LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-489636

Preliminary report for analysis of genome wide mutations from four ciprofloxacin resistant B. anthracis Sterne isolates generated by Illumina, 454 sequencing and microarrays for DHS

C. Jaing, L. Vergez, A. Hinckley, J. Thissen, S. Gardner, K. McLoughlin, P. Jackson

June 23, 2011

Disclaimer

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

LLNL-TR-489636

Preliminary report for analysis of genome wide mutations from four ciprofloxacin resistant *B. anthracis* Sterne isolates generated by Illumina, 454 sequencing and microarrays for DHS

Contributors:
Crystal Jaing, Lisa Vergez, Aubree Hinckley, James Thissen, Shea Gardner, Kevin McLoughlin, Paul Jackson
Lawrence Livermore National Laboratory (LLNL), Livermore, CA

Sally Ellingson, Loren Hauser and Tom Brettin
Computational Biology and Bioinformatics Group
Oak Ridge National Laboratory

Viacheslav Fofanov, Heather Koshinsky
Eureka Genomics, Hercules, CA

Yuriy Fofanov
University of Houston

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

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

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. As the result of a different DHS project, we have selected for and isolated a large number of ciprofloxacin resistant *B. anthracis* Sterne isolates. These isolates vary in the concentrations of ciprofloxacin that they can tolerate, suggesting multiple mutations in the samples. In collaboration with University of Houston, Eureka Genomics and Oak Ridge National Laboratory, we analyzed the ciprofloxacin resistant *B. anthracis* Sterne isolates by microarray hybridization, Illumina and Roche 454 sequencing to understand the error rates and sensitivity of the different methods. The report provides an assessment of the results and a complete set of all protocols used and all data generated along with information to interpret the protocols and data sets.

METHODS

1. *Bacillus anthracis* Sterne ciprofloxacin selections

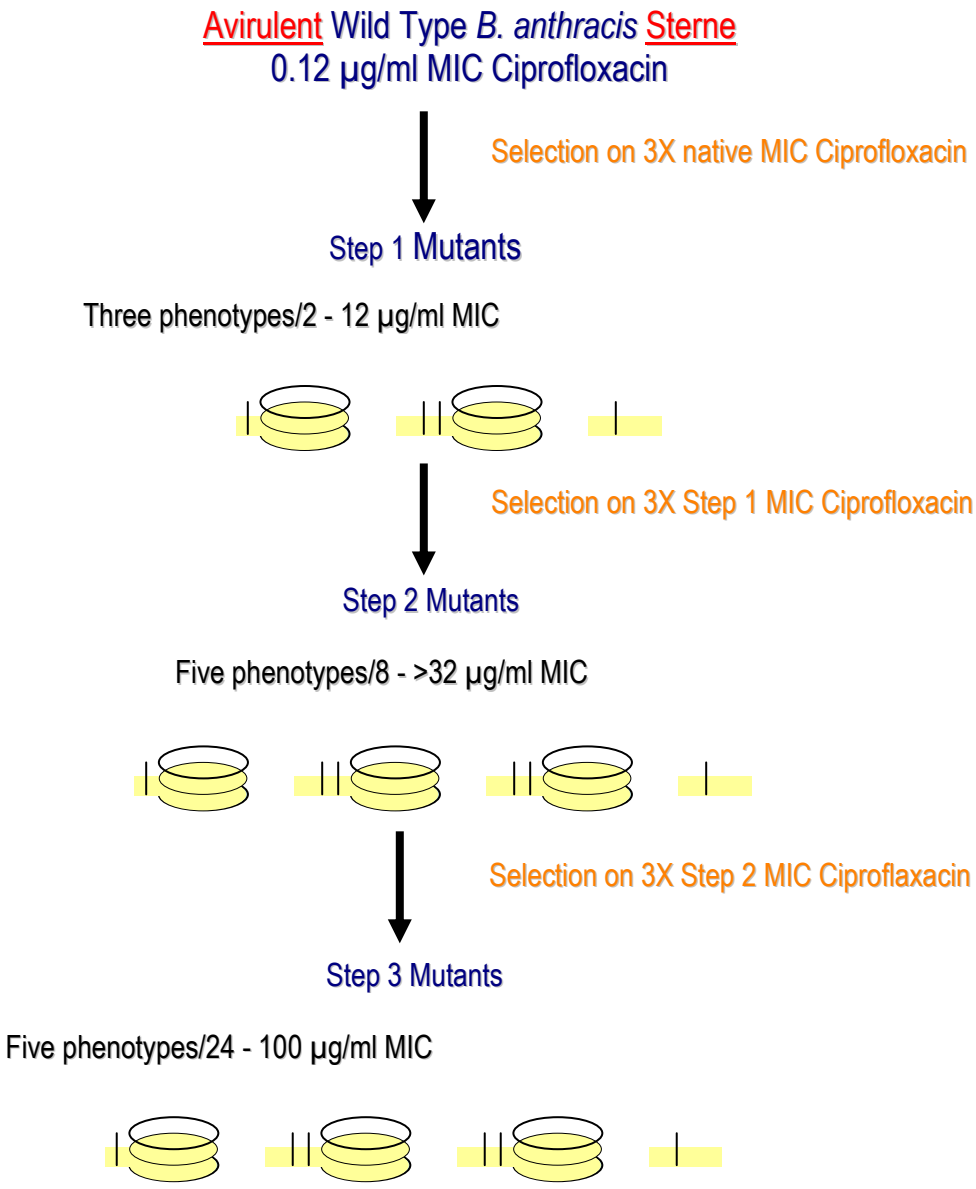
A parental culture of avirulent *Bacillus anthracis* Sterne was streaked onto a nutrient broth agar plate. The wild-type ciprofloxacin minimum inhibitory concentration (MIC) value was determined for *B. anthracis* Sterne by picking a single colony to inoculate 5 mL nutrient broth and incubating overnight 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 was 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.



2. Genomic DNA preparations, 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 using ABI3730 DNA analyzers at the DOE Joint Genome Institute in Walnut Creek, CA.

3. Illumina Sequencing

3.1. Illumina sequencing experimental procedure

Genomic DNA samples from the *Bacillus anthracis* Sterne ciprofloxacin mutants were provided to Eureka Genomics for Illumina sequencing. The samples were sequenced according to standard Illumina sequencing procedures.

3.2 Illumina sequence data generation and quality control

Illumina paired end libraries were prepared from 1 µg of genomic DNA (gDNA) from each of the eleven third round CIP resistant 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 (16), 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.

3.3 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.

4. 454 Sequencing Experimental Procedures.

4.1 454 sequencing procedures

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.

4.2 454 analysis procedures

The 454 sequence reads were provided to Oak Ridge National Laboratory (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.

Mapping the reads from one dataset (one mutant or the parental Dugway isolate of *B. anthracis* Sterne strain) to the *B. anthracis* Sterne genome would take <10 minutes at the highest stringency (100% identity), but would take over an hour at the lowest levels (90% identity). Therefore, once all the scripts were prepared to process the data, all the mappings could be done overnight. This would be followed by a combination of parsing some of the data using scripts and some of it manually.

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. The lowest quality score associated with each variation was obtained to identify detected variations that may be an error due to sequencing quality. 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. 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 to find large regions of the reference genome that had a consensus base of "-" in the mapping. 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. To discern between areas with no consensus base that have a unique depth of 1 and those that do not, the 454AllContigs.fna file, generated by the gsMapper software, was parsed to find regions not mapped.

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.

5. Microarray Experimental Procedures

5.1 Whole genome tiling array design for *B. anthracis* Sterne

Tiling arrays were developed using the NimbleGen 388 K probe format for *B. anthracis* Sterne. We have developed computational tools to design probes that tile across entire bacterial genomes. Probes were isothermal to the extent possible within the allowed length range of 32-40 nucleotides. The T_m ranges were selected to account for the GC% of each organism to center the distribution of probe length around 36 nucleotides. The T_m was calculated using Unafold (http://dinamelt.bioinfo.rpi.edu/download.php), which employs accurate nearest neighbor thermodynamic predictions.

B. anthracis Sterne probes were tiled with an overlap of 55% and T_m 75±3°C across the sequences of
gi|47566322|ref|NC_007322.2| *Bacillus anthracis* str. primeAmes Ancestorprime plasmid pXO1
gi|50163691|ref|NC_007323.3| *Bacillus anthracis* str. primeAmes Ancestorprime plasmid pXO2
gi|10956390|ref|NC_002146.1| *Bacillus anthracis* plasmid pX02
gi|10956247|ref|NC_001496.1| *Bacillus anthracis* plasmid pX01
ref|NC_005945.1|gnl|NCBI_GENOMES|405|gi|49183039|*Bacillus anthracis* str. Sterne
Total # probes: 370,803 with replicate on every 26th position.

5.2 Sample preparation and quantitation.

All genomic 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% amplitude 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).

Cy3-labeled random 9mers (TriLink Biotechnologies, San Diego, CA) were diluted using 125 mM Tris-HCl, 12.5 mM $MgCl_2$, 1.75% (v/v) b-mercaptoethanol at a concentration of 1 OD/42 µL of buffer. Approximately 1 µg sonicated DNA samples in 40 µL were mixed with an equal volume of Cy3-labeled random nonamer primers and the mixture was heat-denatured at 98°C for 10 minutes, then quick chilled on ice for 10 minutes. Ten µL of a solution containing 10 mM dNTPs, 8 µL PCR-grade water and 2 µL of 50 U/µL Klenow DNA polymerase (New England Biolabs, Boston, MA) were added to the labeling reaction in a total volume of 100 µL and the sample was incubated at 37°C for 2.5 hrs. The labeling reaction was stopped by addition of 10 µL 0.5M EDTA. Labeled DNA was precipitated by addition of 11 µL 3 M NaCl and 110 µL isopropanol, collected by centrifugation and the pellet was washed twice in 500 µL of 80% (v/v) ethanol, and then dried. Once ready for hybridization, the pellet was dissolved in 10-20 µL PCR-grade water and the DNA concentration was again determined by A_{260} measurement using

a Nanodrop spectrophotometer. Approximately 20 µg of labeled DNA can be obtained from ~1 µg of starting genomic DNA.

A NimbleGen™ hybridization kit was purchased from Roche NimbleGen (Madison, WI). The kit contains hybridization buffers and alignment oligomers. For each hybridization reaction, 4 µg of labeled DNA in 3.5 µL was mixed with 31.5 µL 2x Hybridization buffer, 8 µL Hybridization component A and 1 µL alignment oligomer (100 nM Cy3 and Cy5-labeled CPK6) in a total volume of 44 µL. The contents were mixed and denatured at 95°C for 5 minutes. The tube was immediately transferred to the MAUI 4-Bay™ Hybridization System (BioMicro Systems, Salt Lake City, UT) at 42°C until ready for sample loading. The MAUI Mixer SL Hybridization Chamber was placed on the array and the sample was loaded. The mixer was sealed and placed in the hybridization stations. Mix mode B was used and the samples were hybridized for 16-17 hrs. NimbleGen™ Wash Buffer kit was purchased from Roche NimbleGen. Each array was washed three times: Wash I, 42°C for 2 min 15 sec, Wash II, room temperature for 1 min and Wash III, room temperature for 15 sec. The arrays were removed from Wash III solution and spin dried in a slide-spinner (Labnet, Edison, NJ) for 1 min.

Arrays were scanned using an Axon 4000B scanner (Molecular Devices, Sunnyvale, CA) at 5 µm resolution. A wavelength of 532 nm was used to scan for Cy3 dyes hybridized on microarray slides. The images of arrays were saved as a .tiff file. NimbleScan software 2.4 was used to extract the array data from .tiff images and convert them into pair file reports. The locations of the alignment oligomers (CPK6 control probes) were used to lay the grid on the array. The pair reports were used for statistical analysis of microarray data.

5.3 Statistical analysis of sequence changes from tiling microarrays.

An algorithm called TAPS (Tiling Array Polymorphism Sensor) to analyze data from two-color hybridizations was developed. The TAPS algorithm is based on a thermodynamic model that predicts the effect of mutations on probe-target hybridization and estimates the likelihood of a mutation at every reference genome position, given the intensities of all probes overlapping the position,. The algorithm superficially resembles the SNPscanner algorithm of Gresham *et al* (3), but requires fewer training parameters (70 vs. 4608), and is thus less susceptible to over fitting; it also works with two-color data, and is not restricted to Affymetrix array designs.

The TAPS algorithm models the effect of a SNP on the intensity of an overlapping probe as a function of several variables: the reference channel probe intensity, the position of the SNP in the probe sequence, the base substitution relative to the reference genome, and the two perfect-match bases on either side of the SNP locus. We assume that probe intensity decreases as the free energy of hybridization increases (becomes less negative), and that the free energy ΔG is a sum of contributions from aligned pairs of nearest-neighbor (NN) nucleotides. A SNP in the target sequence increases the free energy by replacing two perfect-match NN pairs with pairs having a single mismatch. For example, a mutation that changes the sequence AGC to ATC replaces the perfect match pairs AG/TC and GC/CG with the mismatch pairs AT/TC and TC/CG. Since our tiling array only has probes for the reference genome sequence, it does not provide information about the specific base substitution in the target genome. However, the average effect of the 3 possible substitutions can be estimated within a particular base triplet.

The TAPS model also includes a multiplicative position effect, in which SNPs near the middle of a probe cause larger intensity drops than SNPs hitting the ends, especially the 3' region closest to the array surface. We expected to see this based on our earlier work with virulence gene arrays (4). We see that, on average, the intensity drop is almost two-fold when a SNP affects the nucleotides binding near the middle of the probe, but is reduced to zero at either end.

Even in the absence of SNP effects, probe intensities will differ between the two channels due to dye effects, scanner bias and noise. To correct for these effects, each pair of intensities (y_{ref}, y_{mut}) were transformed into the log ratio and log geometric mean:

$$M = \log \frac{y_{mut}}{y_{ref}}$$

$$A = \tfrac{1}{2}(\log y_{ref} + \log y_{mut})$$

A semi-parametric regression model was fit using the M vs. A data for all probes:

$$M = \mu(A) + \varepsilon(A)$$

in which the error term $\varepsilon(A)$ has mean 0 and variance $\sigma^2(A)$, and $\mu(A)$ and $\sigma^2(A)$ are smooth mean and variance functions. The functions $\mu(A)$ and $\sigma^2(A)$ to the M and A values were fit for all probes on each array, using regression on cubic splines to fit $\mu(A)$, and a smoothing spline on binned squared residuals to fit $\sigma^2(A)$. Since SNPs only affect a small fraction of probes on the array, the fitted $\mu(A)$ closely approximates the mean function for perfect match probes (those not overlapping variations between the reference and target strains).

To model the effect of a free energy change $\Delta\Delta G = \Delta G_{mut} - \Delta G_{ref}$ on the log intensity ratio, we assume that the probe oligos within an array feature can be in one of three states: unbound, bound to target DNA from the mutant strain, or bound to target DNA from the reference. At thermodynamic equilibrium at temperature T, the fraction of oligos bound to mutant DNA is given by the Boltzmann equation:

$$\theta_{mut} = \frac{e^{-\Delta G_{mut}/RT}}{1 + e^{-\Delta G_{ref}/RT} + e^{-\Delta G_{mut}/RT}}$$

with a similar equation for the fraction bound to reference DNA θ_{ref}. It follows that

$$\log \frac{\theta_{mut}}{\theta_{ref}} = -\frac{\Delta\Delta G}{RT}$$

Since the probe intensity for each dye (at levels well above background and below saturation) scales with the fraction of oligos bound to target labeled with that dye, we expect the SNP effect on the log intensity ratio to be proportional to $\Delta\Delta G$. That is, for probes overlapping SNPs, our semi parametric regression model was modified to include a term for the SNP effect:

$$M_{obs} = \mu(A_{obs}) + w\,\Delta\Delta G + \varepsilon(A_{obs})$$

where w is a proportionality constant (typically < 0) and the noise term $\varepsilon(A)$ is assumed to be Gaussian with mean 0 and the same variance $\sigma^2(A)$ as was estimated for perfect match probes. The free energy effect $w\,\Delta\Delta G$ is modeled as a product of triplet and position effects:

$$w\,\Delta\Delta G = \beta_\tau\, h(x)$$

where τ indexes the triplet and x is the position of the SNP within the probe, as a fraction of the probe length. The position effect $h(x)$ is approximated by a polynomial function of degree 5:

$$h(x) = \sum_{j=0}^{5} \alpha_j x^j$$

The triplet effects are assumed to be equivalent for reverse complements, so there are 32 β_τ parameters and 6 α_j's. Note that the proportionality constant *w* has been absorbed into the triplet effects. The model parameters were fit to data from experiments in which the target (mutant) strain has known genome sequence and thus has SNPs at known positions relative to the reference genome. To make the parameters identifiable, we scaled the coefficients α_j so that $h(0.5) = 1$.

To apply the model to data from target strains of unknown sequence, a log likelihood ratio test statistic for every position *z* in the reference genome was computed. Let $P(z)$ be the set of probes overlapping position *z*, and let M_i and A_i be the log intensity ratio and average for probe *i*. The semi parametric regression model given above leads to the following expression for the log likelihood:

$$\log L(z) = -\sum_{i \in P(z)} \frac{(M_i - \mu(A_i) - w\Delta\Delta G_i)^2}{2\sigma^2(A_i)}$$

If there is a SNP at position *z*, then $\Delta\Delta G_i$ for each probes was computed using the fitted model parameters; otherwise $\Delta\Delta G_i = 0$ for all probes in $P(z)$. The log likelihood ratio is the difference of the log $L(z)$ values computed under these two assumptions. Candidate SNP positions were identified by looking for regions of the genome where the log likelihood ratio exceeds a fixed threshold. Typically the threshold was set to 20; in tests with targets of known sequence, this threshold provided an optimal balance between false positive and false negative SNP predictions.

RESULTS

1.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 1 contains a compilation of the different resistant isolates and their MIC values.

Table 1. 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	M10:8:1	>32
M10	1.0	M1:6	>32		
M19	1.0	M10:8	12.0		
		M19:2	>32		

2. Evaluation of resistant 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. Table 2 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 2. Total SNPs identified by each technology

B. anthracis **Sterne ciprofloxacin-resistant clones**	**# of SNP's identified by microarray**	**# of SNP's identified by Illumina**	**# of SNP's identified by 454**
1:1	2078	3	14
1:6	42	207	8
10:8:1	86	13	10
19:2	19	4	19

In the following tables (2-5) 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).

Table 3: Mutant 1:1 Sequencing Results Comparison

Gene	Gene Description	Gene Location	SNP Location	SNP Type	Illumina	454
BAS0006	DNA gyrase subunit A	6596-9067	6849	Sub C->T	20 (1.00)	19 (1.00)
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)
BAS3009	penicillin-binding protein, C-terminus	2983668-2984141	2983918	Ins C		10 (1.00)
BAS3391	DNA topoisomerase IV subunit A	3363272-3365695	3365454	Sub G->T	14 (1.00)	10 (1.00)
BAS3658	polyribonucleotide nucleotidyltransferase	3620153-3622291	3621030	Ins AG		10 (1.00)
BAS3720	phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	3687114-3688319	3687446-8	TCT->C		4 (0.40)
BAS5177	modification methylase, HemK family	5056920-5057173	5057173	Ins GAA		18 (1.00)

Table 4: Mutant 1:6 Sequencing Results Comparison

Gene	Gene Description	Gene Location	SNP Location	SNP Type	Illumina	454
BAS0006	DNA gyrase subunit A	6596-9067	6849	Sub C->T	8 (1.00)	11 (1.00)
BAS0425	hypothetical protein	454916-455719	455350	Sub T->C	11 (0.92)	
			455353	Sub T->A	13 (0.93)	
			455365	Sub C->T	9 (1.00)	
BAS0595	sensory box/GGDEF family protein	642479-644182	643376	Ins CCGCG		12 (0.86)
BAS0627	ABC transporter, nucleotide binding domain	677157-678104	677934	Del A		8 (0.50)
BAS0794	transcriptional regulator, TetR family	842297-842875	842614	Ins GCGGGTCTTGC		15 (0.94)
BAS3391	DNA topoisomerase IV subunit A	3363272-3365695	3365454	Sub G->A	8 (0.89)	12 (0.86)
				Sub G->T	1 (0.11)	
BAS5135	ABC transporter ATP-binding protein, N-terminus	5019337-5019846	5019628-30	ACA->C		4 (0.33)
BAS5220	multidrug resistance protein, putative	5109644-5108481	5109171	Sub C->A	15 (0.94)	

Table 5: Mutant 10:8:1 Sequencing and Microarray Results Comparison

Gene	Gene Description	Gene Location	SNP Location	SNP Type	Illumina	454	Micro-array
BAS0006	DNA gyrase subunit A	6596-9067	6849	Sub C->T	22 (1.00)	11 (1.00)	X
tRNA-Leu-3	tRNA-Leu-3	746531-746614	746559	Sub T->C	21 (1.00)		X
			746573	Sub A->G	21 (1.00)		X
BAS0699	thiamine/molybdopterin biosynthesis ThiF/MoeB-like protein	758236-759255	758395	Sub G->A	10 (1.00)		X
			758412	Sub C->T	16 (0.57)		X
			758415	Sub T->A	16 (0.57)		X
			758436	Sub G->A	9 (1.00)		X
BAS0757	hypothetical protein	811432-812592	812171	Sub T->A	19 (1.00)		X
			812197	Sub T->C	22 (1.00)		X
			812296	Sub T->C	32 (1.00)		X
			812299	Sub A->T	36 (1.00)		X
			812300	Sub C->T	36 (0.97)		X
			812338	Sub A->G	12 (1.00)		X
			812339	Sub G->A	12 (1.00)		X
			812455	Sub T->G	15 (0.94)		X
			812476	Sub T->A	46 (1.00)		X
			812485	Sub T->G	39 (0.98)		X
			812524	Sub T->A	10 (1.00)		X
			812535	Sub T->A	21 (1.00)		X
			812560	Sub C->T	23 (1.00)		X
BAS0786	hypothetical protein	837472-837876	837615	Sub T->C	15 (1.00)		X
			837618	Sub C->T	17 (1.00)		X
			837642	Sub T->C	19 (1.00)		X
			837646	Sub A->C	11 (1.00)		X
BAS1107	oligopeptide ABC transporter, oligopeptide-binding protein	1161085-1162689	1162599-600	Del AT		17 (1.00)	
BAS3391	DNA topoisomerase IV subunit A	3363272-3365695	3365454	Sub G->A	33 (1.00)	9 (1.00)	X
BAS4080	hemolysin A	4014548-4015387	4015055	Ins ATC		14 (1.00)	
BAS4814	hypothetical protein	4706252-4707910	4706846	Sub T->C	27 (1.00)	9 (1.00)	
BAS4948	lipoprotein, putative	4829169-4830134	4829181	Del A		8 (0.89)	
BAS5185	stage 0 sporulation protein F	5065837-5066205	5065959-60	Del AA		13 (1.00)	

Table 6: Mutant 19:2 Sequencing Results Comparison

Gene	Gene Description	Gene Location	SNP Location	SNP Type	Illumina	454
BAS0006	DNA gyrase subunit A	6596-9067	6849	Sub C->T	45 (1.00)	12 (1.00)
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	47 (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	44 (1.00)	
			3651013	Del C		6 (1.00)
BAS4278	peptidase, U32 family	4188897-4189826	4188979	Sub T->A		13 (0.93)
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: TAAATATGCCATGAATTATTTAACTGTTATATGAACCAAATAAAAAAAGCATTGCACAAGAGCA ATGCTTTTTTTATATATCCCGATCCAAATAAAGAGGTTA 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.