Popovich et al. Supplementary Materials

Supplemental Methods

Single nucleotide variant identification

Variants were identified by (i) mapping filtered reads to the assembled USA300 reference genome (GenBank accession number NC_002952) using the Burrows-Wheeler short-read aligner, (ii) discarding polymerase chain reaction duplicates with Picard, and (iii) calling variants with SAMtools and bcftools. Variants were filtered from raw results using GATK's VariantFiltration (QUAL, >100; MQ, >50; >=10 reads supporting variant; and FQ, <0.025). In addition, a custom python script was used to filter out single-nucleotide variants that were: (i) <5 base pairs (bp) in proximity to indels that were identified by GATK HaplotypeCaller, (ii) in a recombinant region identified by Gubbins[1], in a phage region identified by Phaster[2] or (iii) they resided in tandem repeats of length greater than 20bp as determined using the exact-tandem program in MUMmer[3]. Note that while Gubbins filtered variants were used for phylogenetic analyses, figures showing pairwise genetic distances are without applying Gubbins filtering so as to prevent inter-clade recombination events from too greatly reducing the core genome size. Intra-clade recombination was minimal, and we confirmed that this decision did not affect any conclusions drawn from these analyses.

Phylogenetic analysis

The alleles at each position that passed filtering in all genomes were concatenated to generate a non-core variant alignment relative to the USA300 reference genome (non-core alignment = 2.87 Mbp). Variant positions in the non-core alignment were used to

reconstruct a maximum likelihood phylogeny with IQTREE v1.5.5[4] using ultrafast bootstrap with 1000 replicates (-bb 1000). ModelFinder limited to ascertainment bias-corrected models (-m MFP-ASC) was used to identify the best model (transversion model with equal base frequency and two rate categories – TVMe + R2) based on Bayesian Information Criterion (BIC)[5].

Selection of representative isolates from each individual

Pairwise single nucleotide variant (SNV) distances were computed by only considering positions present in all sequenced genomes (core genome length = 2.41 Mbp). MRSA genomes from our study were classified into sequence types based on their SNV distances to known reference genomes (USA300, USA500, USA100/ST5, Figure S4) and based on classification scheme outlined by Bowers et al.[6] For patient-level analyses, a single representative isolate was selected for sets of patient isolates deemed to represent a single acquisition. For each individual, we found the maximum sized subtrees that contained all of the isolates from that individual. If all of the isolates from a single individual was present in a single subtree with no other individual, the earliest isolate was chosen as the representative isolate. If isolates from other individual(s) were included in the subtree, a SNV threshold of 40 was used to determine representative isolates from that individual (See Figures S2 and S3). Five individuals had more than one representative isolate. One individual was colonized with two distinct strains simultaneously. Four individuals entered the jail twice and were colonized with at least one distinct strain compared to their previous incarceration. Fisher's exact tests were performed to evaluate enrichment for patient characteristics within each sequence type.

Supplemental references

1. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic acids research. **2015**;43(3):e15.

2. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research. **2016**;44(W1):W16-21.

3. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. Current protocols in bioinformatics. Feb **2003**;Chapter 10:Unit 10 13.

4. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. **2018**;35(2):518-522.

5. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods. **2017**; 14(6):587-589.

6. Bowers JR, Driebe EM, Albrecht V, et al. Improved Subtyping of Staphylococcus aureus Clonal Complex 8 Strains Based on Whole-Genome Phylogenetic Analysis. mSphere **2018**; 3:e00464-17.

7. Frisch, M. B., S. Castillo-Ramírez, R. A. Petit, M. M. Farley, S. M. Ray, V. S. Albrecht, B. M. Limbago, et al. Invasive Methicillin-Resistant Staphylococcus Aureus USA500 Strains from the U.S. Emerging Infections Program Constitute Three Geographically Distinct Lineages. mSphere 3, no. 3: e00571-17.

Supplemental Table

Table S1 – Summary of sequenced MRSA isolates

Table S2- Epidemiologic factors associated with ST5, USA300, and USA500 MRSA

Supplemental Figures

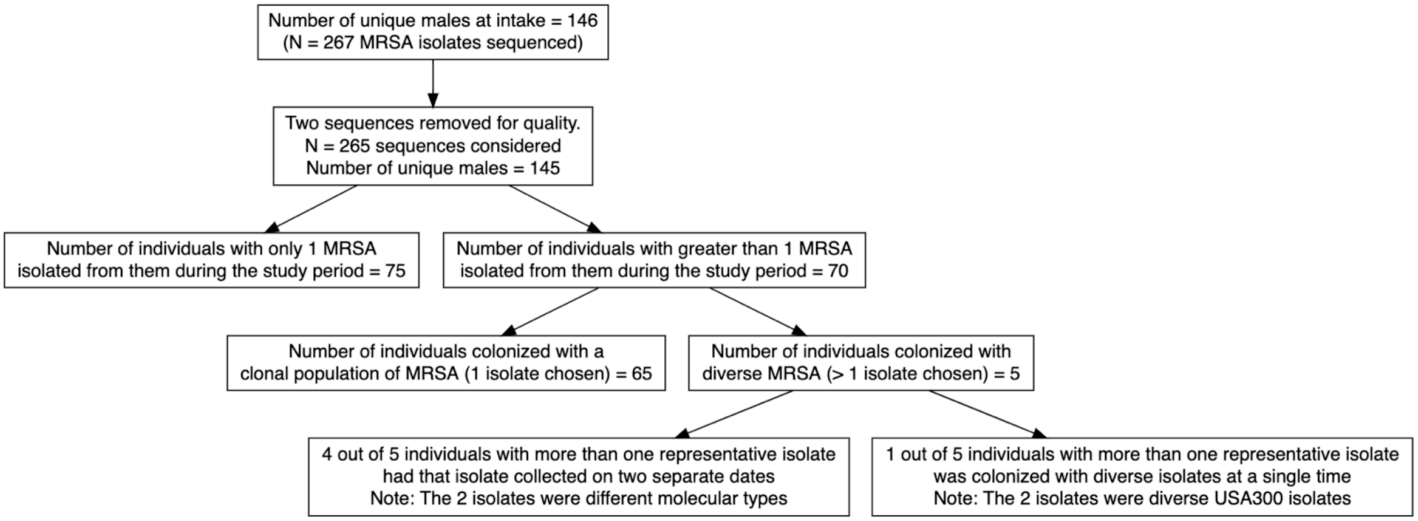

Figure S1 – Flow chart of MRSA isolates sequenced in current study.

* For univariate analysis in the main text only an individual's first incarceration was considered for epidemiologic analysis (n=137). For genomic sequencing, an individual's first MRSA isolate, regardless of whether it was from their first or a subsequent incarceration, were included and therefore the number for unique events was 146. Some enrolled individuals were colonized at more than one body site and these sites were sequenced and therefore, 267 MRSA isolates were sequenced.

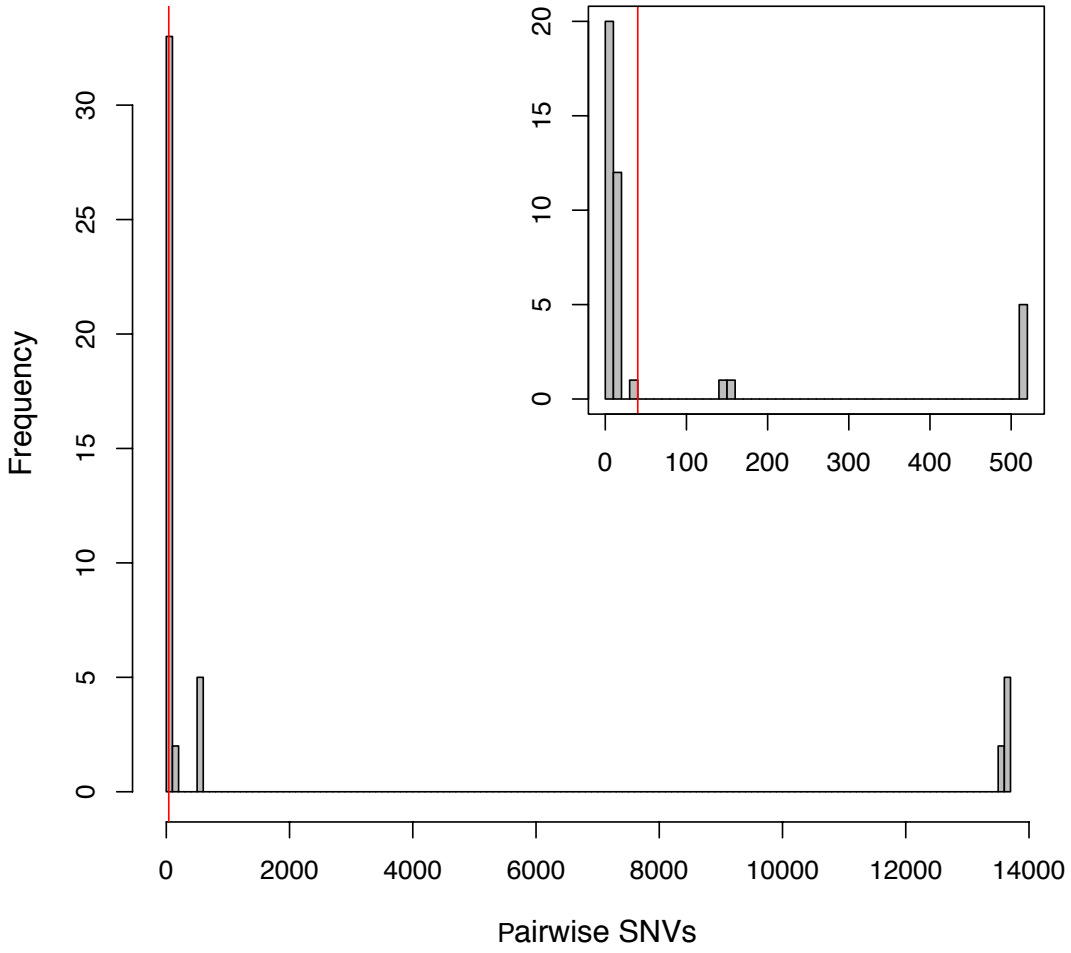

Figure S2 – Selecting representative strains from each individual. Seventy individuals were colonized with more than one MRSA isolate (either colonized across multiple body sites, entered the jail colonized more than once during the study, or a combination of the two). For the analysis of epidemiologic factors associated with acquiring different strain types (**Table 3**, **Table S2** and **Table S3**), we wanted to include individuals more than once if they showed genomic evidence of acquiring MRSA multiple times. If all of an individual's isolates were contained in a monophyletic sub-tree, then the isolates were presumed to represent genetic variation that accumulated during colonization of that individual. If an individual's isolates were inter-mixed with isolates from other individuals, then we used an SNV cutoff to determine if more than one isolate should be chosen to represent multiple independent acquisitions of MRSA. For those individuals, we calculated the SNV distances among all pairs of isolates from an individual (shown as histogram above). Based on the distribution of SNVs within those individuals, a cutoff of 40 was chosen (see red line on histogram). SNVs were determined relative to a core genome of size 2.42 Mb.

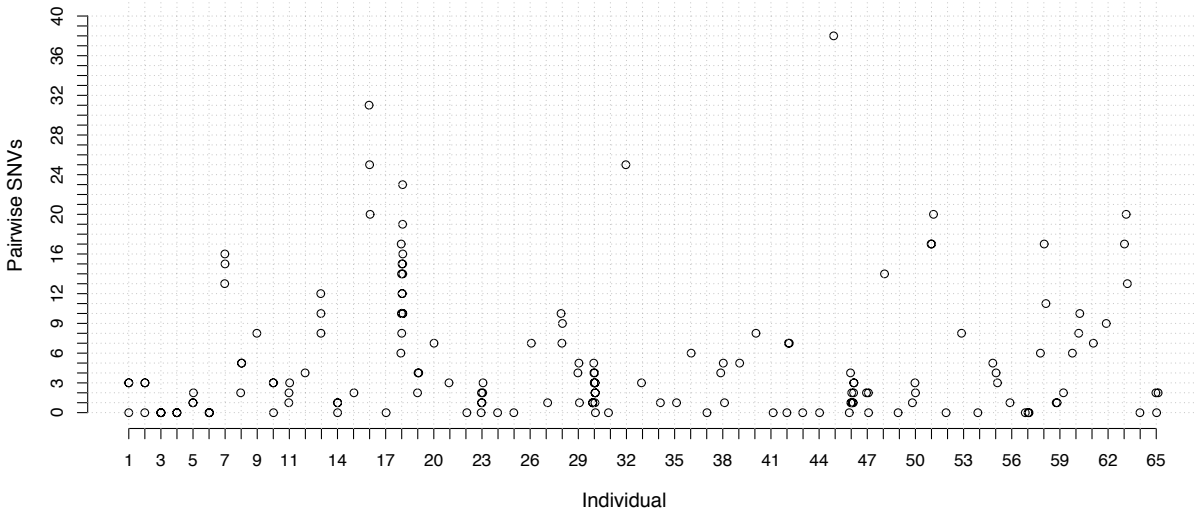

Figure S3 – SNV distance among MRSA isolates from individuals. The x-axis shows the 65 individuals who had multiple sequenced isolates for which between-isolate genetic variation was plausibly the result of intra-host evolution (See **Figure S2**). The y-axis shows the number of SNVs between each intra-patient isolate pair, with each point representing one pair of isolates. SNVs were determined relative to a core genome of size 2.42 Mb.

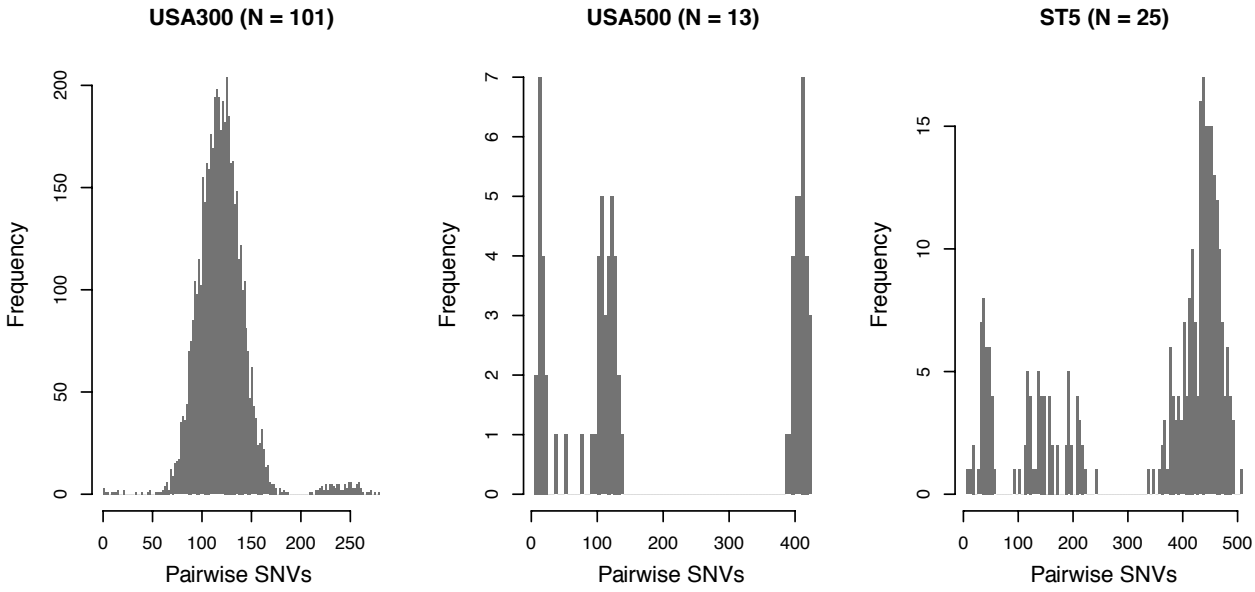

Figure S4 – Within strain variation among in silico typed USA300, USA500 and ST5. Histograms of pairwise SNV distance among strains designated USA300, USA500 or ST5 are shown in the three panels. USA500 isolates fall into one of two clades (USA500-C1 and USA500-C2, see **Figure S5**), which accounts for the bimodal distribution in the USA500 pairwise SNV distribution. SNVs were determined based on a core genome of size 2.42 Mb.

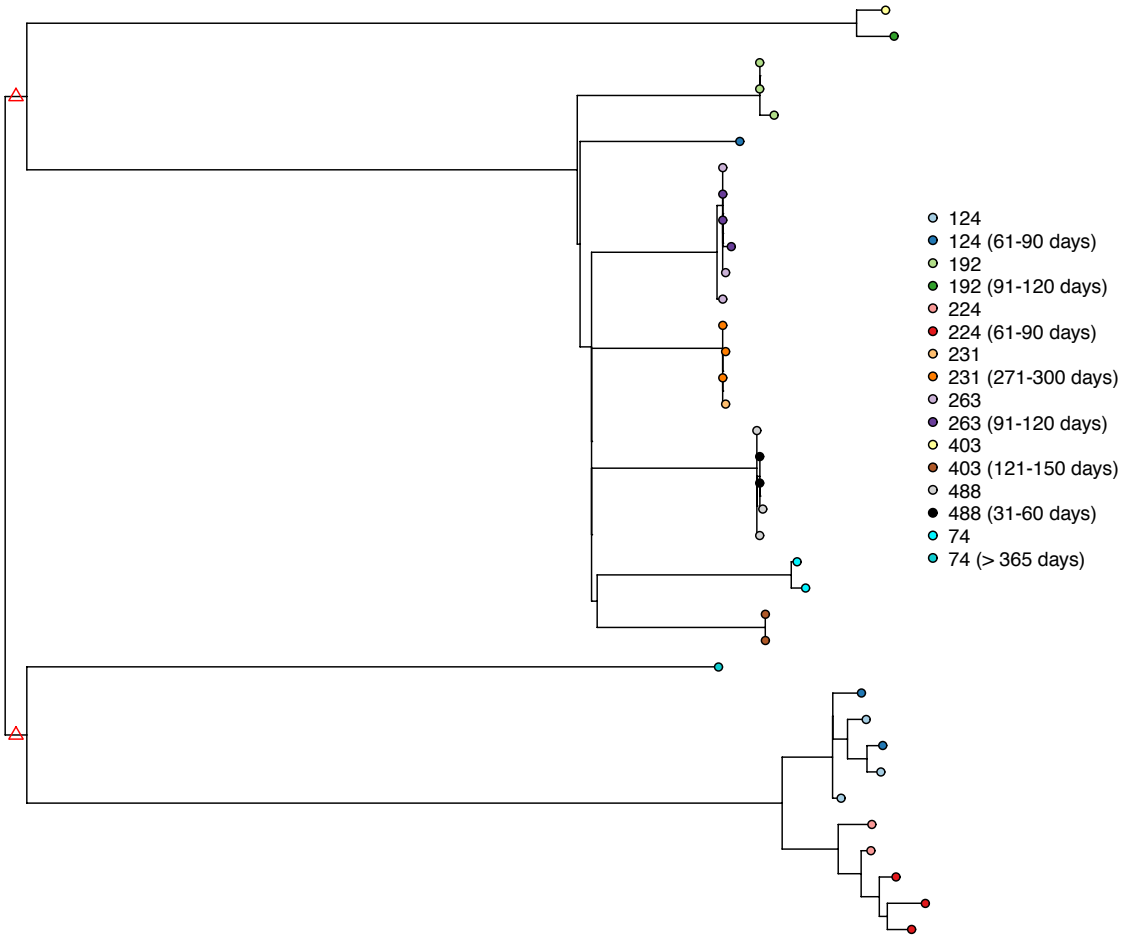

Figure S5. Phylogenetic analysis of individuals colonized with MRSA on multiple jail admissions. A maximum likelihood tree was constructed for all study isolates and reduced to display isolates only from individuals who were admitted to jail multiple times during the study and colonized with MRSA at intake for at least two of those admissions. Each of the eight individuals is represented by a different color, with the two shades of the color representing isolates from their two admissions. The number of days between an individual's admissions is noted in parentheses next to their second admission. Triangles on internal branches are used to represent that long inter-clade branches were collapsed in the figure for visualization purposes.

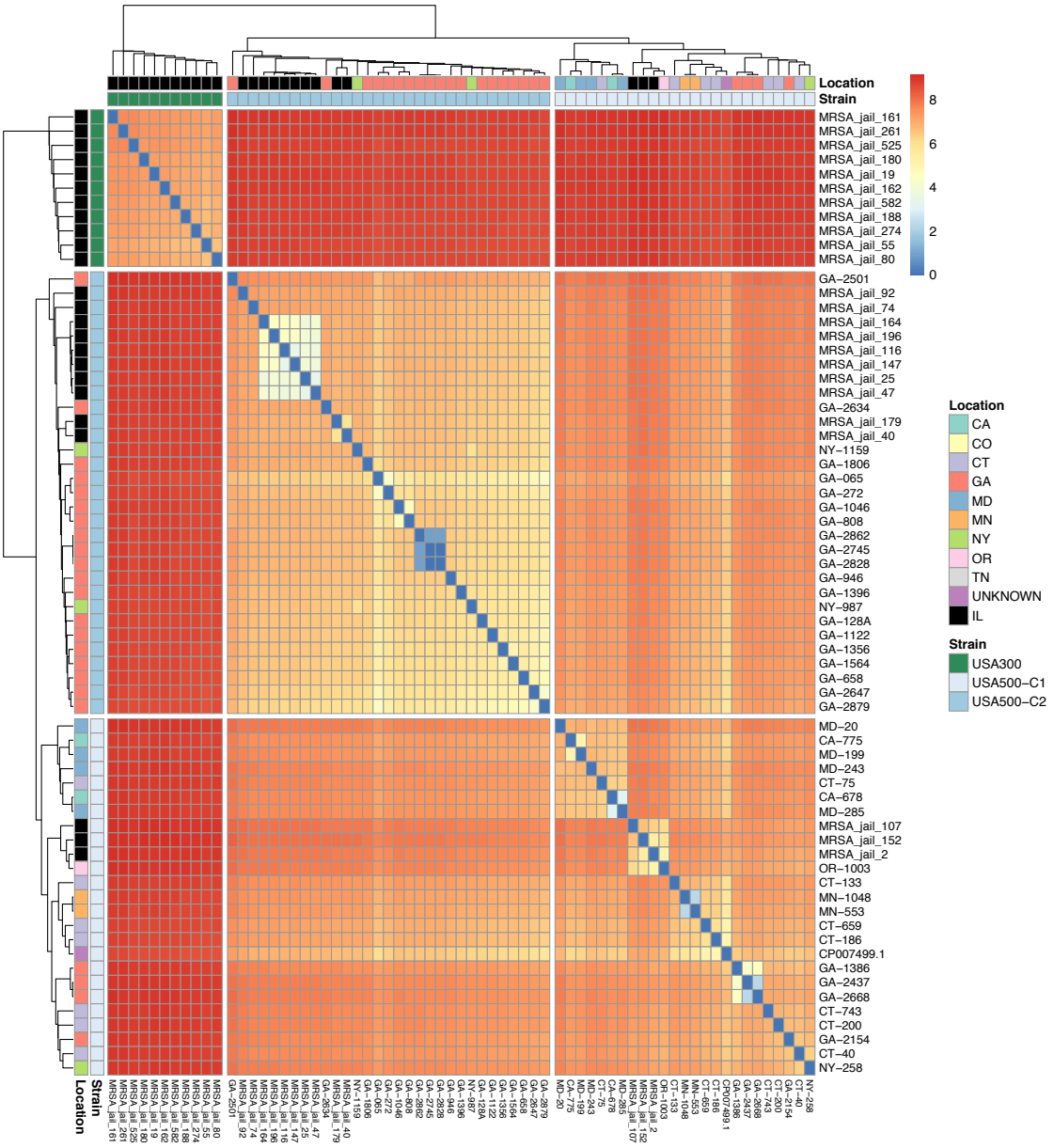

Figure S6 – Genomic comparison of USA500 strains isolated from individuals entering the Cook County Jail relative to USA500 strains isolated from individuals in other US Cities. Isolates from the current study were contextualized with genomes from a previous study that identified three sub-lineages of USA500 (C1, C2, and E1– Frisch et al., Msphere, 2018[7]). MRSA isolates from Illinois are the isolates from the current study (IL - black), with other geographic designations taken from the Frisch et al. manuscript[7]. Colors in the heatmap represent log2 SNV distance between pairs of MRSA isolates. Isolates were grouped via hierarchical clustering. The heatmap was divided into 3 clusters using an SNV cutoff to more clearly separate USA500-C1 (lighter blue), USA500-C2 (darker blue), and USA300 (green). Two isolates in our study were classified as CC8e (likely E1) by Bowers et al. sequence-based classification[6] but E1 strains were not included in this heatmap.