top listed hit is random, whereas, the mapping software did not report reads that mapped equally well to both species.

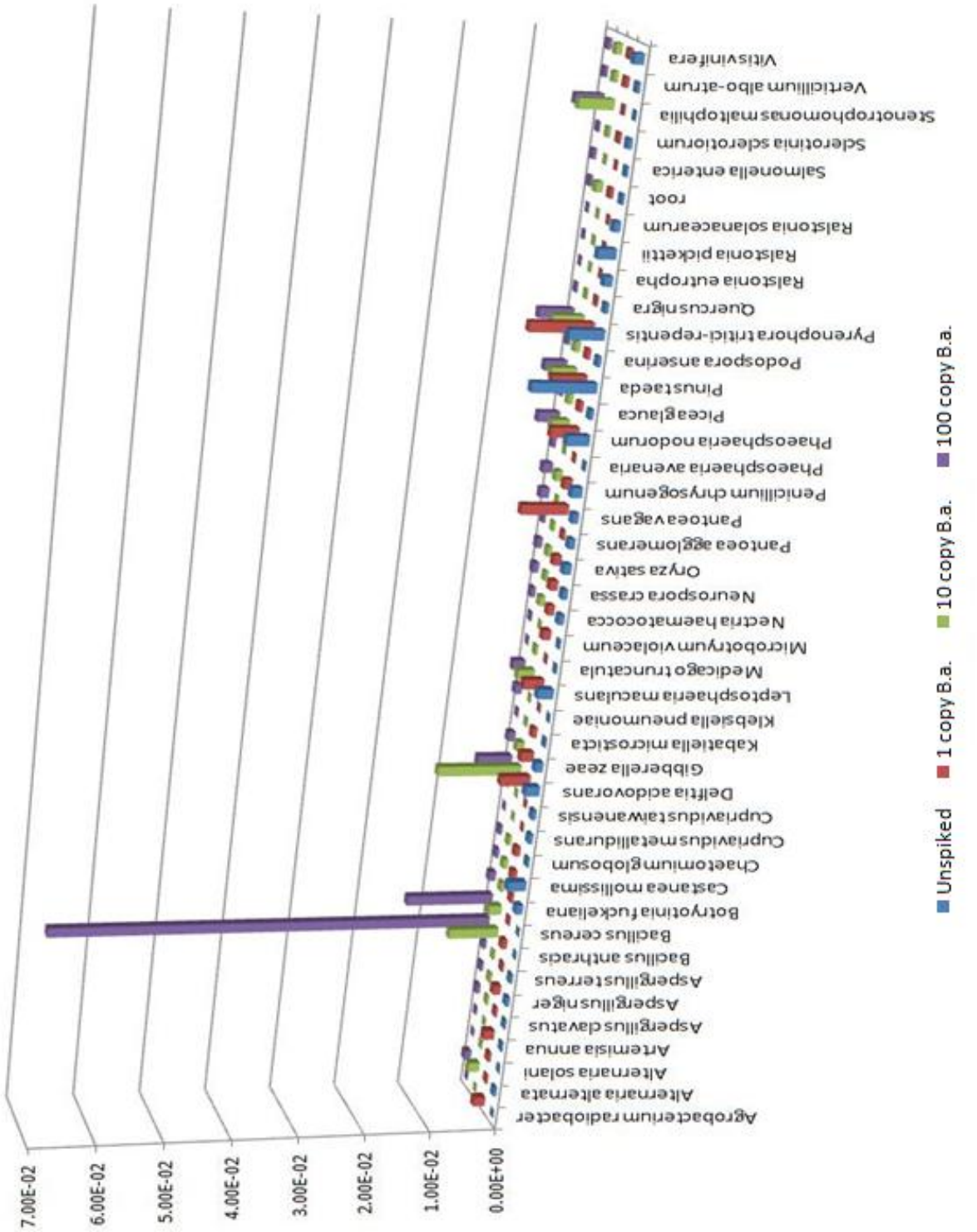



Figure 27. **454**: Distribution of the hits to the union set of the top 25 taxa for the aerosol samples. For each sample, the 25 taxon with the most hits were used to form the union set. The y-axis represents the number of times a read matched the organism named on the x-axis.

**Figure 10** Distribution of the hits to the union set of the top 25 taxa for the Soil samples. For each sample, the 25 taxon with the most hits were used to form the union set. The y-axis represents the number of times a read matched the organism named on the x-axis.

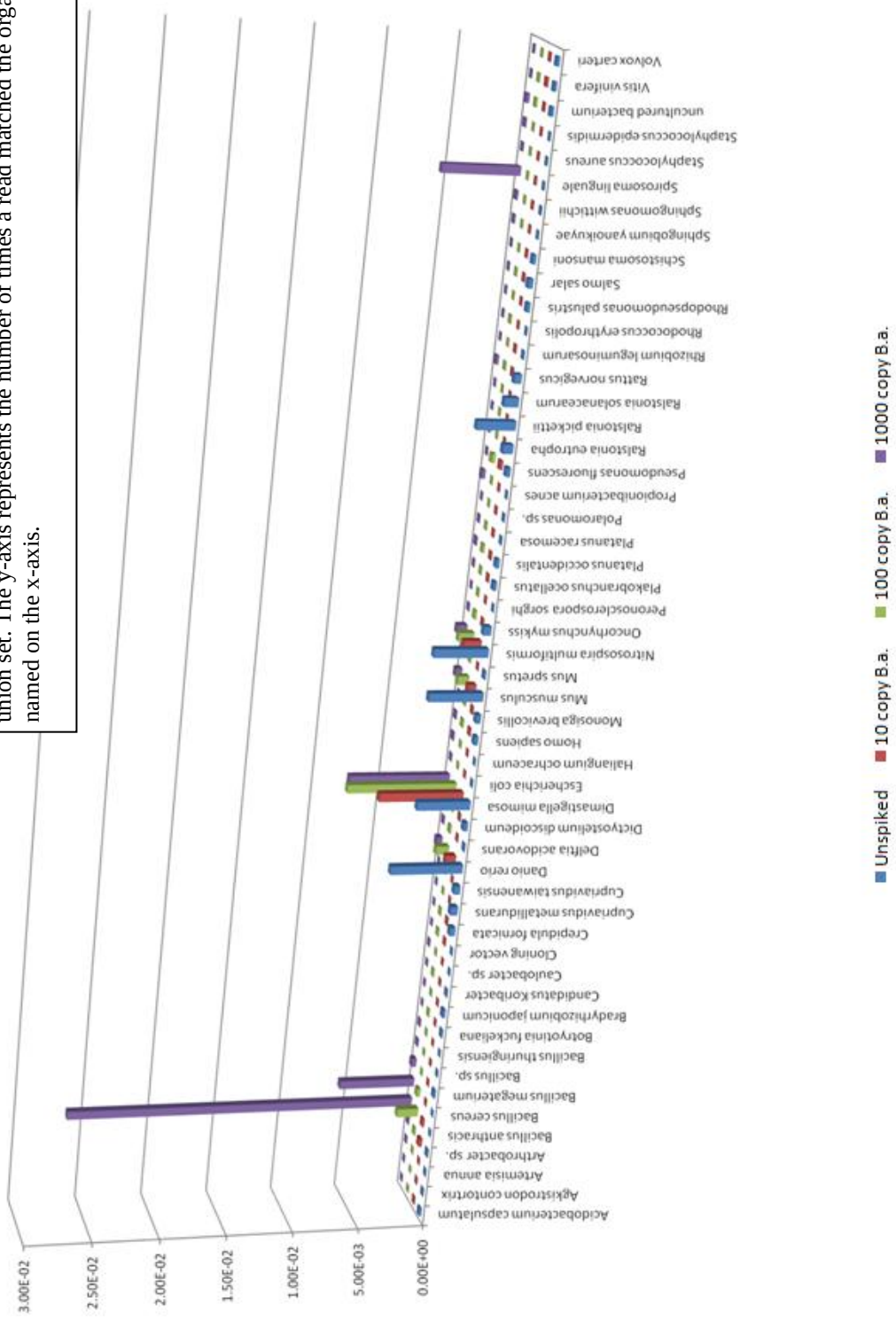



Figure 28. **454**: Distribution of the hits to the union set of the top 25 taxa for the Soil samples. For each sample, the 25 taxon with the most hits were used to form the union set. The y-axis represents the number of times a read matched the organism named on the x-axis.

# *Analysis of B. anthracis* Sterne ciprofloxacin resistant isolates

**Avirulent *B. anthracis* Sterne ciprofloxacin resistant isolates.** Following two rounds of selection by exposure to increasing ciprofloxacin concentrations, 3 ciprofloxacin resistant avirulent *B. anthracis* Sterne isolates were collected for this study. One mutant (10:8:1) was collected after three rounds of selection. The MIC value for the beginning sensitive Sterne strain was 0.047 µg/mL. MIC values for the 4 resistant isolates ranged from 24 µg/mL to >32 µg/mL ciprofloxacin (the limit of the Etest). It is not yet known whether this is a relationship between association with a particular MIC value and particular genomic changes responsible for resistance. Table 38 contains a compilation of the different resistant isolates and their MIC values.

Table 38. Avirulent *B. anthracis* Sterne – Ciprofloxacin resistant isolate MIC summary. All MIC values are in µg/ml.

| **Avirulent *B. anthracis* Sterne wild-type Ciprofloxacin MIC (Etest) = 0.047 µg/ml** | | | | | |
|---|---|---|---|---|---|
| **Round 1** | | **Round 2** | | **Round 3** | |
| **Name** | **MIC** | **Name** | **MIC** | **Name** | **MIC** |
| M1 | 0.75 | M1:1 | 24.0 | | |
| | | M1:6 | >32 | | |
| M10 | 1.0 | M10:8 | 12.0 | M10:8:1 | >32 |
| M19 | 1.0 | M19:2 | >32 | | |

*Evaluation of ciprofloxacin resistant Bacillus anthracis isolates using microarrays and sequencing technologies*

**SNPs identified in ciprofloxacin-resistant avirulent B. anthracis Sterne isolates**. We first tested the *B. anthracis* Sterne tiling microarray on reference (non-selected) *B. anthracis* Sterne. We then tested four ciprofloxacin resistant *B. anthracis* Sterne isolates (1:1, 1:6, 10:8:1, and 19:2) using the *B. anthracis* Sterne tiling array. The number of candidate SNP's on the *B. anthracis* Sterne chromosome that were identified from overlapping probes was: 2078 in clone 1:1, 42 in 1:6, 86 in clone 10:8:1, and 19 in 19:2. In addition, a 93 kb deletion relative to the reference strain was identified in the 10:8:1 isolate. The missing region spans positions 749405 to 842475 on the Sterne chromosome. Two of the overlapping high-scoring probe pairs are located in genes encoding the proteins targeted by ciprofloxacin, DNA gyrase A and topoisomerase IV.

Several other SNP's are located in ABC transporter/permease genes while many others were in genes encoding hypothetical or un-annotated proteins.

The four ciprofloxacin-resistant *B. anthracis* Sterne isolates and the *B. anthracis* Sterne reference isolate were sent to Eureka Genomics to perform Illumina sequencing and BYU to perform 454 sequencing in order to compare the accuracy and cost-effectiveness of these two technologies for SNP detection, and to provide independent confirmation of the microarray results. The strains were sequenced twice via Illumina technology. The initial, version 3, sequencing was performed in 2008. In 2011, when improved chemistry and more accurate data with longer read lengths was possible, the DNA was re sequenced (version 4). Both sets of this sequencing data is listed in this report. Table 39 below displays the total number of SNPs identified by each method. Only those SNPs above a threshold of 0.30 for the sequencing technologies and those identified by 2 or more probes for the microarrays are included in this total number.

Table 39. Total SNPs identified by each technology

| ***B. anthracis*** **Sterne ciprofloxacin-resistant clones** | **# of SNP's identified by microarray** | **# of SNP's identified by Illumina v4** | **# of SNP's identified by Illumina v3** | **# of SNP's identified by 454** |
|---|---|---|---|---|
| 1:1 | 2078 | 6 | 10 | 8 |
| 1:6 | 42 | 2 | 20 | 6 |
| 10:8:1 | 86 | 7 | 11 | 10 |
| 19:2 | 19 | 5 | 7 | 16 |

In the following tables (40-43) the SNPs are detailed for each mutant. SNPs present in intergenic regions and very large deletions are not included here. For the sequencing technologies, the number of reads conferring each SNP is listed along with the proportion this represents. An (X) is marked in the Microarray column if at least two probes identified the SNP in the regions that concurred with the sequencing results. The only mutant with microarray data that matches with the sequencing technologies was 10:8:1 as shown below. *B. anthracis* Sterne tiling array did not product robust data that correlated with sequencing results. The probe design for *B. anthracis* Sterne used an overlap of 55% in order to fit all probes onto a 388K array. The assay showed that this overlap is not sufficient for SNP identification using the tiling array. We have since developed a tiling array for *F. tularensis* LVS genome with 85% overlap of probes. Tiling array data showed that the consistency rate between microarray and sequencing is more than 95% (Jaing et al., manuscript in preparation).

Six mutations identified by either Illumina or 454 sequencing were subjected to Sanger Big Dye Terminator Sequencing performed at LLNL for SNP verification in strains M1:1 and M1:6. In M1:1, 4 of the 6 mutations were confirmed. 2 of the 6 mutations could not be validated with Sanger sequencing. A GC insertion had been detected in gene BAS0794 using 454 sequencing, when sequenced using Big Dye Terminator technology, this region instead contained a T deletion 70 bases downstream and no GC insertion. In gene BAS3720, 454 detected a TCT/C substitution. Sanger did not confirm this mutation; the sequence was determined to match the wild type strain.

In M1:6, 2 of the 6 tested SNPs were also found to be inaccurate using Sanger sequencing. In 454, gene BAS5135 contained a GAA insertion not seen by Sanger analysis. Gene BAS5220, Illumina v3 analysis revealed a G to C substitution also not confirmed with Sanger. The remaining 4 SNPs Sanger tested for strain M1:6 all confirmed the original Illumina or 454 mutations.

Table 40. Mutant 1:1 Sequencing Results Comparison

| Gene | Gene Description | Gene Location | SNP Location | SNP Type | Illumina v4 | Illumina v3 | 454 | Sanger |
|---|---|---|---|---|---|---|---|---|
| BAS0006 | DNA gyrase subunit A | 6596-9067 | 6849 | Sub C->T | 179(1.00) | 13 (1.00) | 19 (1.00) | Verified |
| BAS0627 | ABC transporter, nucleotide binding domain | 677157-678104 | 677934 | Del A | | | 7 (0.50) | |
| BAS0794 | transcriptional regulator, TetR family | 842297-842875 | 842404 | Ins GC | | | 17 (1.00) | Del T @ 842473 |
| BAS0869 | sensor histidine kinase | 932601-934190 | 933,156 | Sub G->T | | 5(0.50) | | |
| BAS1222 | formate/nitrite transporter family protein | 1267131-1267982 | 1,267,631 | Sub C->A | | 6(0.67) | | |
| BAS3009 | penicillin-binding protein, C-terminus | 2983668-2984141 | 2983918 | Ins C | | 12(1.00) | 10 (1.00) | |
| BAS1340 | proton/glutamate symporter protein | 1373442-1373987 | 1,373,973 | Ins T | 101(1.00) | | | |
| | | | 1,373,974 | Ins T | 103(1.00) | | | |
| | | | 1,373,975 | Ins T | 103(1.00) | | | |
| | | | 1,373,976 | Ins T | 103(1.00) | | | |
| BAS2010 | ABC transporter ATP-binding protein | 2015766-2017661 | 2,016,697 | Sub G->T | | 7(0.67) | | |
| BAS3379 | oligopeptide ABC transporter oligopeptide-binding protein | 3352939-3354627 | 3,354,627 | Ins T | | 10(1.00) | | Verified |
| BAS3391 | DNA topoisomerase IV subunit A | 3363272-3365695 | 3365454 | Sub G->T | 95(1.00) | 7 (1.00) | 10 (1.00) | |
| BAS3658 | polyribonucleotide nucleotidyltransferase | 3620153-3622291 | 3621030 | Ins AG | | | 10 (1.00) | Verified |
| BAS3720 | phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase | 3687114-3688319 | 3687446-8 | TCT->C | | | 4 (0.40) | No Mutation |
| BAS4274 | penicillin-binding protein | 4184340-4186094 | 4,185,453 | Sub G->T | | 5(0.50) | | |
| BAS5177 | modification methylase, HemK family | 5056920-5057173 | 5057173 | Ins GAA | | | 18 (1.00) | Verified |
| BAS5068 | hypothetical protein | 4942080-4944080 | 4,942,343 | Sub C->A | | 5(0.50) | | |
| rRNA-23s-11 | 23S ribosomal RNA | 4653344-4656254 | 4,655,363 | Sub C->A | | 17(0.50) | | |

Table 41. Mutant 1:6 Sequencing Results Comparison

| Gene | Gene Description | Gene Location | SNP Location | SNP Type | Illumina v4 | Illumina v3 | 454 | Sanger |
|---|---|---|---|---|---|---|---|---|
| BAS0006 | DNA gyrase subunit A | 6596-9067 | 6849 | Sub C->T | 113(1.00) | 5 (1.00) | 11 (1.00) | Verified |
| BAS0291 | ATP-dependent DNA helicase PcrA | 311188-313443 | 311,272 | Sub C->A | | 5(0.50) | | |
| BAS0595 | sensory box/GGDEF family protein | 642479-644182 | 643376 | Ins CCGCG | | | 12 (0.86) | Verified |
| BAS0627 | ABC transporter, nucleotide binding domain | 677157-678104 | 677934 | Del A | | | 8 (0.50) | |
| BAS0794 | transcriptional regulator, TetR family | 842297-842875 | 842614 | Ins GCGGGTC TTGC | | | 15 (0.94) | |
| BAS0828 | cell wall anchor domain-containing protein | 877745-880654 | 879,482 | Sub C->A | | 7(0.58) | | |
| BAS1009 | hypothetical protein | 1063377-1065986 | 1,065,552 | Sub G->T | | 5(0.63) | | |
| BAS1502 | hypothetical protein | 1523014-1538067 | 1,536,488 | Sub C->A | | 6(0.67) | | |
| BAS3025 | permease | 2999001-3000260 | 2,999,185 | Sub G->T | | 5(0.63) | | |
| BAS3391 | DNA topoisomerase IV subunit A | 3363272-3365695 | 3365454 | Sub G->A | 106(0.98) | 8 (0.89) | 12 (0.86) | |
| | | | | Sub G->T | | 1 (0.11) | | Verified |
| BAS5135 | ABC transporter ATP-binding protein, N-terminus | 5019337-5019846 | 5019628-30 | ACA->C | | | 4 (0.33) | No SNP |
| BAS3452 | hypothetical protein | 3423143-3429484 | 3,429,444 | Sub G->T | | 6(1.00) | | |
| BAS3498 | hypothetical protein | 3472101-3473351 | 3,472,849 | Sub G->T | | 5(0.56) | | |
| argD | acetylornithine aminotransferase | 3973095-3974225 | 3,973,552 | Sub G->T | | 8(0.73) | | |
| BAS4402 | aquaporin Z | 4313089-4313754 | 4,313,705 | Sub G->T | | 8(0.62) | | |
| BAS4422 | long-chain-fatty-acid--CoA ligase | 4330726-4332417 | 4,332,175 | Sub C->A | | 5(0.83) | | |
| BAS4442 | hypothetical protein | 4351157-4353814 | 4,353,323 | Sub G->T | | 5(0.71) | | |
| BAS4536 | hypothetical protein | 4442637-4443629 | 4,442,670 | Sub G->T | | 5(0.71) | | |
| BAS4553 | methionine gamma-lyase | 4458681-4459856 | 4,459,071 | Sub G->T | | 6(0.75) | | |
| BAS4752 | 1,4-dihydroxy-2-naphthoate octaprenyltransferase | 4636149-4637102 | 4,636,570 | Sub G->T | | 5(0.63) | | |
| BAS5136 | stage II sporulation protein D | 5019974-5020993 | 5,020,846 | Sub G->T | | 6(0.67) | | |
| BAS5207 | Collagen adhesion protein | 5094108-5096756 | 5,095,366 | Sub C->A | | 11(0.69) | | |
| BAS5220 | multidrug resistance protein, putative | 5109644-5108481 | 5109171 | Sub C->A | | 6 (0.86) | | No SNP |

Table 42. Mutant 10:8:1 Sequencing Results Comparison

| Gene Description | Gene Location | SNP Location | SNP Type | Illumina v4 | Illumina v3 | 454 | Sanger | Microarray |
|---|---|---|---|---|---|---|---|---|
| DNA gyrase subunit A | 6596-9067 | 6849 | Sub C->T | 91(1.00) | 19 (1.00) | 11 (1.00) | Verified | X |
| xanthine/uracil permease family protein | 714320-715612 | 714,817 | Sub T->C | | 12(1.00) | | | |
| oligopeptide ABC transporter, oligopeptide-binding protein | 1161085-1162689 | 1162599-600 | Del AT | | | 17 (1.00) | | |
| Proton/glutamate symporter protein | 1373442-1373987 | 1,373,973 | Ins T | 86(1.00) | | | | |
| | | 1,373,974 | Ins T | 84(1.00) | | | | |
| | | 1,373,975 | Ins T | 82(1.00) | | | | |
| | | 1,373,976 | Ins T | 82(1.00) | | | | |
| hypothetical protein | 1523014-1538067 | 1,523,233 | Sub A->T | | 5(0.56) | | | |
| hypothetical protein | 2587646-2588470 | 2,588,359 | Sub T->C | | 10(0.63) | | | |
| hypothetical protein | 2696838-2697347 | 2,697,137 | Ins C | | 6(1.00) | | | |
| RocB protein | 2763122-2764762 | 2,764,366 | Sub C->A | | 7(0.54) | | | |
| hypothetical protein | 2864166-2865347 | 2,864,925 | Sub T->C | | 7(0.58) | | | |
| DNA topoisomerase IV subunit A | 3363272-3365695 | 3365454 | Sub G->A | 73(1.00) | 25(0 .93) | 9 (1.00) | | X |
| hemolysin A | 4014548-4015387 | 4015055 | Ins ATC | | | 14 (1.00) | | |
| bifunctional preprotein translocase subunit SecD/SecF | 4218755-4221019 | 4,220,942 | Sub T->C | | 8(.067) | | | |
| D-alanyl-D-alanine carboxypeptidase family protein | 4625357-4626217 | 4,626,111 | Sub T->C | | 9(0.53) | | | |
| hypothetical protein | 4706252-4707910 | 4706846 | Sub T->C | 56(1.00) | 23 (1.00) | 9 (1.00) | | |
| lipoprotein, putative | 4829169-4830134 | 4829181 | Del A | | | 8 (0.89) | | |
| stage 0 sporulation protein F | 5065837-5066205 | 5065959-60 | Del AA | | | 13 (1.00) | | |

Table 43. Mutant 19:2 Sequencing Results Comparison

| Gene | Gene Description | Gene Location | SNP Location | SNP Type | Illumina v4 | Illumina v3 | 454 | Sanger |
|---|---|---|---|---|---|---|---|---|
| BAS0006 | DNA gyrase subunit A | 6596-9067 | 6849 | Sub C->T | 94(1.00) | 21 (1.00) | 12 (1.00) | Verified |
| BAS0276 | phosphoribosylaminoimidazole carboxylase, ATPase subunit | 296037-297188 | 296553 | Sub C->T | | | 17 (1.00) | |
| BAS0361 | DNA topoisomerase III | 391511-393700 | 391751 | Sub G->A | | | 13 (0.68) | |
| BAS0794 | transcriptional regulator, TetR family | 842297-842875 | 842399 | Ins GC | | | 22 (1.00) | |
| BAS2000 | conserved domain protein | 2006552-2007553 | 2006853 | Ins T | | | 6 (0.86) | |
| BAS3391 | DNA topoisomerase IV subunit A | 3363272-3365695 | 3365454 | Sub G->A | 70(1.00) | 21 (1.00) | 16 (1.00) | |
| BAS3618 | DNA mismatch repair protein MutS | 3577813-3580491 | 3579082 | Ins GC | | | 10 (0.91) | |
| BAS3684 | DNA topoisomerase I | 3651076-3648998 | 3651011 | Sub C->T | 59(0.50) | 24 (1.00) | | |
| | | | 3651013 | Del C | 6 (1.00) | | 6 (1.00) | |
| | | | 3,651,005 | Ins T | | 6(1.00) | | |
| | | | 3,651,006 | Ins T | | 6(1.00) | | |
| | | | 3,651,007 | Ins T | | 6(1.00) | | |
| | | | 3,651,008 | Ins T | | 6(1.00) | | |
| BAS4278 | peptidase, U32 family | 4188897-4189826 | 4188979 | Sub T->A | | | 13 (0.93) | |
| BAS5207 | Collagen adhesion protein | 5094108-5096756 | 5,095,745 | Ins G | 78(1.00) | | | |
| BAS4585 | FtsK/SpoIIIE family protein | 4485133-4489068 | 4486921 | Sub T->C | | | 3 (0.60) | |
| | | | 4486948 | Sub T->C | | | 6 (0.86) | |
| | | | 4486987 | Sub T->C | | | 8 (1.00) | |
| | | | 4486990 | Ins TT | | | 8 (1.00) | |
| | | | 4486992 | Ins A | | | 8 (1.00) | |
| | | | 4487009 | Sub T->G | | | 8 (1.00) | |
| | | | 4487014 | Sub C->T | | | 4 (0.50) | |

In addition to those mutations detailed above, 454 sequencing also identified a 51bp deletion present in all 4 resistant mutants. The deletion sequence: TAAATATGCCATGAATTATTTAACTGTTATATGAACCAAATAAAAAAAGCATTGCACAA GAGCAATGCTTTTTTTATATATCCCGATCCAAATAAAGAGGTTA was located in an intergenic region from bases 3930400- 3930451.

A large deletion, 100722 bp long, located from positions 741,694 to 842,468 was identified in mutant 10:8:1. This deletion was identified by both sequencing technologies and the microarray analysis.

# Appendix

**Appendix 1:** Number of reads obtained from Illumina sequencing runs for each sample.

| Sample Name | B.a copy # | # of Reads |
|---|---|---|
| BA_aerorsol_spike_1 | 1 copy B.a | 22581638 |
| BA_aerorsol_spike_2 | 10 copy B.a | 20452007 |
| BA_aerorsol_spike_3 | 100 copy B.a. | 22934146 |
| BA_aerorsol_spike_4 | 1000 copy B.a. | 22394992 |
| BA_aerorsol_spike_5 | 10,000 copy B.a | 29822561 |
| BA_aerorsol_spike_6 | 100,000 copy B.a. | 26063900 |
| BA_soil_spike_1 | 1 copy B.a | 54634550 |
| BA_soil_spike_2 | 10 copy B.a | 56495446 |
| BA_soil_spike_3 | 100 copy B.a. | 52082616 |
| BA_soil_spike_4 | 1000 copy B.a. | 54173124 |
| BA_soil_spike_5 | 10,000 copy B.a | 56620796 |
| BA_soil_spike_6 | 100,000 copy B.a. | 61680142 |

**Appendix 2:** Number of Illumina reads mapped for each sample. The *B. anthracis* reference genome contains both pXO1 and pXO2 plasmids and is referred to as the target.

| # Of reads Mapped to reference | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6 |
|---|---|---|---|---|---|---|
| B.a copy # | 1 copy B.a | 10 copy B.a | 100 copy B.a | 1000 copy B.a | 10,000 copy B.a | 10,0000 copy B.a |
| Aerosol mapped to background | 54317 | 40622 | 68825 | 77827 | 115857 | 123711 |
| Aerosol mapped to target | 28142 | 196147 | 2228084 | 13433595 | 28183408 | 25650528 |
| Soil mapped to background | 17633 | 31490 | 18134 | 29355 | 106003 | 293616 |
| Soil mapped to target | 5424 | 27129 | 142562 | 2383829 | 15914018 | 48366091 |
| | | | | | | |
| % Of total reads mapped to reference | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6 |
| Aerosol mapped to background | 0.240536 | 0.198621 | 0.300098 | 0.34752 | 0.388488 | 0.474645 |
| Aerosol mapped to target | 0.124623 | 0.95906 | 9.715138 | 59.98482 | 94.50365 | 98.414006 |
| Soil mapped to background | 0.032274 | 0.055739 | 0.034818 | 0.054187 | 0.187216 | 0.47603 |
| Soil mapped to target | 0.009928 | 0.04802 | 0.273723 | 4.40039 | 28.10631 | 78.414364 |

**Appendix 3:** Data for the read mapping to each of the sequences for determining the discrimination between *B. anthracis* and *B. thuringiensis* Al Hakam from Illumina sequencing.

| Sample Name | Organism | # Reads mapped | Total Reads | Percent Reads mapped |
|---|---|---|---|---|
| BA_aerosol_spike_1 | Bacillus thuringiensis str. Al Hakam | 21028 | 22581638 | 0.093119906 |
| BA_aerosol_spike_1 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 0 | 22581638 | 0 |
| BA_aerosol_spike_1 | Bacillus cereus biovar anthracis str. Cl | 24258 | 22581638 | 0.107423562 |
| BA_aerosol_spike_1 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 104 | 22581638 | 0.000460551 |
| BA_aerosol_spike_1 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 44 | 22581638 | 0.000194849 |
| BA_aerosol_spike_1 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 22581638 | 0 |
| BA_aerosol_spike_1 | Bacillus anthracis virulence plasmid PX01 | 104 | 22581638 | 0.000460551 |
| BA_aerosol_spike_1 | Bacillus anthracis plasmid pX02 | 44 | 22581638 | 0.000194849 |
| BA_aerosol_spike_1 | Bacillus anthracis str. Ames | 28018 | 22581638 | 0.124074259 |
| | | | | |
| BA_aerosol_spike_2 | Bacillus thuringiensis str. Al Hakam | 152475 | 20452007 | 0.745525855 |
| BA_aerosol_spike_2 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 11 | 20452007 | 5.38E-05 |
| BA_aerosol_spike_2 | Bacillus cereus biovar anthracis str. Cl | 160455 | 20452007 | 0.78454403 |
| BA_aerosol_spike_2 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 4278 | 20452007 | 0.020917263 |
| BA_aerosol_spike_2 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 990 | 20452007 | 0.004840601 |
| BA_aerosol_spike_2 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 20452007 | 0 |
| BA_aerosol_spike_2 | Bacillus anthracis virulence plasmid PX01 | 4293 | 20452007 | 0.020990605 |
| BA_aerosol_spike_2 | Bacillus anthracis plasmid pX02 | 993 | 20452007 | 0.004855269 |
| BA_aerosol_spike_2 | Bacillus Antracis str. Ames | 190819 | 20452007 | 0.933008677 |
| | | | | |
| BA_aerosol_spike_3 | Bacillus thuringiensis str. Al Hakam | 1732087 | 22934146 | 7.552437313 |
| BA_aerosol_spike_3 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 153 | 22934146 | 0.000667128 |
| BA_aerosol_spike_3 | Bacillus cereus biovar anthracis str. Cl | 1798685 | 22934146 | 7.842825279 |
| BA_aerosol_spike_3 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 68979 | 22934146 | 0.300769865 |
| BA_aerosol_spike_3 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 12862 | 22934146 | 0.056082315 |
| BA_aerosol_spike_3 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 22934146 | 0 |
| BA_aerosol_spike_3 | Bacillus anthracis virulence plasmid PX01 | 69235 | 22934146 | 0.301886105 |
| BA_aerosol_spike_3 | Bacillus anthracis plasmid pX02 | 12962 | 22934146 | 0.056518346 |
| BA_aerosol_spike_3 | Bacillus Antracis str. Ames | 2146998 | 22934146 | 9.361578146 |
| | | | | |
| BA_aerosol_spike_4 | Bacillus thuringiensis str. Al Hakam | 10338006 | 22394992 | 46.16213303 |
| BA_aerosol_spike_4 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 1333 | 22394992 | 0.005952224 |
| BA_aerosol_spike_4 | Bacillus cereus biovar anthracis str. Cl | 10740832 | 22394992 | 47.96086554 |
| BA_aerosol_spike_4 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 433639 | 22394992 | 1.936321299 |
| BA_aerosol_spike_4 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 105033 | 22394992 | 0.469002177 |
| BA_aerosol_spike_4 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 22394992 | 0 |
| BA_aerosol_spike_4 | Bacillus anthracis virulence plasmid PX01 | 435435 | 22394992 | 1.944340949 |
| BA_aerosol_spike_4 | Bacillus anthracis plasmid pX02 | 106363 | 22394992 | 0.474941005 |

| | | | | |
|---|---|---|---|---|
| BA_aerosol_spike_4 | Bacillus Antracis str. Ames | 12897471 | 22394992 | 57.59087121 |
| | | | | |
| BA_aerosol_spike_5 | Bacillus thuringiensis str. Al Hakam | 21554078 | 29822561 | 72.27440326 |
| BA_aerosol_spike_5 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 2632 | 29822561 | 0.008825533 |
| BA_aerosol_spike_5 | Bacillus cereus biovar anthracis str. Cl | 22384277 | 29822561 | 75.05819839 |
| BA_aerosol_spike_5 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 1044558 | 29822561 | 3.502576455 |
| BA_aerosol_spike_5 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 215986 | 29822561 | 0.724236929 |
| BA_aerosol_spike_5 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 29822561 | 0 |
| BA_aerosol_spike_5 | Bacillus anthracis virulence plasmid PX01 | 1048466 | 29822561 | 3.515680629 |
| BA_aerosol_spike_5 | Bacillus anthracis plasmid pX02 | 218224 | 29822561 | 0.731741315 |
| BA_aerosol_spike_5 | Bacillus Antracis str. Ames | 26930475 | 29822561 | 90.30235532 |
| | | | | |
| BA_aerosol_spike_6 | Bacillus thuringiensis str. Al Hakam | 19654381 | 26063900 | 75.40844233 |
| BA_aerosol_spike_6 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 2598 | 26063900 | 0.00996781 |
| BA_aerosol_spike_6 | Bacillus cereus biovar anthracis str. Cl | 20386611 | 26063900 | 78.21780701 |
| BA_aerosol_spike_6 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 970305 | 26063900 | 3.722792828 |
| BA_aerosol_spike_6 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 221192 | 26063900 | 0.848652734 |
| BA_aerosol_spike_6 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 26063900 | 0 |
| BA_aerosol_spike_6 | Bacillus anthracis virulence plasmid PX01 | 973941 | 26063900 | 3.736743158 |
| BA_aerosol_spike_6 | Bacillus anthracis plasmid pX02 | 223447 | 26063900 | 0.857304548 |
| BA_aerosol_spike_6 | Bacillus Antracis str. Ames | 24465424 | 26063900 | 93.86708819 |
| | | | | |
| BA_soil_spike_1 | Bacillus thuringiensis str. Al Hakam | 5245 | 54634550 | 0.0096002 |
| BA_soil_spike_1 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 0 | 54634550 | 0.0000000 |
| BA_soil_spike_1 | Bacillus cereus biovar anthracis str. Cl | 5312 | 54634550 | 0.0097228 |
| BA_soil_spike_1 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 8 | 54634550 | 0.0000146 |
| BA_soil_spike_1 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 2 | 54634550 | 0.0000037 |
| BA_soil_spike_1 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 54634550 | 0.0000000 |
| BA_soil_spike_1 | Bacillus anthracis virulence plasmid PX01 | 8 | 54634550 | 0.0000146 |
| BA_soil_spike_1 | Bacillus anthracis plasmid pX02 | 2 | 54634550 | 0.0000037 |
| BA_soil_spike_1 | Bacillus Antracis str. Ames | 5380 | 54634550 | 0.0098472 |
| | | | | |
| BA_soil_spike_2 | Bacillus thuringiensis str. Al Hakam | 23459 | 56495446 | 0.0415237 |
| BA_soil_spike_2 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 1 | 56495446 | 0.0000018 |
| BA_soil_spike_2 | Bacillus cereus biovar anthracis str. Cl | 23667 | 56495446 | 0.0418919 |
| BA_soil_spike_2 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 464 | 56495446 | 0.0008213 |
| BA_soil_spike_2 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 894 | 56495446 | 0.0015824 |
| BA_soil_spike_2 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 56495446 | 0.0000000 |
| BA_soil_spike_2 | Bacillus anthracis virulence plasmid PX01 | 463 | 56495446 | 0.0008195 |
| BA_soil_spike_2 | Bacillus anthracis plasmid pX02 | 900 | 56495446 | 0.0015930 |
| BA_soil_spike_2 | Bacillus Antracis str. Ames | 25646 | 56495446 | 0.0453948 |
| | | | | |
| BA_soil_spike_3 | Bacillus thuringiensis str. Al Hakam | 111685 | 52082616 | 0.2144382 |

| | | | | |
|---|---|---|---|---|
| BA_soil_spike_3 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 7 | 52082616 | 0.0000134 |
| BA_soil_spike_3 | Bacillus cereus biovar anthracis str. Cl | 115592 | 52082616 | 0.2219397 |
| BA_soil_spike_3 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 3652 | 52082616 | 0.0070119 |
| BA_soil_spike_3 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 1235 | 52082616 | 0.0023712 |
| BA_soil_spike_3 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 52082616 | 0.0000000 |
| BA_soil_spike_3 | Bacillus anthracis virulence plasmid PX01 | 3656 | 52082616 | 0.0070196 |
| BA_soil_spike_3 | Bacillus anthracis plasmid pX02 | 1249 | 52082616 | 0.0023981 |
| BA_soil_spike_3 | Bacillus Antracis str. Ames | 137703 | 52082616 | 0.2643934 |
| | | | | |
| BA_soil_spike_4 | Bacillus thuringiensis str. Al Hakam | 1834774 | 54173124 | 3.3868713 |
| BA_soil_spike_4 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 272 | 54173124 | 0.0005021 |
| BA_soil_spike_4 | Bacillus cereus biovar anthracis str. Cl | 1902457 | 54173124 | 3.5118097 |
| BA_soil_spike_4 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 90660 | 54173124 | 0.1673524 |
| BA_soil_spike_4 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 20051 | 54173124 | 0.0370128 |
| BA_soil_spike_4 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 54173124 | 0.0000000 |
| BA_soil_spike_4 | Bacillus anthracis virulence plasmid PX01 | 91011 | 54173124 | 0.1680003 |
| BA_soil_spike_4 | Bacillus anthracis plasmid pX02 | 20249 | 54173124 | 0.0373783 |
| BA_soil_spike_4 | Bacillus Antracis str. Ames | 2273779 | 54173124 | 4.1972455 |
| | | | | |
| BA_soil_spike_5 | Bacillus thuringiensis str. Al Hakam | 12286357 | 56620796 | 21.6993717 |
| BA_soil_spike_5 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 1645 | 56620796 | 0.0029053 |
| BA_soil_spike_5 | Bacillus cereus biovar anthracis str. Cl | 12735128 | 56620796 | 22.4919621 |
| BA_soil_spike_5 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 573200 | 56620796 | 1.0123489 |
| BA_soil_spike_5 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 123965 | 56620796 | 0.2189390 |
| BA_soil_spike_5 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 56620796 | 0.0000000 |
| BA_soil_spike_5 | Bacillus anthracis virulence plasmid PX01 | 575271 | 56620796 | 1.0160066 |
| BA_soil_spike_5 | Bacillus anthracis plasmid pX02 | 125159 | 56620796 | 0.2210478 |
| BA_soil_spike_5 | Bacillus Antracis str. Ames | 15220855 | 56620796 | 26.8820929 |
| | | | | |
| BA_soil_spike_6 | Bacillus thuringiensis str. Al Hakam | 37461447 | 61680142 | 60.7350207 |
| BA_soil_spike_6 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1 | 5114 | 61680142 | 0.0082912 |
| BA_soil_spike_6 | Bacillus cereus biovar anthracis str. Cl | 38782845 | 61680142 | 62.8773601 |
| BA_soil_spike_6 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X01 | 1711505 | 61680142 | 2.7748072 |
| BA_soil_spike_6 | Bacillus cereus biovar anthracis str. Cl plasmid pCl-X02 | 416162 | 61680142 | 0.6747099 |
| BA_soil_spike_6 | Bacillus cereus biovar anthracis str. Cl plasmid pBAslCl14 | 0 | 61680142 | 0.0000000 |
| BA_soil_spike_6 | Bacillus anthracis virulence plasmid PX01 | 1717387 | 61680142 | 2.7843435 |
| BA_soil_spike_6 | Bacillus anthracis plasmid pX02 | 420075 | 61680142 | 0.6810539 |
| BA_soil_spike_6 | Bacillus Antracis str. Ames | 46249059 | 61680142 | 74.9820890 |

**Appendix 4:** Number of reads obtained from sequencing runs for each sample from 454 sequencing.

| Sample | Total Reads |
|---|---|
| Aero1 | 230562 |
| Aero2 | 237769 |
| Aero3 | 73751 |
| Aero4 | 289578 |
| Soil1 | 142707 |
| Soil2 | 208806 |
| Soil3 | 128179 |
| Soil4 | 100298 |

**Appendix 5** Number of reads mapped for each sample from 454 sequencing. The *B. anthracis* reference genome contains both pXO1 and pXO2 plasmids.

| | B. anthracis reads mapped | | Background set | |
|---|---|---|---|---|
| Sample | Reads Mapped | Percent Mapped | Reads Mapped | Percent Mapped |
| Aero1 | 5 | 0.00% | 61 | 0.03% |
| Aero2 | 237 | 0.10% | 271 | 0.11% |
| Aero3 | 697 | 0.95% | 4 | 0.01% |
| Aero4 | 23039 | 7.96% | 85 | 0.03% |
| Soil1 | 6 | 0.00% | 106 | 0.07% |
| Soil2 | 35 | 0.02% | 15 | 0.01% |
| Soil3 | 242 | 0.19% | 7 | 0.01% |
| Soil4 | 3176 | 3.17% | 21 | 0.02% |

**Appendix 6:** 454 sequencing data for the read mapping to each of the sequences for determining the discrimination between *B. anthracis* and *B. thuringiensis* Al Hakam.

| Name | Total Reads | Number of Reads Used | Percent Reads Mapped | Reference Accession | Num Unique Matching Reads | Percent of All Unique Matches | Percent of All Reads | Percent Coverage of Reference | Description |
|---|---|---|---|---|---|---|---|---|---|
| Aero1.fna | 230562 | 34 | 0.01 | AE016879.1 | 2 | 50 | 0.000867 | 6.95 | Bacillus anthracis str. Ames, complete genome |
| Aero1.fna | 230562 | 34 | 0.01 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero1.fna | 230562 | 34 | 0.01 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero1.fna | 230562 | 34 | 0.01 | CP000485.1 | 2 | 50 | 0.000867 | 6.3 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Aero1.fna | 230562 | 34 | 0.01 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Aero2.fna | 237769 | 239 | 0.1 | AE016879.1 | 158 | 99.4 | 0.066451 | 26.64 | Bacillus anthracis str. Ames, complete genome |
| Aero2.fna | 237769 | 239 | 0.1 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero2.fna | 237769 | 239 | 0.1 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero2.fna | 237769 | 239 | 0.1 | CP000485.1 | 1 | 0.6 | 0.000421 | 5.77 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Aero2.fna | 237769 | 239 | 0.1 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Aero3.fna | 73751 | 656 | 0.89 | AE016879.1 | 478 | 89.5 | 0.648127 | 23.18 | Bacillus anthracis str. Ames, complete genome |
| Aero3.fna | 73751 | 656 | 0.89 | AE017336.2 | 43 | 8.1 | 0.058304 | 27.42 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero3.fna | 73751 | 656 | 0.89 | AE017335.3 | 10 | 1.9 | 0.013559 | 17.96 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero3.fna | 73751 | 656 | 0.89 | CP000485.1 | 3 | 0.6 | 0.004068 | 5.1 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Aero3.fna | 73751 | 656 | 0.89 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Aero4.fna | 289578 | 21710 | 7.51 | AE016879.1 | 16874 | 95.9 | 5.827100 | 53.43 | Bacillus anthracis str. Ames, complete genome |
| Aero4.fna | 289578 | 21710 | 7.51 | AE017336.2 | 573 | 3.3 | 0.197874 | 57.08 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero4.fna | 289578 | 21710 | 7.51 | AE017335.3 | 108 | 0.6 | 0.037296 | 34.13 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero4.fna | 289578 | 21710 | 7.51 | CP000485.1 | 44 | 0.3 | 0.015195 | 7.33 | Bacillus thuringiensis str. Al Hakam, complete genome |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Aero4.fna | 289578 | 21710 | 7.51 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Soil1.fna | 142707 | 28 | 0.02 | AE016879.1 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. Ames, complete genome |
| Soil1.fna | 142707 | 28 | 0.02 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil1.fna | 142707 | 28 | 0.02 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil1.fna | 142707 | 28 | 0.02 | CP000485.1 | 1 | 100 | 0.000701 | 10.8 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Soil1.fna | 142707 | 28 | 0.02 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Soil2.fna | 208806 | 44 | 0.02 | AE016879.1 | 25 | 96.2 | 0.011973 | 21.29 | Bacillus anthracis str. Ames, complete genome |
| Soil2.fna | 208806 | 44 | 0.02 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil2.fna | 208806 | 44 | 0.02 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil2.fna | 208806 | 44 | 0.02 | CP000485.1 | 1 | 3.8 | 0.000479 | 5.7 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Soil2.fna | 208806 | 44 | 0.02 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Soil3.fna | 128179 | 245 | 0.19 | AE016879.1 | 175 | 95.6 | 0.136528 | 19.38 | Bacillus anthracis str. Ames, complete genome |
| Soil3.fna | 128179 | 245 | 0.19 | AE017336.2 | 6 | 3.3 | 0.004681 | 16.6 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil3.fna | 128179 | 245 | 0.19 | AE017335.3 | 1 | 0.5 | 0.000780 | 27.92 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil3.fna | 128179 | 245 | 0.19 | CP000485.1 | 1 | 0.5 | 0.000780 | 4.33 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Soil3.fna | 128179 | 245 | 0.19 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |
| Soil4.fna | 100298 | 3109 | 3.11 | AE016879.1 | 2237 | 92.8 | 2.230354 | 24.79 | Bacillus anthracis str. Ames, complete genome |
| Soil4.fna | 100298 | 3109 | 3.11 | AE017336.2 | 141 | 5.9 | 0.140581 | 29.14 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil4.fna | 100298 | 3109 | 3.11 | AE017335.3 | 26 | 1.1 | 0.025923 | 18.72 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil4.fna | 100298 | 3109 | 3.11 | CP000485.1 | 6 | 0.2 | 0.005982 | 5.26 | Bacillus thuringiensis str. Al Hakam, complete genome |
| Soil4.fna | 100298 | 3109 | 3.11 | CP000486.1 | 0 | 0 | 0.000000 | 0 | Bacillus thuringiensis str. Al Hakam, plasmid pALH1, complete sequence |

**Appendix 7:** 454 sequencing data for the read mapping to each of the sequences for determining the discrimination between *B. anthracis* and *B. cereus* biovar anthracis.

| Name | Total Reads | Number of Reads Used | Percent Reads Mapped | Reference Accession | Num Unique Matching Reads | Percent of All Unique Matches | Percent of All Reads | Percent Coverage of Reference | Description |
|---|---|---|---|---|---|---|---|---|---|
| Aero1.fna | 230562 | 32 | 0.01 | AE016879.1 | 2 | 100 | 0.000867 | 6.95 | Bacillus anthracis str. Ames, complete genome |
| Aero1.fna | 230562 | 32 | 0.01 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero1.fna | 230562 | 32 | 0.01 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero1.fna | 230562 | 32 | 0.01 | CP001746.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Aero1.fna | 230562 | 32 | 0.01 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero1.fna | 230562 | 32 | 0.01 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero1.fna | 230562 | 32 | 0.01 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Aero2.fna | 237769 | 240 | 0.1 | AE016879.1 | 151 | 97.4 | 0.063507 | 27.3 | Bacillus anthracis str. Ames, complete genome |
| Aero2.fna | 237769 | 240 | 0.1 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero2.fna | 237769 | 240 | 0.1 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero2.fna | 237769 | 240 | 0.1 | CP001746.1 | 4 | 2.6 | 0.001682 | 18.65 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Aero2.fna | 237769 | 240 | 0.1 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero2.fna | 237769 | 240 | 0.1 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero2.fna | 237769 | 240 | 0.1 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Aero3.fna | 73751 | 657 | 0.89 | AE016879.1 | 469 | 98.1 | 0.635924 | 23.03 | Bacillus anthracis str. Ames, complete genome |
| Aero3.fna | 73751 | 657 | 0.89 | AE017336.2 | 6 | 1.3 | 0.008135 | 21.51 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Aero3.fna | 73751 | 657 | 0.89 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero3.fna | 73751 | 657 | 0.89 | CP001746.1 | 3 | 0.6 | 0.004068 | 6.61 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Aero3.fna | 73751 | 657 | 0.89 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero3.fna | 73751 | 657 | 0.89 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero3.fna | 73751 | 657 | 0.89 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Aero4.fna | 289578 | 21830 | 7.55 | AE016879.1 | 16102 | 99.3 | 5.560505 | 52.26 | Bacillus anthracis str. Ames, complete genome |
| Aero4.fna | 289578 | 21830 | 7.55 | AE017336.2 | 39 | 0.2 | 0.013468 | 19.72 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Aero4.fna | 289578 | 21830 | 7.55 | AE017335.3 | 19 | 0.1 | 0.006561 | 21.52 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Aero4.fna | 289578 | 21830 | 7.55 | CP001746.1 | 48 | 0.3 | 0.016576 | 9.72 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Aero4.fna | 289578 | 21830 | 7.55 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero4.fna | 289578 | 21830 | 7.55 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Aero4.fna | 289578 | 21830 | 7.55 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Soil1.fna | 142707 | 31 | 0.02 | AE016879.1 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. Ames, complete genome |
| Soil1.fna | 142707 | 31 | 0.02 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil1.fna | 142707 | 31 | 0.02 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil1.fna | 142707 | 31 | 0.02 | CP001746.1 | 2 | 100 | 0.001401 | 5.8 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Soil1.fna | 142707 | 31 | 0.02 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil1.fna | 142707 | 31 | 0.02 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil1.fna | 142707 | 31 | 0.02 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Soil2.fna | 208806 | 45 | 0.02 | AE016879.1 | 29 | 93.5 | 0.013888 | 22.85 | Bacillus anthracis str. Ames, complete |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | genome |
| Soil2.fna | 208806 | 45 | 0.02 | AE017336.2 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil2.fna | 208806 | 45 | 0.02 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil2.fna | 208806 | 45 | 0.02 | CP001746.1 | 2 | 6.5 | 0.000958 | 2.57 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Soil2.fna | 208806 | 45 | 0.02 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil2.fna | 208806 | 45 | 0.02 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil2.fna | 208806 | 45 | 0.02 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Soil3.fna | 128179 | 244 | 0.19 | AE016879.1 | 172 | 98.9 | 0.134187 | 18.81 | Bacillus anthracis str. Ames, complete genome |
| Soil3.fna | 128179 | 244 | 0.19 | AE017336.2 | 1 | 0.6 | 0.000780 | 21.18 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil3.fna | 128179 | 244 | 0.19 | AE017335.3 | 0 | 0 | 0.000000 | 0 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil3.fna | 128179 | 244 | 0.19 | CP001746.1 | 1 | 0.6 | 0.000780 | 4.33 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Soil3.fna | 128179 | 244 | 0.19 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil3.fna | 128179 | 244 | 0.19 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil3.fna | 128179 | 244 | 0.19 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |
| Soil4.fna | 100298 | 3129 | 3.13 | AE016879.1 | 2126 | 98.3 | 2.119683 | 24.56 | Bacillus anthracis str. Ames, complete genome |
| Soil4.fna | 100298 | 3129 | 3.13 | AE017336.2 | 16 | 0.7 | 0.015952 | 24.88 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1, complete sequence |
| Soil4.fna | 100298 | 3129 | 3.13 | AE017335.3 | 2 | 0.1 | 0.001994 | 16.89 | Bacillus anthracis str. 'Ames Ancestor' plasmid pXO2, complete sequence |
| Soil4.fna | 100298 | 3129 | 3.13 | CP001746.1 | 17 | 0.8 | 0.016949 | 7.05 | Bacillus cereus biovar anthracis str. CI, complete genome |
| Soil4.fna | 100298 | 3129 | 3.13 | CP001748.1 | 1 | 0 | 0.000997 | 14.73 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |
| Soil4.fna | 100298 | 3129 | 3.13 | CP001748.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pCI-XO2, complete sequence |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Soil4.fna | 100298 | 3129 | 3.13 | CP001749.1 | 0 | 0 | 0.000000 | 0 | Bacillus cereus biovar anthracis str. CI plasmid pBAslCI14, complete sequence |

**Appendix 8:** The tables contain the union of the top 25 taxa in each aerosol sample and the union of the top 25 taxa in each soil sample from 454 sequencing. The complete set of data can be found in these supplemental files:

Aero1.fna.br.gi.taxname.counts.readnormalized

Aero2.fna.br.gi.taxname.counts.readnormalized

Aero3.fna.br.gi.taxname.counts.readnormalized

Aero4.fna.br.gi.taxname.counts.readnormalized

Soil1.fna.br.gi.taxname.counts.readnormalized

Soil2.fna.br.gi.taxname.counts.readnormalized

Soil3.fna.br.gi.taxname.counts.readnormalized

Soil4.fna.br.gi.taxname.counts.readnormalized

**Table of the union of the top 25 taxa for the four aerosol samples, 454 sequencing:**

| | Aero1 | Aero2 | Aero3 | Aero4 |
|---|---|---|---|---|
| Agrobacterium radiobacter | 0.00E+00 | 1.74E-03 | 0.00E+00 | 6.91E-06 |
| Alternaria alternata | 4.55E-04 | 2.78E-04 | 1.69E-03 | 1.02E-03 |
| Alternaria solani | 4.34E-05 | 4.79E-04 | 1.22E-04 | 1.04E-05 |
| Artemisia annua | 2.69E-04 | 1.60E-03 | 4.07E-05 | 1.11E-04 |
| Aspergillus clavatus | 2.99E-04 | 3.45E-04 | 2.71E-04 | 6.11E-04 |
| Aspergillus niger | 5.51E-04 | 1.12E-03 | 3.93E-04 | 4.18E-04 |
| Aspergillus terreus | 4.03E-04 | 2.90E-04 | 2.58E-04 | 4.94E-04 |
| Bacillus anthracis | 0.00E+00 | 7.95E-04 | 7.53E-03 | 6.66E-02 |
| Bacillus cereus | 8.67E-06 | 2.10E-04 | 1.93E-03 | 1.31E-02 |
| Botryotinia fuckeliana | 1.01E-03 | 4.37E-04 | 5.83E-04 | 9.95E-04 |
| Castanea mollissima | 2.72E-03 | 7.61E-04 | 6.92E-04 | 4.97E-04 |
| Chaetomium globosum | 3.38E-04 | 7.28E-04 | 4.20E-04 | 4.21E-04 |
| Cupriavidus metallidurans | 5.86E-04 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Cupriavidus taiwanensis | 5.90E-04 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Delftia acidovorans | 2.13E-03 | 4.42E-03 | 1.25E-02 | 5.20E-03 |
| Gibberella zeae | 1.14E-03 | 1.91E-03 | 1.11E-03 | 1.03E-03 |
| Kabatiella microsticta | 2.04E-04 | 6.60E-04 | 1.08E-04 | 1.14E-04 |
| Klebsiella pneumoniae | 1.73E-05 | 0.00E+00 | 1.36E-05 | 9.84E-04 |
| Leptosphaeria maculans | 2.29E-03 | 2.96E-03 | 2.49E-03 | 1.79E-03 |

| | | | | |
|---|---|---|---|---|
| Medicago truncatula | 1.26E-04 | 3.79E-05 | 3.53E-04 | 1.38E-05 |
| Microbotryum violaceum | 1.56E-04 | 1.18E-03 | 4.07E-05 | 1.62E-04 |
| Nectria haematococca | 7.81E-04 | 9.55E-04 | 7.32E-04 | 6.35E-04 |
| Neurospora crassa | 7.55E-04 | 1.12E-03 | 4.88E-04 | 8.56E-04 |
| Oryza sativa | 1.10E-03 | 1.11E-03 | 6.10E-04 | 7.94E-04 |
| Pantoea agglomerans | 8.63E-04 | 3.07E-04 | 2.98E-04 | 3.45E-04 |
| Pantoea vagans | 8.80E-04 | 6.88E-03 | 1.90E-04 | 1.27E-03 |
| Penicillium chrysogenum | 1.49E-03 | 1.12E-03 | 9.36E-04 | 1.41E-03 |
| Phaeosphaeria avenaria | 6.07E-05 | 6.73E-05 | 4.07E-05 | 4.97E-04 |
| Phaeosphaeria nodorum | 3.31E-03 | 4.10E-03 | 2.62E-03 | 3.15E-03 |
| Picea glauca | 5.25E-04 | 5.93E-04 | 6.78E-04 | 2.52E-04 |
| Pinus taeda | 9.36E-03 | 5.01E-03 | 3.67E-03 | 3.19E-03 |
| Podospora anserina | 4.99E-04 | 5.34E-04 | 7.73E-04 | 4.56E-04 |
| Pyrenophora tritici-repentis | 5.05E-03 | 9.28E-03 | 4.50E-03 | 5.07E-03 |
| Quercus nigra | 4.99E-04 | 2.36E-04 | 1.90E-04 | 1.93E-04 |
| Ralstonia eutropha | 1.18E-03 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Ralstonia pickettii | 2.56E-03 | 5.89E-05 | 5.42E-05 | 4.14E-05 |
| Ralstonia solanacearum | 8.98E-04 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Root | 3.69E-04 | 4.96E-04 | 9.90E-04 | 4.04E-04 |
| Salmonella enterica | 2.82E-04 | 0.00E+00 | 5.42E-05 | 5.01E-04 |
| Sclerotinia sclerotiorum | 5.73E-04 | 3.83E-04 | 3.93E-04 | 2.35E-04 |
| Stenotrophomonas maltophilia | 4.34E-05 | 1.35E-04 | 5.02E-03 | 4.03E-03 |
| Verticillium albo-atrum | 4.25E-04 | 5.13E-04 | 6.10E-04 | 4.97E-04 |
| Vitis vinifera | 1.32E-03 | 4.96E-04 | 8.54E-04 | 4.70E-04 |

**Table of the union of the top 25 taxa for the soil samples, 454 sequencing:**

| | Soil1 | Soil2 | Soil3 | Soil4 |
|---|---|---|---|---|
| Acidobacterium capsulatum | 1.68E-04 | 6.70E-05 | 3.12E-05 | 1.99E-05 |
| Agkistrodon contortrix | 7.01E-05 | 2.39E-05 | 0.00E+00 | 9.97E-06 |
| Artemisia annua | 2.80E-05 | 9.58E-06 | 1.56E-05 | 3.99E-05 |
| Arthrobacter sp. | 1.05E-04 | 2.44E-04 | 1.09E-04 | 7.98E-05 |
| Bacillus anthracis | 0.00E+00 | 1.53E-04 | 1.61E-03 | 2.63E-02 |
| Bacillus cereus | 0.00E+00 | 1.44E-05 | 2.57E-04 | 5.76E-03 |
| Bacillus megaterium | 1.47E-04 | 1.92E-05 | 7.80E-06 | 3.49E-04 |
| Bacillus sp. | 1.40E-05 | 0.00E+00 | 0.00E+00 | 7.98E-05 |
| Bacillus thuringiensis | 0.00E+00 | 0.00E+00 | 1.56E-05 | 6.98E-05 |
| Botryotinia fuckeliana | 9.81E-05 | 4.79E-06 | 7.02E-05 | 0.00E+00 |
| Bradyrhizobium japonicum | 1.96E-04 | 1.44E-05 | 1.56E-05 | 0.00E+00 |
| Candidatus Koribacter | 7.71E-05 | 1.44E-05 | 7.80E-05 | 9.97E-06 |
| Caulobacter sp. | 0.00E+00 | 0.00E+00 | 0.00E+00 | 4.99E-05 |
| Cloning vector | 2.10E-05 | 5.27E-05 | 8.58E-05 | 0.00E+00 |
| Crepidula fornicata | 3.92E-04 | 7.18E-05 | 7.80E-05 | 0.00E+00 |
| Cupriavidus metallidurans | 5.33E-04 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Cupriavidus taiwanensis | 4.41E-04 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Danio rerio | 5.36E-03 | 7.61E-04 | 9.75E-04 | 4.29E-04 |
| Delftia acidovorans | 7.01E-05 | 1.92E-05 | 1.17E-04 | 6.98E-05 |
| Dictyostelium discoideum | 3.08E-04 | 4.79E-05 | 4.68E-05 | 3.99E-05 |
| Dimastigella mimosa | 3.93E-03 | 6.25E-03 | 8.12E-03 | 7.54E-03 |
| Escherichia coli | 2.10E-05 | 4.79E-06 | 0.00E+00 | 4.99E-05 |
| Haliangium ochraceum | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.10E-04 |
| Homo sapiens | 2.73E-04 | 8.14E-05 | 1.09E-04 | 9.97E-05 |
| Monosiga brevicollis | 3.36E-04 | 6.70E-05 | 7.80E-05 | 0.00E+00 |
| Mus musculus | 3.99E-03 | 6.18E-04 | 8.35E-04 | 4.79E-04 |
| Mus spretus | 1.12E-04 | 9.58E-06 | 8.58E-05 | 9.97E-06 |
| Nitrosospira multiformis | 3.97E-03 | 1.29E-03 | 1.15E-03 | 7.48E-04 |
| Oncorhynchus mykiss | 5.47E-04 | 7.66E-05 | 1.56E-04 | 3.99E-05 |
| Peronosclerospora sorghi | 0.00E+00 | 0.00E+00 | 5.46E-05 | 0.00E+00 |
| Plakobranchus ocellatus | 2.31E-04 | 5.75E-05 | 5.46E-05 | 2.99E-05 |
| Platanus occidentalis | 2.31E-04 | 6.70E-05 | 1.33E-04 | 9.97E-05 |
| Platanus racemosa | 3.50E-05 | 4.31E-05 | 4.68E-05 | 3.99E-05 |
| Polaromonas sp. | 2.10E-05 | 2.87E-05 | 0.00E+00 | 0.00E+00 |
| Propionibacterium acnes | 0.00E+00 | 0.00E+00 | 0.00E+00 | 2.49E-04 |
| Pseudomonas fluorescens | 2.87E-04 | 1.82E-04 | 2.57E-04 | 1.99E-05 |
| Ralstonia eutropha | 7.22E-04 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Ralstonia pickettii | 2.77E-03 | 0.00E+00 | 0.00E+00 | 0.00E+00 |
| Ralstonia solanacearum | 1.02E-03 | 0.00E+00 | 7.80E-06 | 0.00E+00 |

| | | | | |
|---|---|---|---|---|
| Rattus norvegicus | 4.98E-04 | 6.70E-05 | 6.24E-05 | 1.79E-04 |
| Rhizobium leguminosarum | 4.20E-05 | 5.27E-05 | 0.00E+00 | 0.00E+00 |
| Rhodococcus erythropolis | 0.00E+00 | 0.00E+00 | 7.80E-05 | 0.00E+00 |
| Rhodopseudomonas palustris | 2.17E-04 | 9.58E-06 | 7.80E-06 | 0.00E+00 |
| Salmo salar | 3.22E-04 | 3.83E-05 | 3.90E-05 | 1.99E-05 |
| Schistosoma mansoni | 2.10E-04 | 2.39E-05 | 7.80E-06 | 9.97E-06 |
| Sphingobium yanoikuyae | 7.01E-06 | 0.00E+00 | 0.00E+00 | 6.98E-05 |
| Sphingomonas wittichii | 1.40E-05 | 9.58E-06 | 0.00E+00 | 1.20E-04 |
| Spirosoma linguale | 0.00E+00 | 0.00E+00 | 0.00E+00 | 5.58E-03 |
| Staphylococcus aureus | 0.00E+00 | 4.79E-06 | 0.00E+00 | 4.99E-05 |
| Staphylococcus epidermidis | 0.00E+00 | 0.00E+00 | 0.00E+00 | 9.97E-05 |
| uncultured bacterium | 1.75E-04 | 4.31E-05 | 9.36E-05 | 2.19E-04 |
| Vitis vinifera | 1.54E-04 | 1.05E-04 | 7.02E-05 | 1.99E-05 |
| Volvox carteri | 1.68E-04 | 3.83E-05 | 7.80E-06 | 0.00E+00 |

# References


1. Van Ert, M.N., Easterday, W.R., Simonson, T.S., U'Ren, J.M., Pearson, T., Kenefic, L.J., Busch, J.D., Huynh, L.Y., Dukerich, M., Trim, C.B. *et al.* (2007) Strain-Specific Single-Nucleotide Polymorphism Assays for the Bacillus anthracis Ames Strain. *J Clin Microbiol.*, **45**, 47–53.

2. Slezak, T., Kuczmarski, T., Ott, L., Torres, C., Medeiros, D., Smith, J., Truitt, B., Mulakken, N., Lam, M., Vitalis, E., Zemla, A., Zhou, C. E., Gardner, S. N. 2003. Comparative genomics tools applied to bioterrorism defense. *Briefings in Bioinformatics*, June 2003, 4: 133-149.

3. Hill, K.K., Ticknor, L.O., Okinaka, R.T., Asay, M., Blair, H., Bliss, K.A., Laker, M., Pardington, P.E., Richardson, A.P., Tonks, M. *et al.* (2003) Fluorescent Amplified Fragment Length Polymorphism Analysis of Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis Isolates. *Appl. and Environ Microbiol.*, **70**, 1068-1080.

4. Brownstein, M.J., J.D. Carpten, J.R. Smith (1996) Modulation of non-templated nucleotide addition by TaqDNA polymerase: Primer modifications that facilitate genotyping. Biotechniques **20:**1004-1010.

5. Jaing, C., Gardner, S.N., McLoughlin, K., Mulakken, N., Alegria-Hartman, M., Banda, P., Williams, P., Gu, P., Wagner, M., Manohar, C. *et al.* (2008) A functional gene array for detection of bacterial virulence elements. *PLoS ONE*, **3(5)**, e2163. doi:2110.1371/journal.pone.0002163.

6. Allen, J.E., Gardner, S.N. and Slezak, T.R. (2008) DNA signatures for detecting genetic engineering in bacteria. *Genome Biology*, **9**, 56.

7. Gardner, S.N., Kuczmarski, T.A., Vitalis, E.A. and Slezak, T.R. (2003) Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. *Journal of Clinical Microbiology*, **41**, 2417-2427.

8. Gardner, S., Jaing, C., McLoughlin, K. and Slezak, T. (2010) A Microbial Detection Array (MDA) for viral and bacterial detection. *BMC Genomics*, **11**, 668doi:610.1186/1471-2164-1111-1668.

9. Wong, C., Heng, C., Wan Yee, L., Soh, S., Kartasasmita, C., Simoes, E., Hibberd, M., Sung, W.-K. and Miller, L. (2007) Optimization and clinical validation of a pathogen detection microarray. *Genome Biology*, **8**, R93.

10. Giegerich, R., Kurtz, S. and Stoye, J. (2003) Efficient implementation of lazy suffix trees. *Software-Practice and Experience*, **33**, 1035-1049.

11. Rozen, S. and Skaletsky, H. (2000) In Krawetz, S. and Misener, S. (eds.), *Bioinformatics Methods and Protocols: Methods in Molecular Biology* Humana Press, Totowa, NJ, pp. 365-386.

12. Markham, N.R. and Zuker, M. (2005) DNAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577-W581.