Supplementary Information for:

A genome sequence based discriminator for vancomycin intermediate *Staphylococcus aureus*

Lavanya Rishishwar, Robert A. Petit, III, Colleen S. Kraft and I. King Jordan

Supplementary Methods

Overview and rationale for the machine learning approach

We chose a supervised machine learning based approach for our efforts to distinguish VISA from VSSA isolates. We chose machine learning because it provides for: 1) maximum potential discriminatory power, 2) the availability of numerous distinct machine learning algorithms that can be evaluated for relatively efficacy, 3) the ability to simultaneously evaluate multiple attributes and 4) the ability to provide information as to which attributes contribute most to the discriminatory power of the algorithm [1]. Supervised machine learning refers to a generic set of algorithmic methods that are used in a two-step process of classification and prediction. In the classification step, input data sets are grouped according to user-specified criteria and a model that can distinguish the data set groups is built. In the prediction step, the model is used to predict the group to which new, *i.e.* previously unseen, data belong.

Details of the machine learning approach used to distinguish VISA and VSSA isolates

The machine learning approach we employed involved the evaluation of 6 different state-of-the-art machine learning algorithms run on 2 meta-parameters and 44 different genomic parameters. Following attribute selection and evaluation of the different machine learning algorithms, the Logistic Regression algorithm was used with the 14 genomic parameters highlighted (*) in Table 3. The scheme of the final machine learning approach used is shown in Supplementary Figure 1. The 4 steps that were used in the machine learning approach are detailed below.

1. Parameterization of S. aureus pairwise genome comparisons

Note that we are using the terms ‘parameter’ and ‘attribute’ synonymously here as is typically done in the machine learning field. A parameter or an attribute is an individual characteristic or feature that is being compared among genomes in order to characterize them.

Meta-parameters: Two meta-parameters were established in order to classify all other genomic parameters that are subsequently compared. These meta-parameters are the class to which the genome

under consideration originally belongs (VISA or VSSA) and the class of genome it is being compared to. The formulation of the two meta-parameters in this way allows all subsequent parameters to be represented as a set of pairwise distances between genomes.

Genome assembly-based parameters: For each complete *S. aureus* genome sequence under consideration, a set of sequence reads was simulated using the 454sim tool [2]. The simulated reads for each genome were used in reference based assembly against all genomes under consideration, and the following assembly statistics from each individual pairwise reference assembly was recorded: number of large contigs (>500bp), number of assembled bases, the N50 value, the percent of aligned reads, an assembly score computed as $\log_{10}\left[\frac{N50 \times number\ assembled\ bases}{number\ large\ contigs}\right]$. *De novo* assemblies of each simulated genome read set were performed, and the pairwise genome-wide average nucleotide identities (ANI) amongst all *de novo* assemblies were computed. This set of 5 pairwise reference assembly statistics, along with the pairwise ANI values, were taken as the genome assembly-based parameters that were used in the model building step of the algorithm.

Gene-based parameters: Pairwise inter-gene percent identities were computed for each individual vancomycin-intermediate susceptibility implicated gene (Figure 1) as well as for 16S rRNA and the 7 MLST genes (Table 3). This set of 38 gene-specific pairwise sequence identities was taken as the gene-based parameters that were used in the model building step of the algorithm.

2. Attribute (parameter) selection

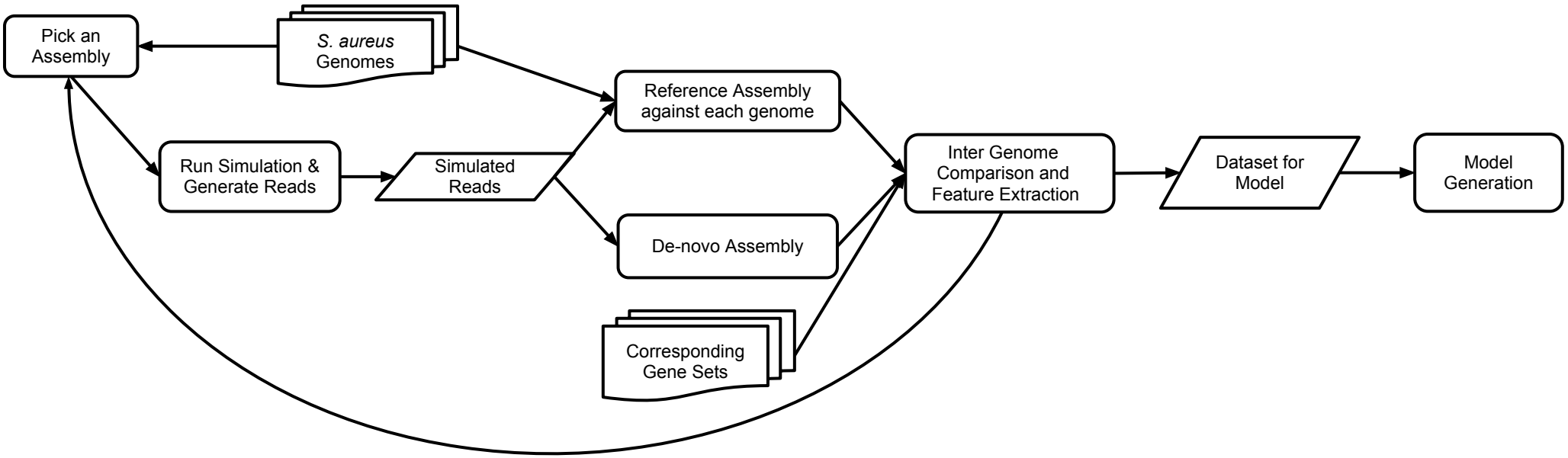
Attribute selection is used in machine learning in order to converge on the minimal set of maximally informative attributes or parameters. These are the parameters that provide the most information with respect to the delineation of the user specified classes, in our case VISA versus VSSA. The attribute selection algorithm implemented in the WEKA collection of machine learning algorithms [3] was used for this purpose resulting in the reduction of the total number of attributes used in classification from 44 to 14.

3. Model generation for classification

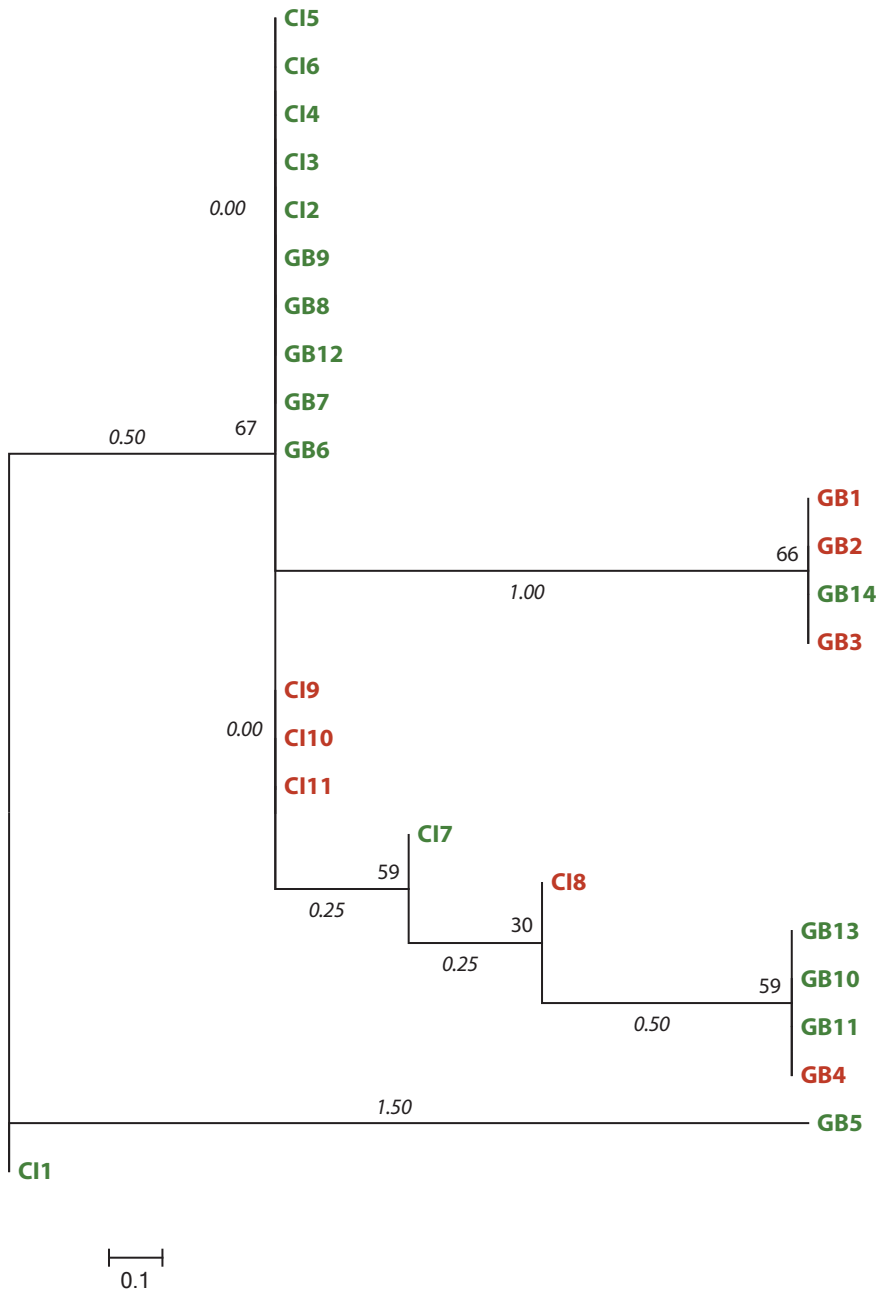
The following 6 different machine learning algorithms, implemented in the WEKA collection, were used to build VISA and VSSA discriminatory model based on the reduced set of 14 attributes: J48, Logistic Regression, Multilayer Perceptron (aka Artificial Neural Networks), Naïve Bayes, RandomForest and Support Vector Machines. The relative performance of each of these algorithms for the discrimination of VISA and VSSA strains was evaluated using the following 3 metrics: accuracy, precision and recall (Supplementary Table 1). All of the algorithms performed well with Logistic Regression and Multilayer Perceptron showing the best results. Logistic Regression was chosen for the final implementation of the machine learning approach given its simplicity and transparency with respect to the formalization of the model.

4. Model testing for prediction

Once the Logistic Regression model was generated using all genomes as described above, the prediction phase of the machine learning approach was performed using cross-validation with K=25, *i.e.* leave-one-out validation. This means that each individual strain, both VISA and VSSA, was removed from the total set of genomes and then tested against a model built using the remaining 24 strains' genomes. In this way, all VISA and VSSA strains are used in testing. For each iteration of the leave-one-out validation procedure, the accuracy of the class assignment for the left out strain was recorded. The overall accuracy was recorded as the fraction of individual strains correctly assigned as VISA or VSSA.



Supplementary Figure 1. **Machine learning scheme used for the genome based discrimination of VISA and VSSA isolates.**



Supplementary Figure 2. **Evolutionary relationships among VISA (red) and VSSA (green) isolates based on 16S rRNA.** Numbers of nucleotide differences between clades are shown along the branches and bootstrap support values are shown for interior nodes.

Supplementary Table 1: Machine Learning Algorithms evaluated on the genome data. The table presents different evaluation metrics calculated by applying a number of machine learning algorithms on a 10-fold cross validation basis. It should be noted here that the performance of the overall system will always be higher as the system is designed to pick the best genome pair match *i.e.*, the one that shows the greatest degree of confidence whereas the metrics shown below describes the efficiency of the system in classifying all the possible genome pair.

Machine Learning Algorithm	Accuracy (%)[a]	Precision (%)[b]	Recall (%)[c]
J48	66.19	66.20	66.20
Random Forest	68.78	68.80	68.80
Naïve Bayes	67.94	77.60	67.90
SVM	73.17	73.20	73.20
Logistic	73.65	73.70	73.70
Multilayer Perceptron	71.90	71.90	71.90

[a] $Accuracy = \frac{number\ of\ true\ positives + number\ of\ true\ negatives}{number\ of\ positives + number\ of\ negatives}$

[b] $Precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ positives}$

[c] $Recall = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$

Supplementary Table 2. **Protein cluster assignment of the genes analyzed.**

Gene	Protein Cluster ID	Protein Function	Protein Cluster Function	Class
agrA	PCLA_885733	Histidine kinase	Signal transduction mechanisms; Transcription	Cellular processes and signaling
agrC	PCLA_885732	Histidine kinase	Signal transduction mechanisms	Cellular processes and signaling
blaZ	PCLA_928930	Beta-lactamase	Defense mechanisms	Cellular processes and signaling
graS	PCLA_884800	Sensor histidine kinase	Signal transduction mechanisms	Cellular processes and signaling
pbp4	PCLA_884782	D-alanyl-D-alanine carboxypeptidase	Cell wall/membrane biogenesis	Cellular processes and signaling
pbpB	PCLA_885360	Transglycosylase	Cell wall/membrane biogenesis	Cellular processes and signaling
rsbU	PCLA_3268752	Serine phosphatase	Signal transduction mechanisms; Transcription	Cellular processes and signaling
spoVG	PCLA_3407896	Stage V sporulation protein G	Cell wall/membrane biogenesis	Cellular processes and signaling
tcaB	PCLA_885979	Bicyclomycin transporter TcaB	Defense mechanisms	Cellular processes and signaling
vraF	PCLA_916605	Bacitracin ABC transporter ATP-binding protein	Defense mechanisms	Cellular processes and signaling
vraG	PCLA_3341615	Bacitracin ABC transporter permease	Defense mechanisms	Cellular processes and signaling
vraS	PCLA_885626	Sensor histidine kinase	Signal transduction mechanisms	Cellular processes and signaling
ccpA	PCLA_883369	Catabolite control protein A	Transcription	Information storage and processing
graR	PCLA_884799	Response regulator GraR	Signal transduction mechanisms; Transcription	Information storage and processing
rpoB	PCLA_2821305	DNA-directed RNA polymerase subunit beta	Transcription	Information storage and processing
rpoD	PCLA_3392123	DNA polymerase	Transcription	Information storage and processing
tcaR	PCLA_885922	MarR family transcriptional regulator	Transcription	Information storage and processing
tgt	PCLA_414015	Queuine tRNA-ribosyltransferase	Translation	Information storage and processing
walK	PCLA_888192	Sensor histidine kinase	Signal transduction mechanisms	Information storage and processing
walR	PCLA_873777	PhoP family transcriptional regulator	Signal transduction mechanisms; Transcription	Information storage and processing
arcC	PCLA_4954367	Carbamate kinase	Amino acid transport and metabolism; General function prediction only; Nucleotide transport and metabolism	Metabolism
aroE	PCLA_3373691	Shikimate 5-dehydrogenase	-	Metabolism

folC	PCLA_885487	Folylpolyglutamate synthase	Coenzyme transport and metabolism	Metabolism
glfP	PCLA_885258	Glycerol transporter	Carbohydrate transport and metabolism	Metabolism
gmk_	PCLA_4868944	Guanylate kinase	Nucleotide transport and metabolism	Metabolism
isdE	PCLA_873576	Heme ABC transporter substrate-binding protein	-	Metabolism
prsA	PCLA_885597	Peptidyl-prolyl cis-trans isomerase	-	Metabolism
pta_	PCLA_429196	Phosphotransacetylase	Energy production and conversion	Metabolism
tpi_	PCLA_413954	Triosephosphate isomerase	Amino acid transport and metabolism; Carbohydrate transport and metabolism; Coenzyme transport and metabolism; General function prediction only; Nucleotide transport and metabolism	Metabolism
yqil	PCLA_209635	Acetyl-CoA acetyltransferase	-	Metabolism
SA1129	PCLA_624509	Ribonuclease	Function unknown; General function prediction only	Poorly characterized
SA1703	PCLA_885627	Transporter	Function unknown	Poorly characterized
SBF	PCLA_894844	Sodium transporter	General function prediction only	Poorly characterized
sigB	PCLA_3619763	RNA polymerase sigma factor SigB	General function prediction only	Poorly characterized
stp1	PCLA_885212	Serine/threonine-protein kinase	Function unknown	Poorly characterized
tcaA	PCLA_885921	Membrane protein	Function unknown	Poorly characterized

References

1. Han J, Kamber M. Data mining : concepts and techniques. 2nd ed. San Francisco, Calif.
Oxford: Morgan Kaufmann ;
Elsevier Science distributor, **2006** The Morgan Kaufmann series in data management systems).

2. Lysholm F, Andersson B, Persson B. An efficient simulator of 454 data using configurable statistical models. BMC research notes **2011**; 4:449.

3. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics **2004**; 20:2479-81.