Genomic Epidemiology of Methicillin-Resistant *Staphylococcus aureus* in Two Cohorts of High-Risk Military Trainees

Robyn S. Lee PhD[1,2]*, Eugene V. Millar PhD[3,4], Alanna Callendrello BA[2], Caroline E. English BA[3,4], Emad M. Elassal MS[3,4], Michael W. Ellis MD[5], Jason W. Bennett MD MSPH[6,7], William P. Hanage PhD[2]

1. Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2. The Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
3. Infectious Disease Clinical Research Program, Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, Bethesda, MD, USA
4. Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD, USA
5. University of Toledo College of Medicine and Life Sciences, Toledo, OH, USA
6. Walter Reed Army Institute of Research, Silver Spring, MD, USA
7. Department of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

*Address correspondence to:
Dr. Robyn S. Lee, PhD
Epidemiology Division
Dalla Lana School of Public Health, University of Toronto
Health Sciences Building
155 College Street, Room HS 586
Toronto, Ontario, Canada
M5T 3M7
Email: robyn.s.c.lee@gmail.com

Alternative corresponding author:
Dr. William P. Hanage, PhD
Harvard T.H. Chan School of Public Health
677 Huntington Avenue
Kresge Building, 5th floor
Boston, MA, USA
02115
whanage@hsph.harvard.edu
617-432-7955

Text word count: 3039
Figures: 5
Tables: 3
Key words: Methicillin-resistant Staphylococcus aureus ; genomic epidemiology ; military ; community-acquired MRSA ; transmission

Summary (40/40 words): US Infantry trainees were followed through training for MRSA skin and soft tissue infection, swabbing for colonization at 2-4 week intervals. Sequencing suggests serial importations of diverse strains on base, followed by transmission mostly within platoons, involving multiple anatomical sites.

45 **Abstract – 246/250 words**

46 *Background.* MRSA skin and soft tissue infection (SSTI) is a significant cause of morbidity in

47 military trainees. To guide interventions, it is critical we understand the epidemiology of MRSA

48 in this population.

49 *Methods.* Two cohorts ('companies') of US Army Infantry trainees (N=343) at Fort Benning,

50 GA, USA, were followed during their training cycles (Jun.-Dec. 2015). Trainees had nares,

51 oropharynx, perianal and inguinal areas swabbed for MRSA colonization at five ~2-4 week

52 intervals, and monitored for SSTI throughout training. Epidemiological data were collected.

53 Isolates were sequenced using Illumina HiSeq and NovaSeq. Single-nucleotide polymorphisms

54 and clusters were identified. Multi-locus sequence type (MLST) and antimicrobial resistance

55 genes were predicted from *de novo* assemblies.

56 *Results.* 87 trainees were positive at least once for MRSA (12 had SSTI, 2 without any

57 colonization). Excluding those positive at baseline, 43.7% were colonized within the first month

58 of training. 244/254 samples were successfully sequenced (including all SSTI). ST8 (n=135,

59 100% of SSTI), ST5 (n=81) and ST87 (n=21) were the most represented. Three main Clusters

60 were identified, largely corresponding to these STs. Sub-analyses within Clusters showed

61 multiple importations of MRSA, with transmission subsequently predominantly within, rather

62 than between, platoons in each company. Over 50% of trainees were colonized only at other

63 anatomical sites; restricting analyses to nares missed substantial transmission.

64 *Conclusions.* Serial importations of MRSA into this high-risk setting likely contribute to the

65 ongoing burden of MRSA colonization and infection among military trainees. Sampling multiple

66 anatomical sites is critical for comprehensive characterization of MRSA transmission.

67

68 **Introduction**

69 Community-associated methicillin-resistant *Staphylococcus aureus* (CA-MRSA) is a major

70 cause of skin and soft tissue infections (SSTI) [1]. Military personnel have an elevated risk of

71 MRSA SSTI compared to the general population with new recruits disproportionately affected

72 [2, 3]. SSTI outbreaks compromise not only training cycles, but also operational readiness

73 among deployed forces [4]. The annual cost of SSTIs to the US military is approximately 14-32

74 million US dollars [5]. Past SSTI prevention strategies in military populations have included

75 nasal decolonization [6] as well as chlorhexidine-based body wash [7, 8], based on the principle

76 that MRSA colonization precedes and is necessary for infection [9]. However, neither strategy

77 prevented SSTI in these high-risk populations [6, 8].

78

79 To develop effective prevention strategies, it is critical we understand the epidemiology and

80 transmission dynamics of MRSA in this population. Recently, whole genome sequencing (WGS)

81 has been used to describe the epidemiology of MRSA colonization [4] and infection among US

82 Army trainees [10]. A key limitation of this study was that WGS was restricted to trainees with

83 pulse-field gel electrophoresis (PFGE) type USA300 MRSA, the most common cause of CA-

84 MRSA SSTI. As such, the conclusions that could be drawn about overall MRSA transmission

85 dynamics, and prevalence of different strains circulating in this population, were greatly limited.

86 To address this concern, we conducted a prospective cohort study of MRSA in the same trainee

87 population, following them throughout their training cycle for colonization and infection events.

88 All MRSA-positive samples from multiple anatomical sites were included.

89

3

90   Our study found a high prevalence of MRSA colonization at baseline, comprising a genetically

91   diverse population imported from across the US by recruits entering basic training. This was

92   followed by substantial person-to-person transmission, particularly among soldiers belonging to

93   the same 'platoons' (i.e., those in the same training company who share the same living

94   quarters). Moreover, the observed transmission events frequently involved different anatomical

95   sites - irrespective of nasal colonization. As nasal swabs are routinely and almost exclusively

96   used to detect MRSA colonization in clinical settings, this suggests MRSA transmission may be

97   drastically under-estimated with current infection control methodology – a finding with

98   significant implications for the control of MRSA.

99

100   **Methods**

101   *Overview.* We conducted a longitudinal cohort study of SSTI among US Army Infantry trainees

102   at Fort Benning, Georgia, from June-December 2015. Two companies (designated A and B)

103   were included, which commenced the 14-week training cycle at two different timepoints during

104   peak SSTI seasons (June and September, respectively). Infantry training companies are

105   comprised of ~200 soldiers, further segregated into four platoons comprised of ~50 soldiers. See

106   **Supplementary Methods** for additional detail.

107

108   *MRSA screening.* MRSA colonization swabs were obtained at baseline and at ~2-4 week

109   intervals over five study visits. [6] Anterior nasal, oropharyngeal, inguinal and perianal swabs

110   were collected (see **Supplementary Methods**). Swabs were initially processed at the Benning

111   Martin Army Community Hospital clinical microbiology laboratory as previously described [8].

112

113   *Epidemiologic data collection.* Trainees completed questionnaires at baseline for demographic

114   and MRSA risk factor information, and at follow-up visits to collect data on personal hygiene

115   practices.

116

117   *SSTI case definition.* Trainees were considered to have probable SSTI if they presented with

118   cellulitis, abscess, folliculitis, impetigo, paronychia, infected blister, or pilonidal cyst; SSTI cases

119   were defined as those with *S. aureus* cultured from the clinical site. Case detection is described

120   in the **Supplementary Methods**.

121

122   *Genome sequencing and bioinformatics.* See **Supplementary Methods** for details. In brief,

123   genomes were sequenced using the Illumina HiSeq and NovaSeq platforms. As most samples

124   were known to be USA300, we initially aligned reads to the USA300 TCH1516 reference

125   genome (National Center for Biotechnology Information [NCBI] Accession CP000730.1) using

126   the Burrows-Wheeler Aligner MEM algorithm (v.0.7.15, [12]) and identified variants compared

127   to this reference. Prophage sequences were identified using Phaster (http://phaster.ca, [13]) and

128   SNPs in these regions were masked.

129

130   Concatenated SNP alignments were used to generate maximum likelihood trees with IQ-TREE

131   (v.1.6.8, [14]). To infer a more accurate phylogeny for the analysis including all STs, we masked

132   predicted recombination sites using Gubbins (v.2.3.4, [15]). Hierarchical Bayesian Analysis of

133   Population Structure (RhierBAPS v.1.1.0, [16]) was used to discriminate clusters of closely-

134   related samples in R (v.3.5.2). Analyses were then repeated for each major Cluster identified

135   using a closely-related reference genome.

136   *De novo* assemblies were also generated, using SPAdes (v.3.8.0, [17]) and Unicycler (v.0.4.7,

137   Wick R, https://github.com/rrwick/Unicycler). Sequence Types (STs) according to the multi-

138   locus sequence typing (MLST) scheme [18] were predicted using 'mlst' (v.2.10, Seemann T,

139   https://github.com/tseemann/mlst) and Abricate (v.0.8, Seemann T,

140   https://github.com/tseemann/abricate) was used with ResFinder [19] and the Virulence Factor

141   DataBase (VFDB, [20]) to detect genes associated with antimicrobial resistance and virulence,

142   respectively. *mupA* and *qacA* genes were predicted using gene-puller.py

143   (https://github.com/kwongj/gene-puller).

144

145   ***Statistical analysis.*** Categorical comparisons were done using the Chi-Square test where cell

146   counts > 5, or Fisher's Exact test if otherwise. Distributions were compared using the Wilcoxon-

147   Mann-Whitney test. Conditional logistic regression was used to compare hygiene variables

148   between trainees who became colonized with time-matched, MRSA-negative controls

149   (**Supplementary Methods**). Analyses were done in Stata15 (StataCorp, College Station, TX,

150   USA).

151

152   ***Data availability.*** Sequence data is on NCBI's Sequence Read Archive under BioProject

153   PRJNA587530.

154

155   ***Ethics approval.*** The study was approved by the Uniformed Services University Institutional

156   Review Board (IDCRP-090) and Harvard T.H. Chan School of Public Health Institutional

157   Review Board (IRB17-0802).

158

159 **Results**

160 We followed two Infantry training companies in the summer and fall of 2015. 343 trainees

161 participated (132 and 211 in companies A and B, respectively), with 86.6% (297/343) completed

162 all study visits (**Supplementary Results**). 87 (25.4%) of trainees were MRSA-positive during

163 the study; 85 had at least one positive colonization swab and 12 were diagnosed with MRSA

164 SSTI. Among the latter, only seven were colonized prior to or at the same time as their SSTI

165 (five with the infecting strain); three were colonized only at later timepoints, and two were never

166 colonized (**Supplementary Table 1**).

167

168 Among those colonized, thirty-five had a single positive sample; the oropharynx and perianal

169 sites were most frequently colonized, each representing 31.4% of swabs, followed by nasal

170 (20.0%), and inguinal (17.1%) sites. Fifty trainees were colonized at multiple anatomical sites

171 and/or timepoints. The number of repeat positive samples did not differ significantly by

172 anatomical site of isolation (p = 0.16), although inguinal swabs tended to be less frequently

173 positive (**Figure 1**).

174

175 *Risk factors for colonization or infection*

176 There were no significant differences in age, self-reported race, or MRSA risk factors at baseline

177 (**Table 1**) comparing those with colonization/infection to those who were never MRSA-positive.

178 Epidemiological characteristics of SSTI cases are described in the **Supplementary Results**.

179

180

181

7

182   *Timing of MRSA acquisition*

183   Twenty-seven trainees were colonized at baseline (see **Supplementary Results**). Excluding

184   these, most trainees appeared to have acquired MRSA shortly after starting training (**Figure 2**);

185   43.7% of all positive samples were collected during the first month of training. There were no

186   significant differences in personal hygiene practices over the same timepoint between trainees

187   with new colonization and those who were never positive (**Supplementary Results**,

188   **Supplementary Table 2**).

189

190   *Genomic analyses*

191   High-quality sequencing data was available for 244/254 (96.1%) MRSA-positive samples from

192   85 individuals; further analyses presented are restricted to this subset. ST8 (USA300) was the

193   dominant strain identified (**Table 2**), and included all cases of SSTI. The virulence factor

194   Panton-Valentine Leukocidin was present in all clinical isolates (as demonstrated by intact *lukS*

195   and *lukF* genes) and nearly half of colonization isolates (**Table 2**). All samples were PCR-

196   positive for *mecA,* while 98.8% were predicted to have *mecA in silico*; 2/3 isolates lacking *mecA*

197   *in silico* were different STs than other isolates from those trainees, indicating that mixed

198   genotypes may have been present at sampling (**Supplementary Table 3**). To assess the potential

199   utility of mupirocin and chlorhexidine for decolonization, *mupA* and *qacA* genes were predicted.

200   *mupA* was only detected in colonization isolates (all ST5) from four trainees (and one ST8

201   clinical isolate from another trainee), while *qacA* was detected in colonization isolates from three

202   trainees. Correlation between PFGE and ST are shown in **Supplementary Table 4**.

203

204

205   *Strain replacement*

206   Strain replacement was rare (**Supplementary Figure 1**), occurring only four times during the

207   study (excluding co-colonization). Two instances involved different STs; trainee 2013 was

208   initially colonized with an ST5 strain but developed an ST8-SSTI, while 2031 was colonized

209   with ST8 but later colonized with ST5. The remaining instances involved trainees 1007 and

210   2061. These trainees had isolates from the same ST that were separated from their later isolates

211   by > 97 core SNPs, suggestive of exogenous re-infection.

212

213   *Overall transmission analyses*

214   See **Supplementary Results** for quality metrics. The recombination-adjusted maximum

215   likelihood tree is shown in **Figure 3**, excluding one highly divergent isolate (1134_5N, see

216   **Supplementary Results**). Five Clusters were identified, with three containing >1 isolate;

217   Cluster-1 (n=139, 97.1% ST8); Cluster-2 (n=81, all ST5); and Cluster-4 (n=21, all ST87).

218   Pairwise adjusted core SNP distances between isolates indicate that transmission was

219   predominantly within, rather than between, platoons in the same Companies (**Figure 4**).

220

221   *Sub-analyses of transmission, by Cluster*

222   To provide the highest resolution, analyses were re-run for each Cluster using a closely-related

223   reference (**Supplementary Table 5**). As ST8s represent the largest proportion of samples

224   collected, the maximum likelihood tree for Cluster 1 is shown in **Figure 5** (see **Supplementary**

225   **Figures 2** and **3** for Clusters 2 and 4, respectively). Within Cluster 1, there were nine sub-groups

226   each representing isolates from 1 to 24 trainees. These sub-groups were separated from one

227   another by a minimum of 34 SNPs (**Table 3**). In nearly all sub-groups, there were isolates from

228   different trainees that shared close genetic relationships, indicative of recent transmission (**Table**

229   **3**). Overlaying company and platoon (**Figure 5**) suggests that such transmission was largely

230   restricted to the same platoons, with a small amount across platoons in the same company.

231   Isolates from different companies were only found in the largest sub-group (1-1), with as few as

232   6 core SNPs separating them (including repeat isolates per trainee, the median core SNP distance

233   was 71 across companies, IQR 35-100). Given these companies were separated in time and

234   space, this suggests undetected transmission via an intermediate source.

235

236   Nearly all transmission appears to be within rather than between platoons for Clusters 2 and 4 as

237   well (**Supplementary Figures 2** and **3**). The exception to this was sub-group 1 in Cluster 4,

238   where multiple platoons are implicated; however, all six trainees in this sub-group were

239   colonized at baseline (**Supplementary Results** and **Supplementary Table 6**). There was no

240   evidence of transmission across companies in either Cluster.

241

242   *Anatomical sites and their role in transmission*

243   The nares are commonly thought to play a key role in transmission, but > 50% of trainees

244   (regardless of Cluster/ST) were positive only at other anatomical sites (**Table 3**). To examine the

245   impact of this, we repeated our transmission analyses only including isolates from the nares and

246   SSTI (**Supplementary Table 7**) and found that transmission was substantially under-estimated,

247   with as many as 100% of trainees being missed in some Clusters/sub-groups.

248

249

250

10

251 *ST8s in context of previous analyses*

252 We compared our contemporaneous samples to historical clinical and colonization isolates from

253 trainees with ST8-associated SSTI at Fort Benning [4, 10] (described further in the

254 **Supplementary Results**), and found historical isolates largely clustered independently of more

255 recent samples (**Supplementary Figure 4**). The median core pairwise distance between these

256 isolates and our 2015 samples was 111 SNPs (IQR 80-137). 99% of pairs (14,187/14,385) were

257 separated by > 15 SNPs, suggesting a single reservoir of MRSA is unlikely to be driving ongoing

258 transmission of ST8-MRSA on base.

259

260 **Discussion**

261 MRSA SSTI continues to be a major source of morbidity in military trainees in the United States

262 [2, 3] and other countries [21, 22]. In this study, we integrated genomics with comprehensive

263 epidemiologic data to investigate MRSA transmission in US Army Infantry trainees at Fort

264 Benning, GA, a group known to be at increased risk for MRSA colonization and infection.

265

266 We followed individuals prospectively through training for MRSA colonization and SSTI,

267 including all positive colonization isolates from these trainees over time. We found there were

268 multiple STs present at baseline, with subsequent transmission predominantly within, rather than

269 between platoons - reinforcing the importance of close proximity for MRSA transmission.

270 Consistent with previous ST8 analyses [4, 10], there appeared to be little transmission across

271 training classes (i.e., companies) over time, suggesting no single endemic strain in this context.

272 By contrast, the persistent burden of MRSA colonization and infection is likely due to the serial

273 importation of CA-MRSA strains with each incoming group of new trainees. With > 30,000

274   trainees from multiple US states and territories cycling through Fort Benning annually, the

275   diversity of MRSA strains being introduced into this single geographic site is likely considerable.

276

277   Surprisingly, a large proportion of trainees (27/343, [7.9%]) were positive at baseline or positive

278   shortly after training onset. This suggests that particular phases of training may serve as critical

279   'control points' wherein interventions to reduce transmission and/or SSTI could be implemented.

280   As no specific risk factors for MRSA colonization were identified, we propose that risk of

281   acquisition may be predominantly mediated by exposure to circulating strains and shared training

282   activities, rather than individual hygiene practices - though it is possible the small numbers may

283   preclude detection of significant differences between our comparison groups.

284

285   The nares are considered the primary reservoir for *S. aureus* [23]. In the current study, however,

286   we found that more than half (55.3%) of those colonized were exclusively positive at other

287   anatomical sites. Screening of these different sites is therefore critical to accurately detect

288   colonization and resolve transmission. This finding may have important implications for the

289   control of MRSA in military trainees, as MRSA prevalence, transmission, and success of

290   decolonization efforts, have previously been largely estimated from nasal carriage alone [24]. A

291   randomized controlled trial among Army combat medics in training [6] found that, while

292   intranasal mupirocin reduced nasal colonization with MRSA, SSTI rates remained unaffected.

293   Similar results were found in a large-scale, field-based evaluation of chlorhexidine-based body

294   wash [7, 8]. One possibility is that inoculation via fomites may contribute to SSTI, however,

295   another possibility is that trainees colonized exclusively at non-nasal sites (not included in these

296   studies due to the design) served as a constant reservoir for re-colonization/infection. These

297 explanations need not be mutually exclusive. This suggests that, given the negligible levels of

298 resistance to either treatment, mupirocin coupled with chlorhexidine gluconate washes may yet

299 be viable tools for SSTI reduction in this population, potentially coupled with oral antimicrobials

300 [25]. However, for such interventions to work, this would likely require *all* colonized individuals

301 are identified and treated, regardless of anatomical site.

302

303 This study has a number of major strengths, including a high participation rate with > 85% of

304 participants completing all visits. Because this was conducted in a military training setting, the

305 schedule, location, type of activities, and risk of MRSA acquisition were relatively uniform for

306 trainees across all platoons at all visits. Additionally, the prospective design, with screening for

307 colonization and collection of corresponding epidemiological data at regular intervals throughout

308 training allowed us to detect such events as they occurred and assess the potential association

309 with specific behaviours that preceded them. This degree of follow-up is rarely feasible outside

310 of a hospital setting. Another strength was the screening of multiple anatomical sites over time;

311 this not only allowed us to more accurately detect colonization than previous studies in this

312 population [6, 7, 24] but examine the role they play in transmission. Finally, previous studies

313 focused exclusively on ST8, with sampling exclusively from cases of SSTI and select controls

314 [4, 10]; by including all STs, we were able to provide the first fully comprehensive analysis of

315 population structure and transmission dynamics of MRSA circulating in this setting.

316

317 There are several limitations of this work. Due to resource requirements, we focused recruitment

318 on two (out of 37) training classes from Jun. - Dec. 2015. Though our analysis suggests limited

319 transmission across companies (i.e., classes), it is possible some transmission events were not

320   detected. Permanent staff on base (e.g. drill sergeants, health care personnel, etc.) may also

321   represent a reservoir of MRSA, though our analysis suggest this is less likely (as there were

322   multiple circulating strains, with evidence of importation from outside the base). Staff may also

323   be source of unrecognized transmission, potentially even during in-processing, but unfortunately,

324   were not included in the current study. Similarly, sampling of high touch common surfaces (e.g.

325   countertops, weight benches, wrestling mats, etc.) was not conducted as part of this investigation.

326   Future studies should address these gaps, to better understand the role these may play in MRSA

327   transmission dynamics.

328

329   We also note that the included cohorts were purposely chosen based on peak MRSA SSTI

330   seasonality [8], with some studies reporting increased colonization in warmer months [26, 27].

331   MRSA prevalence in our classes may therefore be higher than would be found in other cohorts.

332   However, baseline prevalence of MRSA colonization among combat medic trainees [6] in Jan. –

333   Dec. 2005 was only 3.9% (based on nasal carriage) and restricting our estimates to nasal carriage

334   yielded a similar proportion (2.0%, p=0.08), suggesting our findings are likely generalizable to

335   other training courses and/or periods. Another potential limitation of this work is that, for privacy

336   reasons, perianal and inguinal swabs were collected by the trainees themselves, rather than by

337   study personnel. Though they were instructed beforehand, it is possible these swabs were not

338   collected in a consistent fashion and may be more likely to yield false negative results. However,

339   we did not find a significant difference in frequency of positivity across anatomical sites. Finally,

340   personal hygiene practices were self-reported at routine study visits for the preceding month,

341   potentially leading to misclassification due to poor recall. However, as colonization status was

342   not known at the time of collecting this information (and was not reported to trainees during the

343   study), we would not expect this misclassification to be differential.

344

345   Through this in-depth study, we identified critical timepoints in military training associated with

346   MRSA colonization and highlight the importance of different anatomical sites in mediating this

347   transmission. Given our historical reliance on nasal screening for detection of MRSA

348   colonization, and subsequent interventions to prevent infection, this work may have important

349   implications not only for control of MRSA in the military, but in-patient settings as well. In order

350   to halt MRSA transmission in the military, and other vulnerable populations, increased screening

351   to include these other anatomical sites may be necessary; however, cost and feasibility of such

352   interventions must also be investigated.

353

374   **Contributions.** EVM, MWE, and JB designed and ran the parent longitudinal cohort study.

375   WPH and RSL conceived the current genomic study. RSL designed and was responsible for the

376   bioinformatics and epidemiological analyses, designing/preparing the figures and tables, and

377   writing the manuscript. CEE, EE, and AC were responsible for isolating the bacteria, and culture,

378   and AC prepared the sequencing libraries. WPH reviewed the initial draft. All authors were

379   involved in interpreting the data, and reviewed and provided feedback on the manuscript. RSL

380   takes full responsibility for the content of the paper and decision to submit for publication. Please

381   note, the views expressed are those of the authors and do not necessarily reflect the official

382   policy or position of the Henry M. Jackson Foundation for the Advancement of Military

383   Medicine, the Uniformed Services University, the Department of the Army, the Department of

384   Defense, nor the U.S. Government. Mention of trade names, commercial products, or

385   organizations does not imply endorsement by the U.S. Government.

386

387

388

References

1. David MZ, Daum RS. Community-associated methicillin-resistant Staphylococcus aureus: epidemiology and clinical consequences of an emerging epidemic. Clin Microbiol Rev **2010**; 23(3): 616-87.
2. Landrum ML, Neumann C, Cook C, et al. Epidemiology of Staphylococcus aureus blood and skin and soft tissue infections in the US military health system, 2005-2010. JAMA **2012**; 308(1): 50-9.
3. America DoDUSo. Cellulitis and abscess, active components, US Armed Forces, 2002-2005. , **2006**.
4. Millar EV, Rice GK, Elassal EM, et al. Genomic Characterization of USA300 Methicillin-Resistant Staphylococcus aureus (MRSA) to Evaluate Intraclass Transmission and Recurrence of Skin and Soft Tissue Infection (SSTI) Among High-Risk Military Trainees. Clin Infect Dis **2017**; 65(3): 461-8.
5. Lee BY, Singh A, David MZ, et al. The economic burden of community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA). Clin Microbiol Infect **2013**; 19(6): 528-36.
6. Ellis MW, Griffith ME, Dooley DP, et al. Targeted intranasal mupirocin to prevent colonization and infection by community-associated methicillin-resistant Staphylococcus aureus strains in soldiers: a cluster randomized controlled trial. Antimicrob Agents Chemother **2007**; 51(10): 3591-8.
7. Millar EV, Chen WJ, Schlett CD, et al. Frequent use of chlorhexidine-based body wash associated with a reduction in methicillin-resistant Staphylococcus aureus nasal colonization among military trainees. Antimicrob Agents Chemother **2015**; 59(2): 943-9.
8. Ellis MW, Schlett CD, Millar EV, et al. Hygiene strategies to prevent methicillin-resistant Staphylococcus aureus skin and soft tissue infections: a cluster-randomized controlled trial among high-risk military trainees. Clin Infect Dis **2014**; 58(11): 1540-8.
9. Miller LG, Diep BA. Clinical practice: colonization, fomites, and virulence: rethinking the pathogenesis of community-associated methicillin-resistant Staphylococcus aureus infection. Clin Infect Dis **2008**; 46(5): 752-60.
10. Millar EV, Rice GK, Schlett CD, et al. Genomic epidemiology of MRSA infection and colonization isolates among military trainees with skin and soft tissue infection. Infection **2019**.
11. Fritz SA, Hogan PG, Camins BC, et al. Mupirocin and Chlorhexidine Resistance in Staphylococcus aureus in Patients with Community-Onset Skin and Soft Tissue Infections. Antimicrobial Agents and Chemotherapy **2013**; 57(1): 559-68.
12. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv **2013**: 1303.3997v2.
13. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res **2016**; 44(W1): W16-21.
14. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **2015**; 32(1): 268-74.

431   15.   Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of
432   recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res **2015**;
433   43(3): e15.
434   16.   Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. RhierBAPS: An R
435   implementation of the population clustering algorithm hierBAPS. Wellcome Open Res
436   **2018**; 3: 93.
437   17.   Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its
438   applications to single-cell sequencing. J Comput Biol **2012**; 19(5): 455-77.
439   18.   Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for
440   characterization of methicillin-resistant and methicillin-susceptible clones of
441   Staphylococcus aureus. J Clin Microbiol **2000**; 38(3): 1008-15.
442   19.   Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial
443   resistance genes. J Antimicrob Chemother **2012**; 67(11): 2640-4.
444   20.   Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big
445   data analysis--10 years on. Nucleic Acids Res **2016**; 44(D1): D694-7.
446   21.   Aamot HV, Eskonsipo PKJ, Jorgensen SB, Blomfeldt A. Staphylococcus aureus
447   colonization during military service: a prospective cohort study. Clin Microbiol Infect
448   **2018**; 24(7): 744-8.
449   22.   Nivesvivat T, Janthayanont D, Mungthin M, et al. Methicillin-susceptible Staphylococcus
450   aureus skin infections among military conscripts undergoing basic training in Bangkok,
451   Thailand, in 2014. BMC Res Notes **2016**; 9: 179.
452   23.   Mulcahy ME, McLoughlin RM. Host-Bacterial Crosstalk Determines Staphylococcus
453   aureus Nasal Colonization. Trends Microbiol **2016**; 24(11): 872-86.
454   24.   Ellis MW, Hospenthal DR, Dooley DP, Gray PJ, Murray CK. Natural history of community-
455   acquired methicillin-resistant Staphylococcus aureus colonization and infection in
456   soldiers. Clin Infect Dis **2004**; 39(7): 971-9.
457   25.   Simor AE PE, McGeer A, Konvalinka A, LOeb M, Devlin HR, Kiss A. Randomized
458   Controlled Trial of Chlorhexidine Gluconate for Washing, Intranasal Mupirocin,
459   and Rifampin and Doxycycline Versus No Treatment for the Eradication of Methicillin-Resistant
460   Staphylococcus aureus Colonization. Clin Infect Dis **2007**; 44(178-85).
461   26.   Leekha S, Diekema DJ, Perencevich EN. Seasonality of staphylococcal infections. Clin
462   Microbiol Infect **2012**; 18(10): 927-33.
463   27.   Mork RL, Hogan PG, Muenks CE, et al. Comprehensive modeling reveals proximity,
464   seasonality, and hygiene practices as key determinants of MRSA colonization in exposed
465   households. Pediatr Res **2018**; 84(5): 668-76.
466   28.   SNP-sites: rapid efficient extraction of SNPs from multi- FASTA alignments.
467   29.   Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially explicit
468   clustering of DNA sequences with BAPS software. Mol Biol Evol **2013**; 30(5): 1224-8.
469   30.   Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and
470   annotation of phylogenetic and other trees. Nucleic Acids Res **2016**; 44(W1): W242-5.
471   31.   Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y, McInerny G. ggtree: anrpackage for
472   visualization and annotation of phylogenetic trees with their covariates and other
473   associated data. Methods in Ecology and Evolution **2017**; 8(1): 28-36.
474

475

476
477 **Tables.**
478
479 **Table 1. Clinical and epidemiological characteristics of trainees at baseline (N=343)**
480

Characteristic	Colonized and/or SSTI (n=87)	Never colonized and no SSTI (n=256)	P value
Age, average (SD)	19.2 (1.7)	19.4 (2.5)	0.42 [a]
Hispanic ethnicity, n (%)	15 (17.2)	40 (15.6)	0.72 [b]
Race, n (%)			
White	64 (73.6)	182 (71.1)	
Black	8 (9.2)	23 (9.0)	
Asian	0 (0)	6 (2.3)	
American Indian	1 (1.1)	5 (2.0)	
Pacific Islander	0 (0)	3 (1.2)	
Other [c]	7 (8.0)	20 (7.8)	
> 1 race reported	7 (8.0)	17 (6.6)	0.88 [d]
Lived alone prior to training, n (%)	5 (5.7)	10 (3.9)	0.54 [d]
Lived with pets/livestock, n (%)	64 (73.6)	177 (69.1)	0.44 [b]
Antibiotics in previous 6 months, n (%)	6 (6.9) [e]	16 (6.3)	0.83 [b]
Admitted to hospital in past year, n (%)	1 (1.1)	6 (2.3)	0.68 [d]
Worked in a hospital in past year, n (%)	0 (0)	8 (3.1)	1.0 [d]
Lived with someone who was hospitalized/worked in hospital in past year, n (%)	1 (1.1)	69 (27.0)	0.47 [d]
Known/suspected SSTI/MRSA infection in past year, n (%)	2 (2.3)	2 (0.8)	0.27 [b]
Prior contact with MRSA SSTI in 3 months prior to start of training, n (%)	1 (1.1)	4 (1.6)	1.00 [b]

481 Note, all trainees on base were male during the study year. SSTI = skin and soft tissue infection ; MRSA =
482 methicillin-resistant *S. aureus*. [a] T-test with unequal standard deviations. [b] Chi-square test with 1 df. [c] Includes 28
483 individuals who self-reported as "Hispanic" and selected race "Other". [d] Fisher's Exact test. [e] One individual with
484 prior antibiotic exposure had an SSTI; none of those with SSTI had hospitalization, prior SSTI themselves or contact
485 with someone with SSTI in the three months prior to arrival to Fort Benning.
486
487
488

21

489   **Table 2. Genomic characteristics of included isolates (N=244)**

490

In silico prediction	Colonization, n=232	Clinical, n=12
MLST, n (%)		
ST8	123 (53.0)	12 (100)
ST5	81 [a] (34.9)	0 (0)
ST87	21 (9.1)	0 (0)
Other STs [b]	4 (1.7)	0 (0)
SLV of ST8	1 (0.4)	0 (0)
Novel STs [c]	2 (0.9)	0 (0)
lukS/lukF-positive, n (%)	108 (46.6)	12 (100)

491   'MLST' refers to multi-locus sequence type, while 'SLV' refers to a single locus variant, wherein one allele differs
492   from the listed ST. Three colonization samples were PCR-positive for *mecA* but negative *in silico:* 1133_2P,
493   1134_5N, 2069_2I. [a] This includes isolate 2014_3P which has two alleles of the *tpi* gene (1,99). [b] This includes
494   ST15, ST30, ST45, ST72. [c] Includes a triple locus variant of ST8.

495

496

497 **Table 3. Pairwise core single nucleotide polymorphism (SNP) distances within sub-groups of Clusters 1, 2 and 4**
498

Cluster	Sub-group	Number of isolates	Number of trainees	SNP distance between trainees, median (range)	SNP distance within trainees, median (range)	Number of trainees with positive at baseline, n (%)	Number of trainees with nasal samples, n (%)	Number of trainees with clinical samples, n (%)
1	1-1	69	24	7 (0-383)	0 (0-9)	4 (16.7)	12 (50.0)	6 (25.0)
	1-2	16	8	4 (0-9)	0 (0-4)	0 (0)	3 (37.5)	-
	1-3	7	2	2 (1-3)	0 (0-1)	1 (50.0)	1 (50.0)	-
	1-4	11	5	1 (0-3)	0 (0-2)	0 (0)	2 (40.0)	3 (60.0)
	1-5	4	2	159 (157-161)	3 (2-4)	1 (50.0)	0 (0)	-
	1-6	8	3	21 (0-22)	1 (0-7)	1 (33.3)	1 (33.3)	2 (66.7)
	1-7	14	2	0 (0-8)	1 (0-10)	1 (50.0)	1 (50.0)	1 (50.0)
	1-8	6	3	0 (0-2)	1.5 (0-3)	3 (100)	0 (0)	-
	1-9	4	1	-	8.5 (2-15)	1 (100)	0 (0)	-
2	2-1	9	5	4 (1-116)	0 (0-3)	1 (100)	2 (40.0)	-
	2-2	51	23	1 (0-655)	1 (0-3)	3 (13.0)	7 (30.4)	-
	2-3	11	2	0 (0-4)	1 (0-5)	1 (50.0)	1 (50.0)	-
	2-4	3	1	-	1 (0-1)	1 (100)	0 (0)	-
	2-5	2	1	-	0 (0-0)	1 (100)	0 (0)	-
	2-6	5	1	-	0 (0-0)	1 (100)	0 (0)	-
4	4-1	12	6	0 (0-2)	1 (0-3)	6 (100)	1 (16.7)	-
	4-2	9	5	0 (0-0)	0 (0-0)	0 (0)	4 (80.0)	-

499 Clusters were identified in the analysis including all samples, using hierBAPS [16]. Core SNPs within each Cluster were then identified by repeating the analysis,
500 but aligning the reads from isolates within each Cluster to a closely-related reference from the most represented sequence type in that Cluster. hierBAPS [16] was
501 then run again for each Cluster using these new alignments as input (to obtain the highest resolution), and sub-groups of closely-related isolates were identified
502 within each Cluster. The number of core SNPs were tabulated within each sub-group and are comparable within, but not across, Clusters (as each Cluster has a
503 different core genome). Cluster 1 sub-groups were separated by a minimum of 34 core SNPs, Cluster 2 sub-groups were separated by a minimum of 150 core
504 SNPs, and Cluster 4 sub-groups were separated by a minimum of 100 core SNPs. Cluster 1 was predominantly ST8 (n=136/139); three non-ST8 isolates were
505 included based on hierBAPS analysis, including 1133_2P (a single locus variant of ST8), 1105_5P (a triple locus variant of ST8) and 1133_2P (ST72, a triple
506 locus variant of ST8). These three non-ST8 isolates were included in sub-group 1-1, and were a maximum of 383 core SNPs to another other sample within this
507 sub-group. Cluster 2 was comprised of all ST5 isolates, while Cluster 4 was comprised of all ST87 isolates.

508   **Figure legends.**

509

510   **Figure 1. Number of isolates per trainee, by anatomical site of isolation.** 85 trainees had at

511   least one non-clinical sample positive for MRSA during the study. Includes samples that were

512   not available/sequenced.

513

514   **Figure 2. Timing of first MRSA positive samples collected from each trainee, by company**

515   **at Fort Benning, Georgia, 2015.** In total, 87 trainees were positive for MRSA from the two

516   Companies under study. 27 trainees were positive at baseline; these samples were collected

517   within 1-2 days of arrival to base, while all trainees were at one central location.

518

519

520
521 **Figure 3. Maximum likelihood tree for colonization and clinical samples from trainees**

522 **undergoing basic training at Fort Benning, Georgia in 2015.** One isolate (1134_5N) was

523 highly divergent from all others, at > 18,500 core SNPs compared to < 1000 SNPs for the next

524 most divergent. Therefore, this isolate was excluded from this analysis. A concatenated SNP

525 alignment of the remaining isolates was produced using snp-sites [28] and IQ-Tree [14] (v1.6.8)

526 was used to produce the maximum likelihood tree, using the TVMe model with ascertainment

527 bias correction. Model selection was based on the lowest Bayesian Information Criterion.

528 Hierarchical Bayesian Analysis of Population Structure (hierBAPS) [16, 29] was used to identify

529 clusters with the full alignment as input. The tree was visualized using the Interactive Tree of

530 Life [30]. Samples from skin and soft tissue infection are indicated.

531

532

533

534   **Figure 4. Recombination-adjusted core pairwise SNP distributions by company and**

535   **platoon.**

536   **A. Pairwise core SNPs across and within companies.**

537   **B. Pairwise core SNPs across and within platoons.**

538   Core SNPs were identified compared to the USA300 reference (excluding isolate 1134_5N due

539   to its high divergence from all others isolates in the dataset). This resulted in 29,403 pairwise

540   comparisons between trainees, including repeat samples from the same individuals. Retaining

541   closely-related samples from the same trainees would be expected to artificially reduce the

542   pairwise distances within companies and platoons. Therefore, where repeat samples from a

543   single participant were within 5 SNPs of one another, only one of these was retained for this

544   analysis. This threshold was chosen because it represented 90% of pairs from the same trainees.

545   In total, 3,739 pairwise comparisons were therefore included in this analysis. Note that all

546   trainees in the same platoon are, by default, in the same company.

547

548

549   **Figure 5. In-depth analysis of Cluster-1 and potential transmission within and across**

550   **platoons.** There were 139 isolates in Cluster-1, as classified by hierBAPs [16, 29] in the overall

551   analysis (Figure 3), 136 of which were ST8. A concatenated core SNP alignment was produced

552   with these isolates. A maximum likelihood tree was produced using IQ-Tree [14] (v1.6.8) and

553   visualized using ggtree [31] in R (v.3.5.2). hierBAPS was again used to identify sub-groups of

554   closely-related isolates within this Cluster, which are indicated in colour. Note that this analysis

555   includes repeat isolates from all trainees; sample names are indicated using the format

556   *Participant Identifier + Visit Number + Anatomical Site of Isolation* (where N=nasal, I=Inguinal,

557   P=Perianal and O=Oropharyngeal), such that repeat samples can be identified. The heatmap

558   indicates epidemiological contact between trainees; the letters (A and B) represent the company

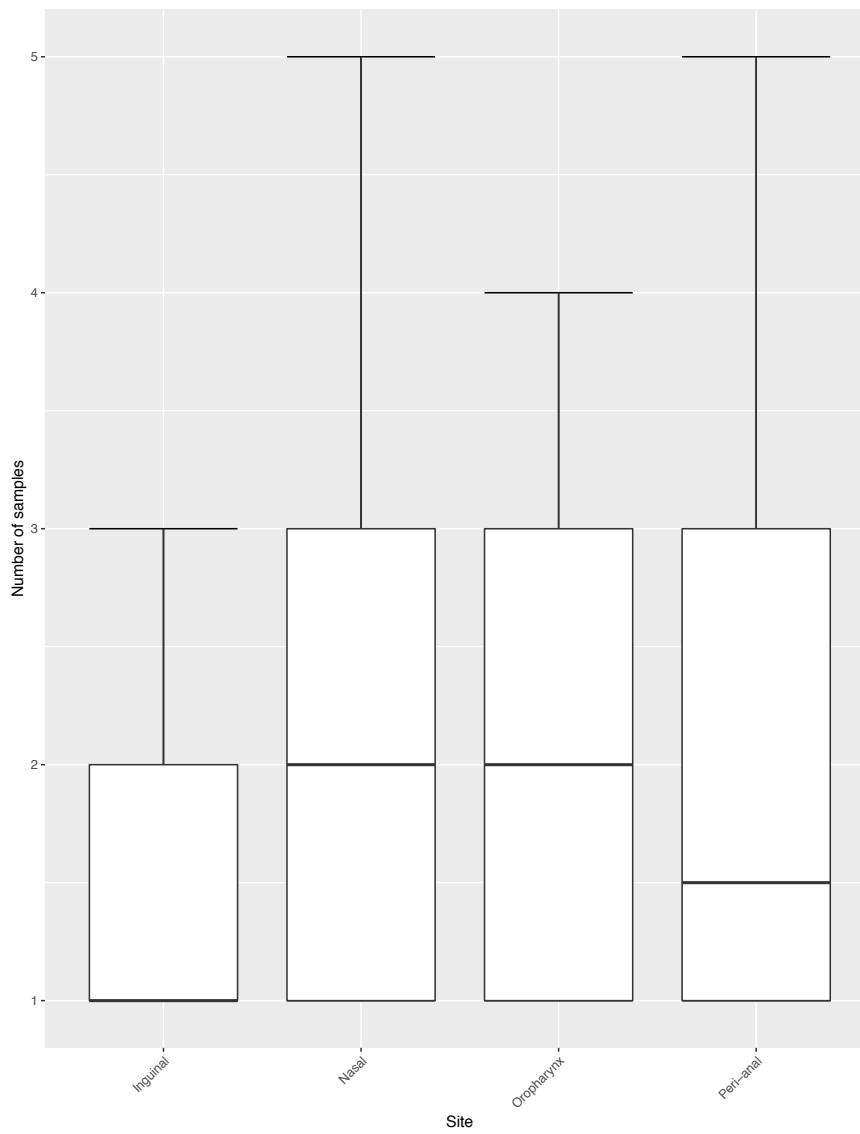559   and the numbers represent the specific platoon within that company.

560

561

28
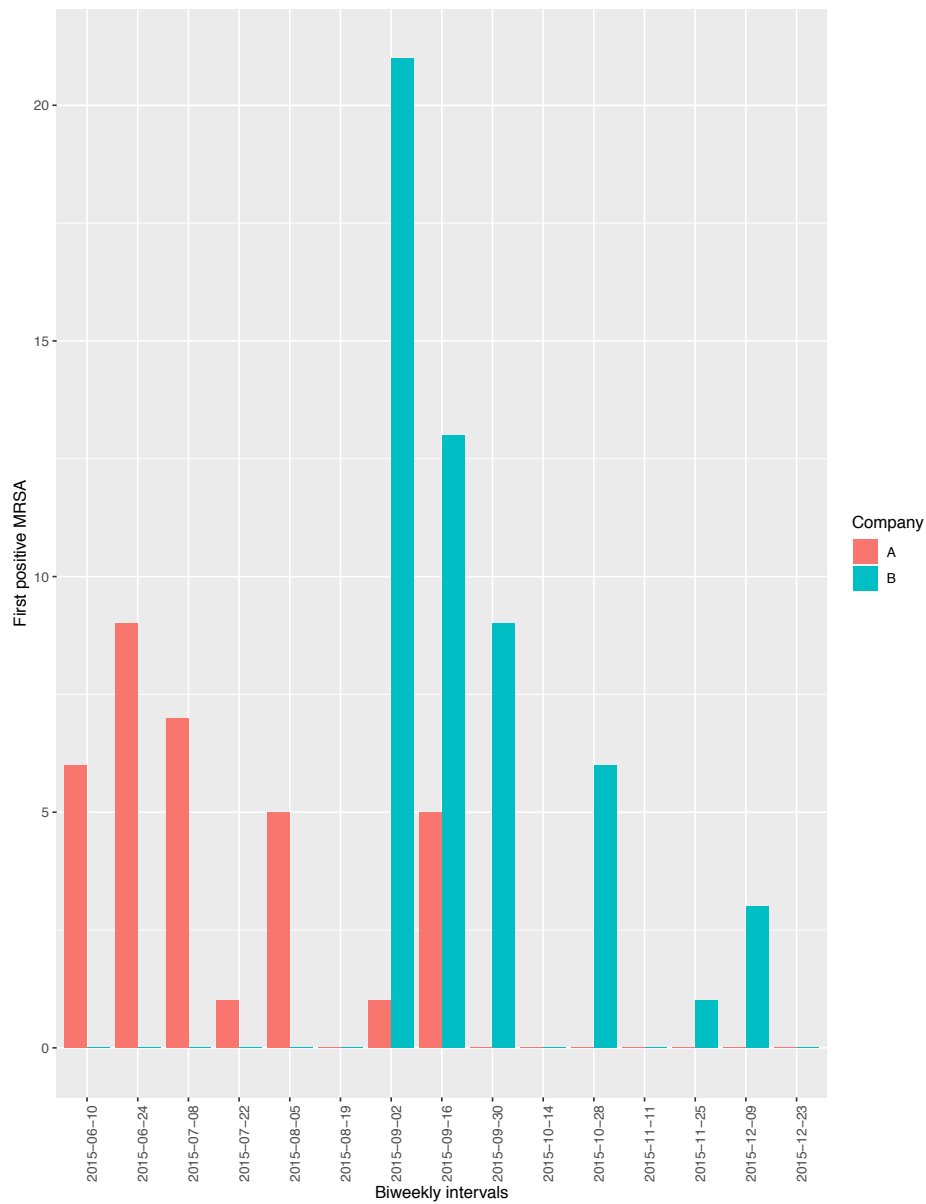
Figure 1.

Figure 2.

Figure 3.

Figure 4.

A.

B.

Figure 5.