MICROBIAL GENOMICS

RESEARCH ARTICLE

Zhuang *et al.*, *Microbial Genomics* 2021;7:000610

DOI 10.1099/mgen.0.000610

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

A random forest model based on core genome allelic profiles of MRSA for penicillin plus potassium clavulanate susceptibility prediction

Hemu Zhuang, Feiteng Zhu, Peng Lan, Shujuan Ji, Lu Sun, Yiyi Chen, Zhengan Wang, Shengnan Jiang, Linyue Zhang, Yiwei Zhu, Yan Jiang, Yan Chen* and Yunsong Yu*

Abstract

Treatment failure of methicillin-resistant *Staphylococcus aureus* (MRSA) infections remains problematic in clinical practice because therapeutic options are limited. Penicillin plus potassium clavulanate combination (PENC) was shown to have potential for treating some MRSA infections. We investigated the susceptibility of MRSA isolates and constructed a drug susceptibility prediction model for the phenotype of the PENC. We determined the minimum inhibitory concentration of PENC for MRSA (n=284) in a teaching hospital (SRRSH-MRSA). PENC susceptibility genotypes were analysed using a published genotyping scheme based on the *mecA* sequence. *mecA* expression in MRSA isolates was analysed by qPCR. We established a random forest model for predicting PENC-susceptible phenotypes using core genome allelic profiles from cgMLST analysis. We identified S2-R isolates with susceptible *mecA* genotypes but PENC-resistant phenotypes; these isolates expressed *mecA* at higher levels than did S2 MRSA (2.61 vs 0.98, P<0.05), indicating the limitation of using a single factor for predicting drug susceptibility. Using the data of selected UK-sourced MRSA (n=74) and MRSA collected in a previous national survey (NA-MRSA, n=471) as a training set, we built a model with accuracies of 0.94 and 0.93 for SRRSH-MRSA and UK-sourced MRSA (n=287, NAM-MRSA) validation sets. The AUROC of this model for SRRSH-MRSA and NAM-MRSA was 0.96 and 0.97. Although the source of the training set data affects the scope of application of the prediction model, our data demonstrated the power of the machine learning approach in predicting susceptibility from cgMLST results.

DATA SUMMARY

The sequence of SRRSH-MRSA had been deposited in National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) under project number PRJNA733242. The sequence of NA-MRSA had been deposited in The National Genomics Data Center (NGDC) (https://bigd.big.ac.cn/) of China under project number PRJCA004763. The sequence of NAM-MRSA were downloaded from European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) by using the accession numbers provided in the Harrison et al. 2019 (1).

INTRODUCTION

Staphylococcus aureus is a common pathogen that causes a wide range of clinical infections, from superficial soft tissue infections to deep infective endocarditis, which seriously endanger public health [1]. The resistance of *S. aureus* to

Received 23 November 2020; Accepted 15 May 2021; Published 23 September 2021

Author affiliations: [1]Department of Infectious Diseases, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine; Key Laboratory of Microbial Technology and Bioinformatics of Zhejiang Province; Regional Medical Center for National Institute of Respiratory Diseases, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, PR China.

***Correspondence:** Yan Chen, chenyan@zju.edu.cn; Yunsong Yu, yvys119@zju.edu.cn

Keywords: CC59; machine learning; methicillin-resistant *Staphylococcus aureus*; *mecA*; penicillin; random forest.

Abbreviations: AST, antibiotic susceptibility test; AUC, area under receiver operating curve; CC, cloning complex; cgMLST, core genome multilocus sequence typing; CLSI, Clinical and Laboratory Standards Institute; ENA, European Nucleotide Archive; ISA, Iso-Sensitest agar; MIC, minimum inhibitory concentrations; MRSA, methicillin-resistant *Staphylococcus aureus*; NAM-MRSA, MRSA isolates collected in the UK; NA-MRSA, MRSA isolates collected nationwide in China; NCBI, The National Center for Biotechnology Information; NGDC, The National Genomics Data Center; OD, optical density; PBP2a, penicillin-binding protein2a; PENC, penicillin plus potassium clavulanate combination; qPCR, real-time quantitative PCR detecting system; RNA, ribonucleic acid; ROC, receiver operating curve; SRRSH-MRSA, MRSA isolates collected in SIR RUN RUN SHAW Hospital; ST, sequence type.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary tables and three supplementary figures are available with the online version of this article.

most β-lactams is mediated by *mecA*-encoded PBP2a, which has a low binding affinity for β-lactam drugs. Compared to methicillin-susceptible *S. aureus*, methicillin-resistant *S. aureus* (MRSA) is associated with a more significant disease burden and higher mortality due to *S. aureus* bacteraemia [2]. Although the incidence of hospital-related MRSA infections has recently declined, community-related MRSA infections have become more prevalent [3].

Vancomycin, daptomycin, and linezolid are the most commonly prescribed drugs to treat MRSA infections [4]. However, vancomycin has poor tissue permeability and causes slow bacterial clearance, and has been associated with a high risk of failure in the treatment of MRSA infections. Previous studies also showed that daptomycin and linezolid have many limitations in treating MRSA infection, including the emergence of resistance and potential side effects [5]. Recently, different approaches to restoring the susceptibility to β-lactams have been investigated in MRSA, including combined administration of β-lactamase inhibitors such as clavulanic acid. Harrison *et al*. demonstrated that a considerable proportion of MRSA of different lineages were susceptible to penicillin plus 15 mg l^{-1} potassium clavulanate (PENC) *in vitro* and *in vivo*, which was mediated by mutations in *mecA* genes and their promoters [6]. However, this study did not include the data of MRSA lineages from China.

In addition to identifying more effective treatment options, rapid drug susceptibility analysis of MRSA is also needed. Machine learning has shown great potential for diagnostics in clinical microbiology laboratories, such as in predicting the susceptibility of clinical bacterial isolates [7–11]. The random forest is a widely used method for ensemble learning. Several previous studies built machine learning models for predicting AMR phenotypes by using assembled genomes or pan genomes as training sets [8-10], indicating that a susceptibility prediction model based on whole-genome sequencing data and machine learning has value as a diagnostic tool. As the mechanism promoting β-lactam susceptibility in MRSA remains unclear, the machine learning model may also be useful for discovering new resistance mechanisms [12].

In the present study, we investigated the susceptibility to PENC in different MRSA lineages from China. We also created a random forest model using the core genome allelic profiles of MRSA isolates to predict drug susceptibility and evaluated other key genetic determinants that may affect MRSA susceptibility to PENC.

METHODS

MRSA isolates and genome sequencing

We collected 292 MRSA (SRRSH MRSA) isolates at a teaching hospital in 2013–2015, and the clinical data and genome sequences of these isolates were obtained from a previous study [13]. Eight isolates were excluded because the complete *mecA* sequences were not captured by second-generation sequencing.

Impact Statement

With the development of genome sequencing technology, access to whole-genome information has become increasingly rapid and at a low cost, and the prediction of microbial phenotypes based on genomics has become feasible and efficient. In this study, we investigated the susceptibility of MRSA isolates to penicillin combined with potassium clavulanate and constructed a powerful drug susceptibility prediction model based on the core genome information of MRSA using the random forest algorithm. Our research showed the strong potential of genomics-based drug sensitivity prediction, which deserves the attention of more clinical microbiologists and computer scientists.

Antimicrobial susceptibility testing

The minimum inhibitory concentrations (MIC) of penicillin (0.03125–64 mg l^{-1}) and amoxicillin (0.03125–64 mg l^{-1}) were determined with reference to the Clinical and Laboratory Standards Institute guidelines [14] and Harrison *et al*. [6] using agar (ISO-sensitest agar) dilution for all MRSA in this study. The ECOFF of susceptibility to penicillin was 2 mg l^{-1} [6], and the breakpoint of amoxicillin was 8 mg l^{-1} [15] in the presence of 15 mg l^{-1} potassium clavulanate.

Verifying the relationship between *mecA* genotype and PENC susceptibility of MRSA

Based on previous study, the genotype of *mecA* based on two mutation sites in the gene (E246G, M122I) and two in its promoter ([−33]:C-T, [−7]:G-T) was classified as resistant and susceptible. In addition to the six previously reported *mecA* genotypes (S1([−7]:G–T), S2([−7]:G–T|E246G), S3([−33]:C–T[−7]:G|E246G),S4([−7]:G|M122I), R1([−7]:G), and R2([−7]:G|E246G)) [6], we identified numerous new genotypes with different mutation profiles. The numbering of the genotypes in this study represents the order in which they were typed. The drug susceptibility distribution of PENC in different *mecA* genotypes of MRSA was investigated.

RNA isolation and reverse transcription-polymerase chain reaction (RT-qPCR)

MRSA isolates were incubated in tryptone soya broth overnight, and then 50 µl of the cultures were incubated in 5 ml tryptone soya broth at 37 °C until the OD$_{600}$ reached 0.3. The cultures were then incubated with 10 mg l^{-1} oxacillin for 1 h at 37 °C to induce *mecA* expression. Total RNA was extracted from 3 ml of treated MRSA cultures using the E.Z.N.A. Bacterial RNA Kit (Omega Bio-Tek, Norcross, GA, USA). cDNA was synthesised using an Evo M-MLVRT Premix for qPCR Kit (Accurate Biotechnology Co., Ltd., Hunan, China). RT-qPCR was conducted using Premix Pro TaqHS qPCR Kits (Accurate Biotechnology Co., Ltd.) using *mecA* primers (q*mecA*-F:CTCAGGTACTGCTATCCACC; q*mecA*-R: GGAACTTGTTGAGCAGAGG) [6] and a

LightCycler 480 II (Roche Holdings AG, Basel, Switzerland). The fold-changes in gene expression in the cultures were calculated relative to SA268 (an ST59 MRSA isolate with the *mecA* S2 genotype) using the $2^{-\Delta\Delta Ct}$ method, with *gyrB* (qgyrB-F: ATAATTAT GGTGCTGGGCAAAT; qgyrB-R: AACCAGCTAATGCT CATCGATA)as the reference [6].

Core genome multilocus sequence typing (cgMLST) analysis

We loaded FASTA files of genome assemblies into the Ridom SeqSphere+ software (version 5.0) for cgMLST analysis [16]. The cgMLST analysis result was a cgMLST scheme containing the core genome allelic profile of all the input genomes. A cgMLST scheme was developed for *S. aureus*, in which 1861 target genes were named with allelic nomenclature and with COL as the seed genome. The cgMLST scheme was used for further training and verification of the algorithm model.

Random forest algorithm trained for prediction

To ensure the random forest predictive model has a larger scope of application, it is better to use the training set that includes various MRSA isolates with different genetic backgrounds. The MRSA isolates collected during a 2.5 year national epidemic survey (NA-MRSA, *n*=471), (sheet Table S1) comprising 240 PENC-susceptible and 231 PENC-resistant MRSA with ten different clonal complexes, served as a training set to establish a random forest predictive model which named random forest model 1 in this article. The whole genomes of these MRSA samples have been sequenced, and their PENC susceptibility was tested in our previous study (unpublished data). FASTA files of the genome assemblies were loaded into the Ridom SeqSphere+ software (version 5.0) for cgMLST analysis [16]. The results of cgMLST and PENC AST results of NA-MRSA were used for further training of the algorithm. The random forest method proceeded using the R package randomForest (R version 4.0.2 for windows; http://www.r-project.org/). The accuracy, 95% confidence interval (CI), sensitivity, specificity, and receiver operating characteristics (ROC) curve of the random forest prediction model were exported. In addition, the top ten genes in the MRSA core genome that were most closely associated with the susceptibility of MRSA to PENC were determined (see Supplementary data for full details).

To make the random forest model 2, we downloaded the data of 298 MRSA isolates reported in the United Kingdom [6]. Among these 298 isolates, we eliminated 11 isolates whose cloning complex results did not match the results reported before [6]. From these 287 isolates (NAM-MRSA (sheet Table S2), we randomly selected up to five PENC-resistant and five PENC-susceptible isolates in different clonal complexes using the Random extraction tool in excel. A total of 16 PENC-resistant MRSA isolates and 58 PENC-susceptible MRSA isolates were selected and listed in (sheet Table S3), available in the online version of this article). Then we added them to the training set to retrain the model, which was assigned as random forest model 2 in this article. (see Supplementary data for full details).

The SRRSH-MRSA (sheet Tabale S4) and NAM-MRSA served as a validation set to verify the reliability of the random forest model. (see Supplementary data for full details).

Statistical methods

Continuous variables were compared using Student's *t*-tests. Data were statistically analysed using R software (version: 4.0.2; The R Project for Statistical Computing; www.r-project.org). Differences were considered as statistically significant when $P<0\ .05$.

RESULTS

Susceptibility of SRRSH-MRSA isolates to PENC

To explore the feasibility of using PENC to treat MRSA, the PENC susceptibility of MRSA isolates from SRRSH hospital was investigated. In this study, we designated MRSA isolates with PENC MIC<=2 mg l^{-1} as PENC susceptible isolates. In total, 106 of 284 (37%) MRSA isolates were susceptible to PENC with MICs ranging from <0.03125 to 32 mg l^{-1} (Fig. 1a, Table 1). Because amoxicillin/potassium clavulanate is frequently applied in the clinical setting, we also assessed amoxicillin susceptibility in the presence of 15 mg l^{-1} potassium clavulanate. The amoxicillin MIC distribution tended to be similar to that of penicillin (Fig. 1b, c). In the presence of 15 mg l^{-1} of potassium clavulanate, 114 of 284 (40%) MRSA isolates were inhibited when the concentration of amoxicillin reached 8 mg l^{-1} (Fig. 1b, Table 1).

The present study included 284 isolates comprising 155 CC5, 73 CC59, 16 CC239, ten CC8, eight CC88, and three CC1 isolates as well as 19 other clone complexes. To determine the relationship between the antibacterial activity of penicillin and MRSA genetic background, we analysed the MIC distribution of penicillin and amoxicillin among different lineages (Fig. 1a, b). We found that 85% (62/73) of CC59 isolates were susceptible to PENC, whereas the susceptible rate was only 5% (7/155) in CC5 isolates. Compared with CC5 isolates (MIC_{50}=32 mg l^{-1}; MIC_{90}=32 mg l^{-1}), CC59 isolates (MIC_{50}=0.25 mg l^{-1}; MIC_{90}=4 mg l^{-1}) were more susceptible to PENC. CC239 was one of the most important hospital-acquired MRSA (HA-MRSA) lineages in China. Here, all 16 CC239 isolates were PENC-resistant with an MIC_{50} and MIC_{90} of 32 mg l^{-1}. Three CC1 isolates, ten CC8 isolates, and all eight CC88 isolates were susceptible to PENC (Fig. 1a, Table 1). These results indicate a higher susceptibility to PENC in the community-associated MRSA (CA-MRSA) lineage. With respect to amoxicillin, 69 of 73 (95%) CC59 MRSA isolates (MIC_{50}=1 mg l^{-1}; MIC_{90}=8 mg l^{-1}) were inhibited by 8 mg l^{-1} amoxicillin in the presence of 15 mg l^{-1} potassium clavulanate, whereas only 7 of 155 (5%) CC5 isolates (MIC_{50}=32 mg l^{-1}; MIC_{90}=64 mg l^{-1}) were inhibited under these conditions (Fig. 1b, Table 1).

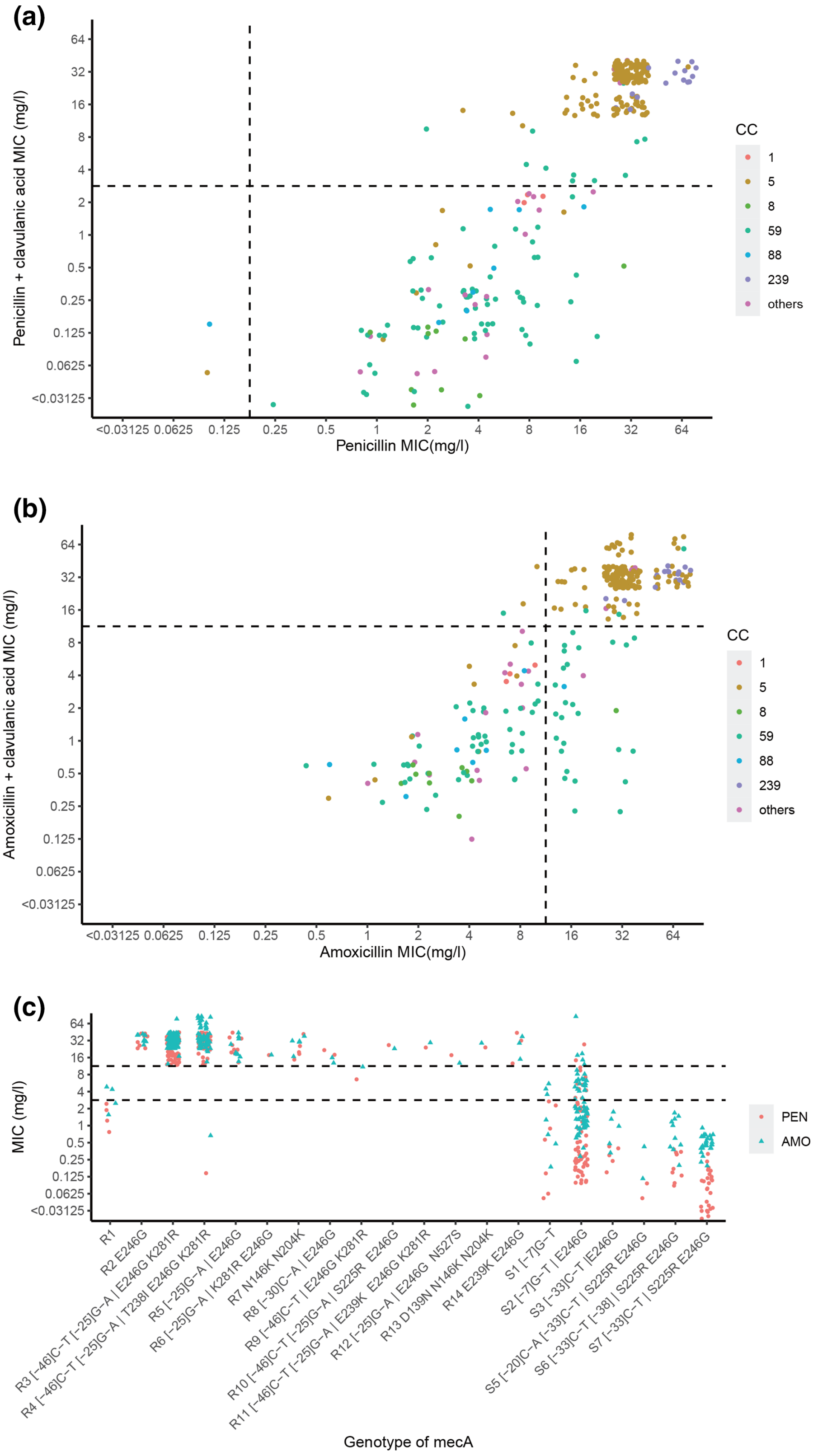


Fig. 1. Distribution of MIC (mg l^{-1}) of SRRSH-MRSA isolates (a) MIC of Penicillin and penicillin plus 15 mg l^{-1} clavulanic acid in the main cloning complex. (b) MIC of Amoxicillin and amoxicillin plus 15 mg l^{-1} clavulanic acid in the main cloning complex. (c) MIC distribution of different MRSA *mecA* genotypes with 15 mg l^{-1} clavulanic acid.

Table 1. Distribution of *mecA* mutations and its promoter in different lineages

		Promoter of *mecA*							PBP2a									Genotype	Phenotype		no. of isolates
		−46	−38	−33	−30	−25	−20	-7	Allosteric domain								TP domain				
									139	146	204	225	238	239	246	281	527		Susceptible	Resistant	
	COL	C	A	C	C	G	C	G	D	N	N	S	T	E	E	K	N	R1	–	–	–
CC1	ST1	C	A	C	C	G	C	G	D	N	N	S	T	E	E	K	N	R1	2	0	2
		C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	0	1
CC5	ST5	T	A	C	C	A	C	G	D	N	N	S	T	E	G	R	N	R3	0	80	80
		T	A	C	C	A	C	G	D	N	N	S	I	E	G	R	N	R4	1	50	51
		C	A	C	C	A	C	G	D	N	N	S	T	E	G	K	N	R5	0	8	8
		C	A	C	A	G	C	G	D	N	N	S	T	E	G	K	N	R8	0	2	2
		C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	1	2
		C	A	C	C	A	C	G	D	N	N	S	T	E	G	R	N	R6	0	1	1
		T	A	C	C	G	C	G	D	N	N	S	T	E	G	R	N	R9	0	1	1
		C	A	C	C	G	C	G	D	N	N	S	T	E	G	K	N	R2	0	1	1
		T	A	C	C	A	C	G	D	N	N	R	T	E	G	K	N	R10	0	1	1
		T	A	C	C	A	C	G	D	N	N	S	T	K	G	R	N	R11	0	1	1
		C	A	C	C	A	C	G	D	N	N	S	T	E	G	K	S	R12	0	1	1
	ST965	C	A	C	C	G	C	T	D	N	N	S	T	E	E	K	N	S1	5	0	5
	ST3194	T	A	C	C	A	C	G	D	N	N	S	T	E	G	R	N	R3	0	1	1
CC59	ST59	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2*	37	11	48
		C	G	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S6	10	0	10
		C	A	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S7	8	0	8
		C	A	C	C	G	C	G	D	N	N	S	T	E	E	K	N	R1	2	0	2
	ST338	C	A	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S7	4	0	4
	ST3195	C	A	C	C	G	C	T	D	N	N	S	T	E	E	K	N	S1	1	0	1
CC239	ST239	C	A	C	C	G	C	G	D	N	N	S	T	E	G	K	N	R2	0	7	7
		C	A	C	C	G	C	G	D	K	K	S	T	E	E	K	N	R7	0	5	5
		C	A	C	C	G	C	G	D	N	N	S	T	K	G	K	N	R14	0	3	3
		C	A	C	C	G	C	G	N	K	K	S	T	E	E	K	N	R13	0	1	1
CC8	ST630	C	A	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S7	9	0	9
		C	A	T	C	G	C	G	D	N	N	S	T	E	G	K	N	S3	1	0	1

Continued

Table 1. Continued

		Promoter of *mecA*							PBP2a									Genotype	Phenotype		no. of isolates
		−46	−38	−33	−30	−25	−20	-7	Allosteric domain								TP domain				
									139	146	204	225	238	239	246	281	527		Susceptible	Resistant	
CC88	ST88	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	4	0	4
		C	A	T	C	G	C	G	D	N	N	S	T	E	G	K	N	S3	4	0	4
others	ST22	C	A	T	C	G	A	G	D	N	N	R	T	E	G	K	N	S5	2	0	2
	ST398	C	A	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S7	3	0	3
	ST772	C	A	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S7	1	0	1
	ST950	C	A	C	C	G	C	T	D	N	N	S	T	E	E	K	N	S1	1	0	1
	ST1661	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	2	0	2
	ST1611	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	0	1
	ST6190	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	0	1
	ST4513	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	0	1
	ST6174	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	0	1
	ST6175	C	A	C	C	G	C	T	D	N	N	S	T	E	G	K	N	S2	1	0	1
	ST6192	C	A	T	C	G	C	G	D	N	N	R	T	E	G	K	N	S7	1	0	1
	ST6191	T	A	C	C	A	C	G	D	N	N	S	T	E	G	R	N	R3	0	1	1
	ST5530	T	A	C	C	A	C	G	D	N	N	S	I	E	G	R	N	R4	0	1	1
	ST4988	C	A	C	C	A	C	G	D	N	N	S	T	E	G	K	N	R5	0	1	1

*In this article isolates whose *mecA* genotype was S2 but phenotype was PENC-resistant was named as S2-R.

Table 2. Performance of *mecA* genotyping and random forest model to predict susceptibility of MRSA to PENC

Method	Sensitivity, %	Specificity, %	Accuracy (95% CI)
mecA	95.3	93.3	0.940 (0.904–0.963)
Random forest model 1*			
In training set	100	99.1	0.996 (0.985–1.00)
In validating set (SRRSH-MRSA)	88.9	93.8	0.919 (0.881–0.948)
In validating set (NAM-MRSA)	70.2	2.4	0.509 (0.449–0.568)
Random forest model 2 (retrained)*			
In training set	100	99.1	0.996 (0.987–1.00)
In validating set (SRRSH-MRSA)	95.4	93.8	0.944 (0.910–0.968)
In validating set (NAM-MRSA)	94.6	90.2	0.934 (0.899–0.960)

*Random forest model 1: the model trained by using NA-MRSA; Random forest model 2 (retrained): the model trained by using NA-MRSA and selected UK-sourced MRSA.

Relationship between *mecA* genotype and PENC susceptibility of MRSA

To investigate the relationship between the *mecA* genotype and PENC susceptibility, we analysed the second-generation sequencing data of SRRSH-MRSA based on the previously reported genotype scheme (Table 1). The results showed that *mecA* polymorphisms were detected in SRRSH-MRSA isolates, with 14 resistant and six susceptible types. The most prevalent genotypes among PENC-resistant MRSA was R3 ([-46]C-T [-25] G-A | E246G K281R), which was mainly distributed in the CC5 MRSA isolates. The most prevalent genotypes among PENC-susceptible MRSA was S2 ([-7]G-T | E246G), which was mainly distributed in the CC59 MRSA isolates. We did not detect the known S4 genotype ([-7]: G | M122I) among our isolates. As observed for the phenotype distribution, the *mecA* genotype is associated with the clonal complex of isolates. We found that 147 of 155 (95%) CC5 isolates had resistant genotypes and 71 of 73 (97%) CC59 had susceptible genotypes. Two of three CC1 and all (100%) CC239 isolates had resistant genotypes. All ten CC8 and eight CC88 isolates had susceptible genotypes (Table 1).

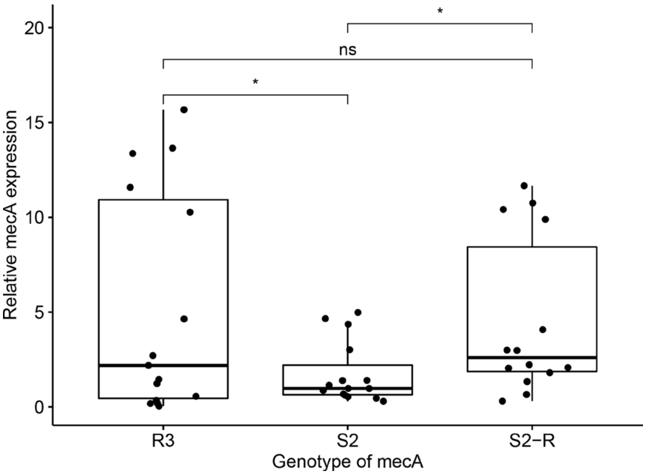


Fig. 2. Relative *mecA* expression measured by RT-qPCR after oxacillin induction, relative to that of *mecA* of SA268. R3, R3 *mecA* genotype of ST5 isolates; S2, S2 *mecA* genotype of ST59 MRSA isolates with matching phenotype and genotype; S2-R, S2 *mecA* genotype of ST59 MRSA isolates with mismatched phenotype and genotype. *P<0.05 (two-tailed unpaired *t*-tests).

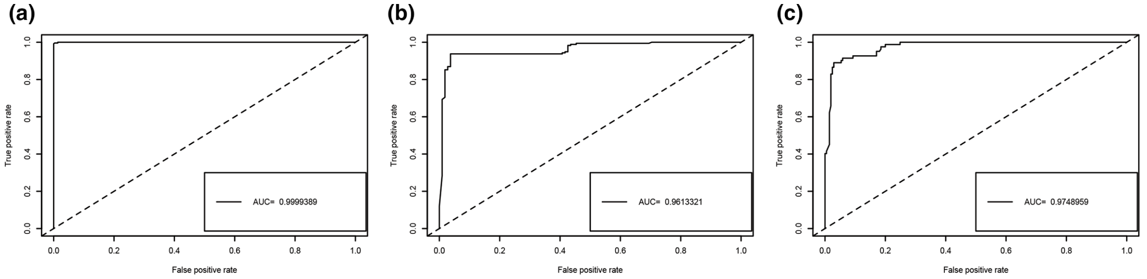
We investigated the MIC distribution of penicillin and amoxicillin in relation to the *mecA* genotypes of SRRSH-MRSA (Fig. 1c). The results were mostly consistent with those of previous reports. Genotyping of *mecA* predicted the phenotype of MRSA (sensitivity: 95.3%; specificity: 93.3%; accuracy [95% CI]: 0.940 (0.904–0.963); Table 2). However, using the *mecA* genotype to predict drug sensitivity has some limitations. We found one MRSA isolate with the R4 genotype and four with the *mecA* R1 genotype that were susceptible to PENC. In contrast, 12 MRSA isolates with the *mecA* S2 genotype were PENC-resistant (Fig. 1c, Table 1). Among these isolates whose phenotypes were incorrectly predicted by *mecA* genotyping, 11/17 isolates belonged to CC59, which was the most prevalent CA-MRSA lineage in China.

mecA expression in ST59 SRRSH-MRSA isolates

Of the 73 CC59 MRSA isolates with the *mecA* S2 genotype, 11 had the PENC-resistant phenotype (Fig. 1c, Table 1). To determine the mechanism of PENC resistance in these isolates, we randomly selected five S2-R isolates among the ST59 MRSA lineage whose *mecA* genotype was S2 but phenotype was PENC-resistant. For comparison, five ST59 MRSA isolates with the *mecA* S2 genotypes had matching phenotypes and five ST5 MRSA isolates with the R3 genotypes were randomly selected as reference groups. Compared with the *mecA* S2 genotype in ST59 MRSA isolates (median relative expression 0.98), with matching phenotypes and genotypes, the levels of *mecA* expression were higher in the S2-R isolates (median relative expression: 2.61, P<0.05). *mecA* expression was similar between the S2-R and R3 isolates (median relative expression: 2.19, P>0.05; Fig. 2). These results indicate the limitation of using a single factor to predict PENC susceptibility.

Phenotype predicted by random forest model

In order to establish an accurate and convenient model for predicting PENC susceptibility, we built random forest models with the core genome allelic profiles and the susceptibility of MRSA to PENC. First, we used NA-MRSA data to train a random forest model 1 (Fig. S1). In the NA-MRSA training set, 100% of 240 and 229 (99%) of 231 PENC-susceptible and PENC-resistant MRSA isolates, respectively, were correctly predicted (Fig. S2a). The accuracy in the training set reached 0.996 with a sensitivity of 100% and specificity of



Random Forest Model (Training Set)	AST	
	S	R
S	298	2
R	0	245

Random Forest Model (SRRSH-MRSA)	AST	
	S	R
S	103	11
R	5	165

Random Forest Model (NAM-MRSA)	AST	
	S	R
S	194	8
R	11	74

Fig. 3. Receiver operating characteristics curves of random forest prediction model 2. ROC of (a) training (b) SRRSH-MRSA validation sets (c) NAM-MRSA validation sets. AST, tests of susceptibility to penicillin with potassium clavulanate; R, resistant; S, susceptible; ROC, receiver operating characteristics curves; AUC, Area Under Curve.

99.1% (Table 2). In the validation set of 284 SRRSH-MRSA isolates, 96 of 108 (88.9%) and 165 of 176 (94%) were PENC-susceptible and PENC-resistant, respectively, were correctly predicted (Fig. S2b). The accuracy in the SRRSH-MRSA validation set reached 0.919 with a sensitivity of 88.9% and specificity of 93.8% (Table 2). As for the validation set collected in the previous study from the UK, the prediction model trained with domestic isolates in China had relatively low ability predict foreign isolates. The value of the AUROC was 0.58 (Fig. S2c).

To make the random forest model reliable for predicting the UK-derived isolates, we added the UK-derived isolates in the training set and built random forest model 2. The results from the retrained model showed that all the (298/298) PENC-susceptible and 99.2%(245/247) of the PENC-resistant MRSA isolates were correctly predicted in the training set (Fig. 3a). The accuracy in the training set reached 0.996 with a sensitivity of 100% and specificity of 99.1% (Table 2). In the validation set comprising 284 SRRSH-MRSA isolates, 95.4% (103/108) PENC-susceptible and 94% (165/176) PENC-resistant MRSA isolates were correctly predicted (Fig. 3b). The accuracy in the validation set of 284 SRRSH-MRSA reached 0.944 with a sensitivity of 95.4% and specificity of 93.8% (Table 2). In the validation set of 287 NAM-MRSA isolates, 94.6% (194/205) PENC-susceptible and 90.2% (74/82) PENC-resistant MRSA isolates were correctly predicted (Fig. 3c). The accuracy in the validation set of 287 NAM-MRSA reached 0.934 with a sensitivity of 94.6% and specificity of 90.2% (Table 2).The predictive ability of the model was verified using the values of the AUROC. The AUROC of the training set, SRRSH-MRSA validation set, and NAM-MRSA validation set were 1.00, 0.96, and 0.97, respectively, in the random forest model 2. The results showed that the predictive power of this model was excellent (Fig. 3).

The top ten genes in the MRSA core genome with the most important roles in PENC susceptibility prediction for MRSA were exported (Fig. S3). Among them, SACOL2352, SACOL0764, SACOL1122 and SACOL1522 had top ten values in both the mean decrease accuracy and mean decrease Gini (Fig. S3). SACOL2352 encodes tcaA protein. SACOL0764 encodes glycosyl transferase. SACOL1122 encodes cell cycle protein FtsW. SACOL1522 encodes elastin binding protein.

DISCUSSION

Bacterial drug resistance has become a growing concern. MRSA, a common pathogen, has imposed tremendous medical and economic burdens [17]. A severe CA-MRSA infection can kill a young, properly treated patient within days [18]. Therefore, safe, effective, economical, and convenient treatments for MRSA infections, and rapid analyses of MRSA susceptibility to drugs are needed. Penicillin is an established beta-lactam antibiotic that is widely used to treat infections in humans because of its wide spectrum of activity, oral availability, excellent pharmacokinetics, lack of toxicity, and bactericidal action [19]. The most prevalent CA-MRSA in the USA (USA300) was found to be susceptible to PENC [6]. Similar to USA300, CC59 is the most prevalent CA-MRSA in China and is gradually penetrating hospitals. With the ongoing development of sequencing and computer technology, rapid MRSA susceptibility analyses will soon be available. The present study investigated whether PENC could be used to treat MRSA, particularly CA-MRSA, infection in China. We created a random forest model for predicting MRSA susceptibility to PENC using the core genome allelic profiles of MRSA.

Our data showed that PENC susceptibility varied in different MRSA lineages. The overall proportion of PENC-susceptible MRSA was not high, but a high proportion of PENC susceptible in CC59 MRSA was observed, according to the breakpoint of 8 mg l^{-1} for amoxicillin/clavulanate from a previous study [20]. To maintain the blood concentration of clavulanic acid above 15 mg l^{-1} for 8 h, the dosage of potassium clavulanate in the amoxicillin-potassium clavulanate compound must be increased. This result indicates the potential of using PENC to treat CA-MRSA infection in China.

The development of rapid drug susceptibility analyses will render PENC treatment feasible. To predict MRSA susceptibility to PENC more accurately than that using five genotypes in the previous study [6], more details on the *mecA* sequence in MRSA isolates collected at our hospital were analysed, and the penicillin MICs of different *mecA* genotypes were determined. However, no mutations other than the known *mecA* E246G, [-7]G-T, and [-33]C-T mutations that could help predict the susceptibility of MRSA to PENC were found. One of the foundations of *mecA* genotyping for predicting drug susceptibility is a mutation in its promoter that affects the level of PBP2a expression. However, our findings for S2-R expression in isolates indicated that factors other than *mecA* can also affect PBP2a expression levels. Previous study showed PrsA and HtrA1 can affect the function of PBP2a [21] and high-level β-lactam resistance in *Staphylococcus aureus* is associated with RNA Polymerase alterations and fine tuning of gene expression [22]. These issues decreased the accuracy of prediction based on one gene. Then, we constructed a model for predicting MRSA susceptibility to PENC using the core genome of our MRSA isolates. For SRRSH-MRSA, the predicted power of the random forest model 2 was equivalent to that of single *mecA* gene prediction (0.944 VS 0.940). However, adding the UK-sourced MRSA data into the training set, we found that the AUROC of the NAM-MRSA validating set increased from 0.58 to 0.97. This showed the strong 'learning' ability of the random forest algorithm model. As more MRSA data are added to the training set, PENC susceptibility predictions will become more reliable. On the other hand, our results indicate the source of the training set data does affect the scope of application of the prediction model. In this case, including more MRSA isolates from different lineages significantly increase the AST prediction power of random forest model.

The random forest model can also identify key genes associated with the predicted phenotype [12]. Here, we detected genes that may be most closely associated with MRSA susceptibility to PENC. They may contribute to the regulation of *mecA* expression, facilitate PBP2a protein transportation and localisation on the cell membrane, or/and correct PBP2a folding. Further studies are needed to confirm the roles of these genes in recovering the susceptibility to β-lactams in MRSA.

Though the random forest model showed a high prediction power for MRSA susceptibility to PENC in our isolates, there are limitations to our work. First, although the machine learning algorithm showed learning ability, a weakness of most machine learning methods is that they cannot predict what they have not learnt. We noticed the importance of including more isolates from different regions to build a reliable prediction model using genomic data. Second, the whole genome may theoretically predict a series of drugs in the future. In this study, only one drug, which was not routinely tested for susceptibility, was analysed and studied. The application of our prediction model to other drugs or other AST results such as MIC values should be studied further. Despite these limitations, our data illustrate the potential of using machine learning for drug susceptibility prediction with cgMLST results obtained from commercial software.

In conclusion, our data showed that machine learning can be applied to the prediction of antimicrobial susceptibility from cgMLST results. Although the source of the training set data affects the scope of application of the prediction model, we propose that the machine learning approach used in combination with microbiological genomic data can play an important role in the ongoing effort to reduce the burden of antimicrobial resistance worldwide.

Supplementary data

Supplementary materials included the cgMLST results of MRSA isolates (Tables S1) in this study are available at Microbial Genomics online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Funding information
This study was funded by the National Natural Science Foundation of China (grant 81971977), Zhejiang Provincial Natural Science Foundation of China (grant LQ20H190005), and National Science and Technology Major Project of the 13th 5 year Plan (Project No: 2018ZX10714002).

Acknowledgements
We thank Xiaoliang Ba for helping us revise the manuscript. We thank Zhouqi Li for guidance with our drawing. We thank Editage for English language editing.

Author contributions
Y.Y. and Y.C., designed the study and developed the approach. H.Z., F.Z., P.L., S.J., L.S., Y.C., Z.W., and S.J., collected the data. H.Z. and P.L., created the random forest models. H.Z., and Y.Z., participated in the drawing of the article. H.Z., F.Z., and Y.J., performed the genetic context analysis. All authors analysed the results. H.Z., performed the experimental verification. All authors discussed the results and its implications. H.Z., F.Z., and Y.C., drafted the manuscript. All authors edited and approved the final manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev* 2015;28:603–661.
2. Geriak M, Haddad F, Rizvi K, Rose W, Kullar R, *et al*. Clinical data on daptomycin plus ceftaroline versus standard of care monotherapy in the treatment of methicillin-resistant *Staphylococcus aureus* bacteremia. *Antimicrob Agents Chemother* 2019;63.
3. Mediavilla JR, Chen L, Mathema B, Kreiswirth BN. Global epidemiology of community-associated methicillin resistant *Staphylococcus aureus* (CA-MRSA). *Curr Opin Microbiol* 2012;15:588–595.
4. Liu C, Bayer A, Cosgrove SE, Daum RS, Fridkin SK, *et al*. Clinical practice guidelines by the infectious diseases society of america

for the treatment of methicillin-resistant *Staphylococcus aureus* infections in adults and children: executive summary. *Clin Infect Dis* 2011;52:285–292.

5. Monaco M, Pimentel de Araujo F, Cruciani M, Coccia EM, Pantosti A. Worldwide epidemiology and antibiotic resistance of *Staphylococcus aureus*. *Curr Top Microbiol Immunol* 2017;409:21–56.
6. Harrison EM, Ba X, Coll F, Blane B, Restif O, *et al*. Genomic identification of cryptic susceptibility to penicillins and beta-lactamase inhibitors in methicillin-resistant *Staphylococcus aureus*. *Nat Microbiol* 2019;4:1680–1691.
7. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, *et al*. Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci Rep* 2018;8:421.
8. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, *et al*. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLoS Comput Biol* 2019;15:e1007349.
9. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, *et al*. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *J Clin Microbiol* 2019;57.
10. Hyun JC, Kavvas ES, Monk JM, Palsson BO. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput Biol* 2020;16:e1007608.
11. Aytan-Aktug D, Clausen PTLC, Bortolaia V, Aarestrup FM, Lund O. Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks. *mSystems* 2020;5.
12. Recker M, Laabei M, Toleman MS, Reuter S, Saunderson RB, *et al*. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nat Microbiol* 2017;2:1381–1388.
13. Chen Y, Sun L, Wu D, Wang H, Ji S, *et al*. Using core-genome multilocus sequence typing to monitor the changing epidemiology of methicillin-resistant *Staphylococcus aureus* in a teaching hospital. *Clin Infect Dis* 2018;67:S241–S248.
14. CLSI. *Performance Standards for Antimicrobial Susceptibility Testing*. Wayne, PA: Clinical and Laboratory Standards Institute; 2019.
15. Bronner S, Murbach V, Peter JD, Levêque D, Elkhaïli H, *et al*. Ex vivo pharmacodynamics of amoxicillin-clavulanate against beta-lactamase-producing *Escherichia coli* in a yucatan miniature pig model that mimics human pharmacokinetics. *Antimicrob Agents Chemother* 2002;46:3782–3789.
16. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, *et al*. Updating benchtop sequencing performance comparison. *Nat Biotechnol* 2013;31:294–296.
17. Vuong C, Yeh AJ, Cheung GY, Otto M. Investigational drugs to treat methicillin-resistant *Staphylococcus aureus*. *Expert Opin Investig Drugs* 2016;25:73–93.
18. Chen Y, Hong J, Chen Y, Wang H, Yu Y, *et al*. Characterization of a community-acquired methicillin-resistant sequence type 338 *Staphylococcus aureus* strain containing a staphylococcal cassette chromosome mec type V(T). *Int J Infect Dis* 2020;90:181–187.
19. Foster TJ. Can β-lactam antibiotics be resurrected to combat MRSA? *Trends Microbiol* 2019;27:26–38.
20. Rubin JE, Ball KR, Chirino-Trejo M. Antimicrobial susceptibility of *Staphylococcus aureus* and *Staphylococcus pseudintermedius* isolated from various animals. *Can Vet J* 2011;52:153–157.
21. Roch M, Lelong E, Panasenko OO, Sierra R, Renzoni A, *et al*. Thermosensitive PBP2a requires extracellular folding factors PrsA and HtrA1 for *Staphylococcus aureus* MRSA β-lactam resistance. *Commun Biol* 2019;2:417.
22. Panchal VV, Griffiths C, Mosaei H, Bilyk B, Sutton JAF, *et al*. Evolving MRSA: High-level β-lactam resistance in *Staphylococcus aureus* is associated with RNA Polymerase alterations and fine tuning of gene expression. *PLoS Pathog* 2020;16:e1008672.