Journal of Infectious Diseases Advance Access published June 5, 2015

Parallel Epidemics of Community-Associated Methicillin-Resistant *Staphylococcus aureus* USA300 in North and South America

Paul J. Planet[1,2,*], Lorena Diaz[3], Sergios-Orestis Kolokotronis[2,4], Apurva Narechania[2], Jinnethe Reyes[3], Galen Xing[1], Sandra Rincon[3], Hannah Smith[1], Diana Panesso[3,5], Chanelle Ryan[1], Dylan P. Smith[3], Manuel Guzman[6], Jeannete Zurita[7], Robert Sebra[8], Gintaras Deikus[8], Rathel L. Nolan[9], Fred C. Tenover[10], George M. Weinstock[11], D. Ashley Robinson[9] and Cesar A. Arias[3,5,*]

[1]Department of Pediatrics, Division of Pediatric Infectious Diseases, Columbia University, College of Physicians and Surgeons, New York, NY, USA

[2]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA

[3]International Center for Microbial Genomics, Molecular Genetics and Antimicrobial Resistance Unit, Universidad El Bosque, Bogotá, Colombia

[4]Department of Biological Sciences, Fordham University, NY, USA

[5]Department of Internal Medicine, Division of Infectious Diseases and Department of Microbiology and Molecular Genetics, University of Texas Medical School at Houston, TX, USA

[6]Centro Médico Caracas, Caracas, Venezuela

[7]Hospital Vozandes, Pontificia Universidad Catolica, Quito, Ecuador

[8]Genome Center, Mount Sinai Hospital, 1425 Madison Avenue, New York, NY 10029, USA

[9]Department of Internal Medicine, Division of Infectious Diseases and Department of Microbiology and Division of Infectious Diseases, University of Mississippi Medical Center, Jackson, MS, USA

[10]Cepheid, Sunnyvale, CA, USA

[11]The Jackson Laboratory for Genomic Medicine, CT, USA

[*]Corresponding authors: Cesar A. Arias, MD PhD, Division of Infectious Diseases, University of Texas Medical School at Houston, 6431 Fannin St. MSB 2.112. Houston, TX 77030, Tel: 713.500.6738, Fax: 713.500.5495, e-mail: caa22@cantab.net, cesar.arias@uth.tmc.edu, Paul J. Planet MD, PhD, Pediatric Infectious Disease Division, Columbia College of Physicians and Surgeons, 630 West 168th Street, New York, NY 10032, Tel: 212-305-1296, Fax: 212 342-5218, Email: pjp23@columbia.edu

Abstract

Background

The community-associated methicillin-resistant *Staphylococcus aureus* (CA-MRSA) epidemic in the United States is attributed to the spread of the USA300 clone. An epidemic of CA-MRSA closely related to USA300 has occurred in northern South America (USA300 Latin-American variant, USA300-LV). Using phylogenomic analysis we aimed to understand the relationships between these two epidemics.

Methods

We sequenced the genomes of 51 MRSA clinical isolates collected between 1999 and 2012 from United States, Colombia, Venezuela and Ecuador. Phylogenetic analysis was used to infer the relationships and times since the divergence of the major clades.

Results

Phylogenetic analyses revealed two dominant clades that segregated by geographical region with a putative common ancestor in 1975, individually originating in 1989 and 1985 for North American and South American clades, respectively. Emergence of these parallel epidemics coincides with the independent acquisition of the arginine catabolic mobile element (ACME) in North American isolates and a novel copper and mercury resistance (COMER) mobile element in South American isolates.

Conclusions

Our results reveal the existence of two parallel USA300 epidemics that shared a recent common ancestor. The simultaneous rapid dissemination of these two epidemic clades suggests the presence of shared, potentially convergent, adaptations that enhance fitness and ability to spread.

Background

Staphylococcus aureus is a major human pathogen causing life-threatening infection both in the hospital and community. In the last few decades, acquisition of several antibiotic resistance determinants by *S. aureus* has made the treatment of such infections an important clinical challenge. Most notably, the emergence and dissemination of methicillin-resistant *Staphylococcus aureus* (MRSA) throughout the world appears to follow an epidemic pattern with MRSA clones disseminating in specific geographical areas. Prior to the 1990’s most MRSA was associated with hospital settings. Since that time, community-associated MRSA (CA-MRSA) has grown increasingly frequent in the United States [1,2], an epidemiological change that coincides with an overall increase of skin and soft tissue infections (SSTI) and *S. aureus*-related hospitalizations [3,4]. In North America, this CA-MRSA epidemic is widely attributed to the spread of a clone designated USA300 [5].

The earliest report of the USA300 clone is from an outbreak of infections starting in November 1999 in a state prison in Mississippi in the United States [6]. Within a 3-year period, additional outbreaks were documented from several other States including California, Colorado, Georgia, Pennsylvania, and Texas, where USA300 appeared in community settings such as correctional facilities, military bases, daycare centers and sports teams. USA300 rapidly became more widespread in the community and, by 2004, it had become the major cause of SSTIs in the U.S. [1]. USA300 also causes other life-threating diseases such as community-acquired pneumonia, osteomyelitis, and blood stream infections [8,9]. The spread of USA300 has been linked to large increases in hospitalizations for severe skin and soft tissue infection from 2000 to 2009 [4,7] and its replacement of other, possibly less virulent *S. aureus* clones [10].

First identified in 2005 [11,12], a USA300 Latin American variant (USA300-LV) has spread through community and hospital settings in Colombia, Venezuela and Ecuador replacing the prevalent, but unrelated, hospital-associated MRSA designated as the Chilean/Cordobes clone [12,13]. USA300-LV appears to cause the same spectrum of disease as USA300 from North America, and it has become the most prevalent CA-MRSA strain in *S. aureus* infections in northern South America [13]. Isolates belonging to the USA300-LV clone are close relatives of North American USA300 based on standard molecular typing techniques, and they possess some of the key genetic signatures of the USA300 lineage including a pathogenicity island that encodes the enterotoxin genes *sek* and *seq* (SaPI5) and the genes for Panton-Valentine Leukocidin (PVL), *lukSF-PV* [12,13]. USA300-LV isolates differ from North American USA300 isolates in that they lack a genomic locus called the arginine catabolic mobile element (ACME) [14,15], which is thought to be an important determinant for the success of USA300 [16-18]. Most USA300-LV isolates also harbor a different variant of the methicillin resistance cassette (SCC*mec* IVc-E) [12,13,19,20].

Since USA300-LV infections first were characterized 6 years after the initial description of USA300 isolates in North America, it has been assumed that South American CA-MRSA strains likely disseminated southwards from a North American origin. In this study, we aimed to delineate the specific genetic relationships between USA300 and USA300-LV using whole genome sequencing in order to provide insights into the epidemiology and evolution of the CA-MRSA epidemics in the Americas. We show that the CA-MRSA South American epidemic is not an extension of the North American epidemic of USA300, but rather occurred simultaneously with the two lineages sharing a common ancestor prior to their epidemic spread.

Methods

Bacterial strains

MRSA isolates from South America were collected from a surveillance study performed in tertiary hospitals in Colombia, Ecuador, and Venezuela between 2006-2007 [13]. We also included the first two characterized USA300-LV strains isolated in 2005 [11,12] and the recently reported vancomycin-susceptible and resistant MRSA strains recovered from a patient in Brazil, also related to the USA300 lineage [21]. Isolates from the US were collected from multiple US States from 1999 to 2012. Identification of all MRSA isolates, antimicrobial susceptibility profiles, SCC*mec* typing, pulsed-field gel electrophoresis (PFGE) and amplification of genes encoding Panton-Valentine Leucocidin (PVL) were performed as described before [13,21]. All bacterial details of isolation are shown in Table S1.

Genome Sequencing

For the Illumina platform, genomic DNA was prepared using either the DNeasy Blood & Tissue kit (Qiagen) or the Wizard Genomic DNA Purification Kit (Promega) after lysostaphin treatment. Genomic DNA libraries were prepped using the NexteraXT DNA Sample Preparation kit and sequenced on a MiSeq Desktop Sequencer (Illumina) with 250bp paired-end reads. Genome assembly was done using the paired-end implementation of ABySS [22] and CLCGenomics Workbench (CLCBio, Aarhus, Denmark) v8.1. Genbank accession codes are shown in the supplementary table S1.

For the PacBio platform, genomic DNA was prepared from concentrated overnight cultures treated with lysostaphin. The Genomic DNA tips 500/G and Genomic Buffer Set (Qiagen) was then used for initial preparation of the DNA. Approximately 3.5 µg of DNA was used to construct SMRTBell libraries for the PacBio RS II DNA sequencing system (Pacific Biosciences) using polymerase enzyme-DNA bound complexes with an average insert size of ~20 kb. The binding chemistry was done using the PacBio P5-C3 DNA/Polymerase binding kit. DNA/Polymerase complex of the sample was prepared using 0.5 nM of the SMRTbell library and 10X excess DNA polymerase. All 20 kb samples were MagBead-loaded prior to immobilization on SMRTcells. The PacBio RS II DNA sequencing system used 180 min continuous collection times and C3 sequencing chemistry, allowing collection of subreads of up to ~36,000 bp. We used the HGAP2 v2.1 de novo assembly pipeline [23]. Genbank accession codes are shown in the supplementary table S2.

Antibiotic resistance genes (resistome)

The presence of the most frequent genetic resistance determinants present in MRSA was evaluated using the BLASTn tool from NCBI (e-value =0, identity >98%, alignment coverage >95%). Protein alignments were performed using ClustalW for GyrA and ParC. Accession numbers of the query genes are provided in the Table S3.

Single nucleotide polymorphisms calling and phylogenetic matrix construction

Comparison of single nucleotide polymorphisms (SNP) between isolates was done using short read alignment to the genome of USA300 strain TCH1516 as reference using the Burrows-Wheeler Alignment (bwa) tool (http://bio-bwa.sourceforge.net). SNP calls were made using samtools (http://samtools.sourceforge.net). SNPs were identified as “high quality" if they were not listed as "heterozygous" and had a per base Q score greater than or equal to 20. For preassembled genomes from public databases, we used whole-genome alignment with reference to the *S. aureus* TCH1516 genome using the show-snps utility of NUCmer (http://mummer.sourceforge.net). For Bayesian analysis, only SNPs defined by the samtools approach were used. All regions from the reference genome annotated as mobile genetic elements were excluded (Table S4). We also applied a mask that excluded repetitive sequences from the reference genome that were more than 80% identical over at least 100 nucleotides to other genomic loci using a pairwise MegaBLAST-based analysis.

Establishment of gene orthology and presence/absence matrix

We determined orthologous gene sets from assembled isolate genomes using a modified version of the OrthologID pipeline [24] that now uses OrthoMCL for gene family clustering [25]. Presence and absence of genes was determined from the resulting concatenated alignment matrix of orthologs. Select genes were confirmed by BLAST (http://blast.ncbi.nlm.nih.gov).

Phylogenetic and time of divergence analyses

Maximum likelihood (ML) phylogenies were constructed with the POSIX-threads RAxML v8.0.19 [26]. We used an ascertainment bias correction and a general time-reversible (GTR) substitution model [27] accounting for among-site rate heterogeneity using the Γ distribution and four rate categories [28] (ASC_GTRGAMMA model) for 100 individual searches with maximum parsimony random-addition starting trees. Node support was evaluated with 1000 nonparametric bootstrap pseudoreplicates [29]. The number of bootstrap pseudoreplicates over which node support is not expected to be significantly altered was evaluated using the frequency-based (300 bootstrap iterations were sufficient) and extended majority-rule (550 bootstrap iterations were sufficient) bootstopping criteria [30]. The program SplitsTree4 was used to create network of bootstrap trees (Fig.1A). The program FigTree (http://tree.bio.ed.ac.uk/software/figtree/) was used to make Figure 1B.

The time of divergence of genetic lineages was estimated in a Bayesian framework in BEAST v1.7.5 [32] by calibrating the tree using the dates of isolation of the strains (Fig.S1) [33]. Since only SNPs were included in the alignment, we accounted for the lack of constant sites by editing the XML input file following recommendations made by the main BEAST developer (Andrew Rambaut, University of Edinburgh). Evolutionary rates along the tree were allowed to vary, as accommodated by the uncorrelated lognormal relaxed clock [34], and a constant-size coalescent tree process was assumed [35]. Each run of the Markov chain Monte Carlo (MCMC) procedure consisted of 300 million steps with a 5,000-step thinning. After inspection of the MCMC traces and the expected sample size (ESS) values of each run (ESS>200), 10% of the first posterior samples were removed as burn-in, and the posterior estimates from three individual runs were combined to form a posterior sample density based on 810 million samples. Departure from a strict molecular clock was confirmed by observing the posterior estimate of the coefficient of rate variation (mean=1.0587, 95% HPD: 0.7825-1.3718). The chronogram was plotted based on the maximum clade credibility tree. A median rate of 8.7201E-7 [HPD 95% 3.6595E-7, 1.4463E-6] was observed for this analysis (Fig.S3). Tree files for Bayesian and ML analysis are included as supplementary data.

Phylogeographic and gene presence/absence reconstruction

All phylogenetic reconstructions were done using parsimony and likelihood based techniques in the application Mesquite v.275 [36]. Likelihood reconstruction was based on the "Mk1 (est.)" model, a one-parameter Markov k-state model [37]. Parsimony reconstruction was based on Fitch Optimization procedure [38]. States of presence or absence were assigned “1” or “0” for each taxon. For phylogeography North America (0) and South America (1) were also coded as binary variables. In all instances the outgroup taxon character state was treated as missing data.

Whole genome alignment

We constructed whole genome multiple alignments of PacBio Assemblies using mummer (http://mummer.sourceforge.net) with each genome compared to the USA300 reference strain FPR3757, averaging the percent identity score for all nucmer alignments that overlap a given coordinate in the reference. Areas of no alignment are assigned a percent ID score of zero. We then used the alignments to plot these as scores in Circos (http://circos.ca).

Genealogy of the *copB* gene

We used the *copB* gene from USA300 FPR3757 (ABD21730.1) to search for closely related genes in Genbank (wgs and nr databases) using BLAST. Genes with greater than 99% nucleotide identity were included (Table S5). Nucleotide sequences were aligned using the Geneious 6.1.4 alignment algorithm with default settings. The phylogenetic tree was constructed using RAxML v8.0.19 with the GTR substitution model, among-site rate heterogeneity using the Γ distribution and four rate categories, and 100 individual searches with maximum parsimony random-addition starting trees. Node support was evaluated with 500 nonparametric bootstrap pseudoreplicates.

Results

To delineate the relationship between USA300 and USA300-LV, we sequenced the genomes of 51 MRSA isolates recovered from clinical samples between 1999 and 2012 from various locations throughout the United States, Colombia, Venezuela and Ecuador. We used a collection of the earliest known USA300-LV isolates that were recovered during 2005-2012 (n=25), and selected USA300 isolates (n=25) covering a wide geographic and temporal distribution within the United States, including the earliest-known, existing isolate of USA300 isolated from Mississippi in 1999. We included the first CA-MRSA isolate belonging to the USA300-LV lineage that was identified in Colombia (Bogota) in 2005 [11,12] and the recently reported vancomycin -susceptible and -resistant strains recovered from a patient in Brazil, also related to the USA300 lineage [39]. Initial comparative genomic analysis suggested that isolates belonging to USA300-LV are very closely related to USA300 isolates. The average number of high quality SNPs comparing South American genome sequences to the core genome of USA300 reference strain TCH1516 was 303 (standard deviation=47, range 265-385), whereas comparison of the USA500 strain 2395 (USA300's putative closest relative) [40] to TCH1516 revealed 533 SNP differences suggesting that USA300-LV are more closely related to North American USA300 isolates than USA500.

This close relationship was confirmed by phylogenetic analysis of 94,847 SNPs from 88 *S. aureus* genomes, including 80 clonal complex 8 (CC8) genomes (Fig.1A), and revealed two distinct sister clades that segregate by geographical region (Fig.1A). The genomic profiles of isolates from each of these clades match the profiles of the most prevalent epidemic strains in North and South America (for instance, ACME is only found in the North American clade) and antibiotic resistance gene profiles were also similar (Fig.S1). We designated these clades NAE and SAE, for North and South American epidemics, respectively. The range of isolation dates in the NAE clade (from 2001 to 2012) reflects almost the entire time frame of the epidemic. Therefore, it appears that the SAE and NAE clades diverged before the emergence of the USA300 epidemic in North America and, most importantly, that the SAE clade does not represent dissemination of the U.S. epidemic.

To further explore the timing of these events, we used a Bayesian-tip-calibration approach [34] to co-estimate the evolutionary rate of change and the times since the divergence of the major clades (Fig.1B). We estimated that the parallel epidemics in North and South America shared a common ancestor in approximately 1975 (HPD95%: 1932-1993) and individually originated in 1989 (HPD 95%: 1971-1999) and 1985 (HPD95%: 1962-1997) for NAE and SAE, respectively. Importantly, the credibility intervals on these dates place the origins of the two clades securely before the first reports of USA300 in North America (Fig.S2) and suggest that members of the USA300 clade have been present in the Western Hemisphere since the 1930's (median: 1933, HPD95%=1846-1977).

Interestingly, the early branches of the USA300 tree that diverge prior to NAE and SAE represent a mixed collection of isolates from both regions suggesting that the USA300 lineage was repeatedly transmitted throughout the Americas long before the current epidemics. Because of this branching pattern, phylogeographic reconstructions using both parsimony and likelihood approaches were unable to definitively assign a North or South

American origin to these early branch points (Fig.1B). Of interest, no early branching isolates or members of the SAE clade harbor the ACME locus (Fig.1). This absence strongly suggests that ACME was first acquired by an ancestor within the NAE clade. The estimated date, 1994 (HPD 95%: 1981-2000), at the node that represents the first ACME-containing ancestor (Fig.1B), is consistent with previous estimates of the acquisition of this element by horizontal gene transfer [18,41] and immediately precedes the North American epidemic.

To more fully characterize the genomic repertoires of the USA300-LV strains, we completed closed-circular genome sequences of five strains (V2200, M121, HUV05, CA15, CA12) chosen based on their phylogenetic position to capture the broadest phylogenetic diversity possible. These genomes are extremely similar to the TCH1516 reference both in sequence and gene order (Fig.2). The most divergent isolate within this sample is V2200 which harbors 431 SNPs compared to TCH1516, and is characterized by a different spa-type (t932) and pulsed gel electrophoresis (PFGE) type from other USA300-LV isolates [13].

Most of the major differences between the genomes derive from regions that are integration sites of mobile genomic elements (Fig.2). To search for whole gene differences that distinguish the NAE and SAE clades, we queried all of our sequences for genes that were exclusively found in the NAE and SAE clades. Several genes distinguished NAE and SAE individually from earlier branching strains. Most, but not all, members of these clades could be unambiguously classified by the combined presence or absence of these genes (Fig.1A). Only two genes were found that were exclusive to the combined NAE/SAE clade and found in more than 50% of the genomes; a gene encoding a putative ATPase copper exporter, *copB* (present in 83% [15/18] and 72% [13/18] of NAE and SAE genomes, respectively) and a putative lipoprotein (present in 83% [15/18] and 94% [17/18] of NAE and SAE genomes, respectively). Very similar genes are found in genomes of other staphylococcal species such as *S. epidermidis*, and are found sporadically in *S. aureus* genomes in variants of the SCC*mec* mobile element [42] and on plasmids [43]. In USA300, the two genes are located in a single genetic locus at one end of ACME (Fig.3A). In USA300-LV strains, the two genes are located in a novel region that takes the place of the ACME immediately adjacent to the SCC*mec* element. This locus also contains genes for an abortive phage infection system and mercury resistance. We refer to this novel locus as the COMER mobile element for its predicted function in copper and mercury resistance. Like ACME, phylogenetic analysis suggests that this locus is closely related to regions from *S. epidermidis* (Fig.3), but also suggests that the *copB* gene was acquired in two distinct events in the NAE and SAE clades (Fig.3B).

Discussion

The worldwide emergence and dissemination of CA-MRSA is an example of the ability of drug-resistant bacteria to establish a community reservoir and infect otherwise healthy individuals. In the United States, spread of MRSA strains belonging to the USA300 lineage occurred in a relatively short time, becoming an important public health problem [1]. Interestingly, apart from the USA, the only other part of the world where USA300-like strains have become endemic is the northern region of South America (specifically Colombia and Ecuador) [13]. Although genetically closely related to the prototypical USA300 strain [12], South American strains have distinct genetic characteristics including a variant of the SCC*mec* IV cassette and the absence of the ACME locus [12,13].

Since the report of the first patient infected with a CA-MRSA belonging to the USA300-LV lineage in 2005, dissemination and spread of this strain has occurred rapidly [13,20], initially leading to the assumption that the USA300 epidemic had spread from North to South [12]. However, our phylogenetic findings suggest that the CA-MRSA epidemic in South America began to disseminate independently of USA300 from North America. It is notable that by 2006 USA300-LV was already the predominant clone in Ecuador and represented 30% of hospital-associated isolates in Colombian hospitals [13], suggesting that USA300-LV likely was circulating in the region before 2005. Although the USA300-LV epidemic is less well characterized than its North American counterpart, its clinical and epidemiological characteristics appear to be similar [13,20]. A recent study of the transmission dynamics of CA-MRSA USA300 in northern Manhattan, where there is a large Latin American population, reported a small number of isolates with characteristics of USA300-LV (ACME-, PVL+, SaPI5+, SCC*mec* IVc) [44], raising the possibility of spread of the SAE clade to the United States.

Based on the tip-calibrated phylogeny, genomic elements encoding Panton Valentine Leukocidin (*lukFS-PVL*), SEK and SEQ enterotoxins (SaPI5; *sek, seq*) were present in the USA300 lineage many years before the emergence of the epidemic clades. Therefore, these genes are probably not the key evolutionary factors that explain the remarkable success of the epidemic lineages. Likewise, differences in gene regulation of phenol soluble modulins (*psm* genes) and alpha toxin (*hla*) may predate the recent epidemics since these genes are expressed at increased levels in other close relatives such as USA500 [10,40,45].

While the acquisition of ACME does coincide with the beginning of the NAE clade, this locus is not found in the SAE clade. Therefore, ACME-encoded adaptations that have been described so far [16-18] cannot explain the success of the South American USA300-LV clone. However, the *copB* locus, which appears to have been acquired independently as part of the ACME and COMER loci, is uniquely shared between the two epidemics. It is likely that these genes are involved in copper homeostasis/resistance [43], and given the antimicrobial properties of copper and its important role in innate immune defense, it is tempting to speculate that this locus may provide an important fitness advantage to both epidemic clades [46,47].

USA300-LV now appears to be the predominant circulating MRSA clone in both Colombia and Ecuador [13,20]. Further sampling will help elucidate the relationships of these isolates to other USA300-like CA-MRSA that are circulating in the Americas. For instance, the specific relationship of the USA300-like vancomycin-resistant isolate (BR-VRSA) from a patient in São Paulo, Brazil [39] to USA300 and USA300-LV remains unclear. Indeed, in the present analysis this genome appears to belong to a distinct, albeit closely related genetic lineage.

Conclusions

Our identification of highly virulent and closely related MRSA clades that have spread over two large regions of the Americas in parallel over a fairly short period of time suggests either convergent acquisition of traits that enhanced virulence or significant adaptations that occurred in their common ancestor. Therefore, understanding the common attributes of SAE and NAE strains could help to uncover the molecular determinants of the rapid dissemination of these lineages. While the exact geographic origin of the NAE and SAE clades remains obscure, the repeated transmission of USA300 throughout the Americas highlights both the longstanding fitness and success of this lineage. Our study also has implications for public health surveillance and preparation. It seems clear that whole genome sequencing (WGS) constitutes a new gold standard for tracking epidemics, identifying emerging strains, and distinguishing robust molecular markers (eg, *copB*). WGS can be used to more precisely target interventions or identify the possible sources of emerging disease. Moreover, it may quickly lead to identification of new genetic determinants of disease.

References

1. Talan DA, Krishnadasan A, Gorwitz RJ, et al. Comparison of *Staphylococcus aureus* from skin and soft-tissue infections in US emergency department patients, 2004 and 2008. Clin Infect Dis **2011**; 53:144-9.

2. Klein E, Smith DL, Laxminarayan R. Community-associated methicillin-resistant *Staphylococcus aureus* in outpatients, United States, 1999-2006. Emerg Infect Dis **2009**; 15:1925-30.

3. Klein E, Smith DL, Laxminarayan R. Hospitalizations and deaths caused by methicillin-resistant *Staphylococcus aureus*, United States, 1999-2005. Emerg Infect Dis **2007**; 13:1840-6.

4. Yu H, Wier LM, Elixhauser A. Hospital Stays for Children, 2009. HUCP **2011**; Statistical Brief #118.

5. Tenover FC, McDougal LK, Goering RV, et al. Characterization of a strain of community-associated methicillin-resistant *Staphylococcus aureus* widely disseminated in the United States. J Clin Microbiol **2006**; 44:108-18.

6. Centers for Disease Control and Prevention (CDC). Methicillin-resistant *Staphylococcus aureus* skin or soft tissue infections in a state prison--Mississippi, 2000. MMWR Morb Mortal Wkly Rep **2001**; 50:919-22.

7. Tenover FC, Goering RV. Methicillin-resistant *Staphylococcus aureus* strain USA300: origin and epidemiology. J Antimicrob Chemother **2009**; 64:441-6.

8. Klevens RM, Morrison MA, Nadle J, et al. Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. JAMA **2007**; 298:1763-71.

9. Klevens RM, Edwards JR, Tenover FC, et al. Changes in the epidemiology of methicillin-resistant *Staphylococcus aureus* in intensive care units in US hospitals, 1992-2003. Clin Infect Dis **2006**; 42:389-91.

10. Montgomery CP, Boyle-Vavra S, Adem PV, et al. Comparison of virulence in community-associated methicillin-resistant *Staphylococcus aureus* pulsotypes USA300 and USA400 in a rat model of pneumonia. J Infect Dis **2008**; 198:561-70.

11. Alvarez CA, Barrientes OJ, Leal AL, et al. Community-associated methicillin-resistant *Staphylococcus aureus*, Colombia. Emerg Infect Dis **2006**; 12:2000-1.

12. Arias CA, Rincon S, Chowdhury S, et al. MRSA USA300 clone and VREF--a U.S.-Colombian connection? The New England journal of medicine **2008**; 359:2177-9.

13. Reyes J, Rincón S, Díaz L, et al. Dissemination of methicillin-resistant *Staphylococcus aureus* USA300 sequence type 8 lineage in Latin America. Clin Infect Dis **2009**; 49:1861-7.

14. Diep BA, Gill SR, Chang RF, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus.* Lancet **2006**; 367:731-9.

15. Goering RV, McDougal LK, Fosheim GE, Bonnstetter KK, Wolter DJ, Tenover FC. Epidemiologic distribution of the arginine catabolic mobile element among selected methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* isolates. J Clin Microbiol **2007**; 45:1981-4.

16. Joshi GS, Spontak JS, Klapper DG, Richardson AR. Arginine catabolic mobile element encoded *speG* abrogates the unique hypersensitivity of *Staphylococcus aureus* to exogenous polyamines. Mol Microbiol **2011**; 82:9-20.

17. Thurlow LR, Joshi GS, Clark JR, et al. Functional modularity of the arginine catabolic mobile element contributes to the success of USA300 methicillin-resistant *Staphylococcus aureus*. Cell Host Microbe **2013**; 13:100-7.

18. Planet PJ, LaRussa SJ, Dana A, et al. Emergence of the epidemic methicillin-resistant *Staphylococcus aureus* strain USA300 coincides with horizontal transfer of the arginine

catabolic mobile element and *speG*-mediated adaptations for survival on skin. MBio **2013**; 4:e00889-13.

19. Portillo BC, Moreno JE, Yomayusa N, et al. Molecular epidemiology and characterization of virulence genes of community-acquired and hospital-acquired methicillin-resistant *Staphylococcus aureus* isolates in Colombia. Int J Infect Dis **2013**; 17:e744-9.

20. Jiménez JN, Ocampo AM, Vanegas JM, et al. CC8 MRSA strains harboring SCC*mec* type IVc are predominant in Colombian hospitals. Plos One **2012**; 7:e38576.

21. Robinson DA, Enright MC. Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother **2003**; 47:3926-34.

22. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res 2009; 19:1117-23.

23. Chin CS, Alexander DH, Marks P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Mehods **2013**; 10:563-9.

24. Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics **2006**; 22:699-707.

25. Li L, Stoeckert CJ, Jr R, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003; 13:2178-89.

26. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **2006**; 22:2688-90.

27. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol 1984; 20:86-93.

28. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **1994**; 39:306-14.

29. Felsenstein J. Confidence Limits on Phylogeneties: An Approch Usig the Bootstrap. Evolution **1985**; 39:783-91.

30. Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. How many bootstrap replicates are necessary? J Comput Biol 2010; 17:337-54.

31. PAUP*. PAUP*: Phylogenetic analysis using parsimony (*and other methods) v. 4b10 (Sinauer, Sunderland, Massachussetts, **2003**.

32. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol **2012**; 29:1969-73.

33. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **2002**; 161:1307-20.

34. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. Plos Biol 2006; 4:e88.

35. Kingman J. The Coalescent. Stochastic Processes and their Applicaitons **1982**; 13:235-48.

36. Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org 2011.

37. Lewis PO. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol **2001**; 50:913-25.

38. Fitch WM. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst Zoology 1971; 20:406-16.

39. Rossi F, Diaz L, Wollam A, et al. Transferable vancomycin resistance in a community-associated MRSA lineage. The New England journal of medicine **2014**; 370:1524-31.

40. Benson MA, Ohneck EA, Ryan C, et al. Evolution of hypervirulence by a MRSA clone through acquisition of a transposable element. Mol Microbiol **2014**; 93:664-81.

41. Uhlemann AC, Kennedy AD, Martens C, Porcella SF, DeLeo FR, Lowy FD. Toward an Understanding of the Evolution of Staphylococcus aureus Strain USA300 during Colonization in Community Households. Genome Biol Evol **2012**; 4:1275-85.

42. Shore AC, Coleman DC. Staphylococcal cassette chromosome *mec*: recent advances and new insights. Int J Med Microbiol **2013**; 303:350-9.

43. Baker J, Sengupta M, Jayaswal RK, Morrissey JA. The *Staphylococcus aureus* CsoR regulates both chromosomal and plasmid-encoded copper resistance mechanisms. Environ Microbiol **2011**; 13:2495-507.

44. Uhlemann AC, Dordel J, Knox JR, et al. Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. Proc Natl Acad Sci U S A **2014**; 111:6738-43.

45. Li M, Diep BA, Villaruz AE, et al. Evolution of virulence in epidemic community-associated methicillin-resistant *Staphylococcus aureus*. Proc Natl Acad Sci U S A **2009**; 106:5883-8.

46. Festa RA, Thiele DJ. Copper at the front line of the host-pathogen battle. Plos Pathog 2012; 8:e1002887.

47. Soutourina O, Dubrac S, Poupel O, Msadek T, Martin-Verstraete I. The pleiotropic CymR regulator of *Staphylococcus aureus* plays an important role in virulence and stress response. Plos Pathog **2010**; 6:e1000894.

Dr. Planet has received lecture fees, research support and consulting fees from Pfizer, Inc. Dr. Arias has received lecture fees from Pfizer, Actavis, Novartis, Cubist (Merck), The Medecins Company and Astra-Zeneca; consulting fees from Cubist, Bayer, Theravance and the Medecins Company and research support from Theravance, Actavis Pharmaceuticals and Cubist.

This work was supported by R01 AI093749 (Dr. Arias) and K08AI101005 (Dr. Planet) from the National Institute of Allergy and Infectious Diseases (NIAID) and R01 GM080602 (Dr. Robinson) from the National Institute of General Medical Sciences. Dr. Planet was also supported by the Doris Duke Clinical Scientist Development Award, the Pediatric Infectious Disease Society/St. Jude Award, and the John M. Driscoll, Jr., MD Children's Fund.

Parts of the results of the present study were presented at the 14th International symposium of on staphylococci and staphylococcal infections. Chicago, IL, 2014.

Corresponding authors´ contact information

Cesar A. Arias, MD PhD.

Division of Infectious Diseases, University of Texas Medical School at Houston, 6431 Fannin St. MSB 2.112. Houston, TX 77030

Tel: 713.500.6738 Fax: 713.500.5495

e-mail: caa22@cantab.net, cesar.arias@uth.tmc.edu

Paul J. Planet MD, PhD

Pediatric Infectious Disease Division, Columbia College of Physicians and Surgeons

630 West 168th Street, New York, NY 10032

Tel: 212-305-1296 Fax: 212 342-5218

email: pjp23@columbia.edu

Figures

Figure 1. Phylogeny shows relationships between whole genome isolates. Phylogenetic network **(A)** is based on 1000 maximum likelihood nonparametric bootstrap iterations for 88 *S. aureus* genomes. Only sequence type 8 (ST8) strains are shown (8 genomes used as outgroups have been excluded for clarity). Areas of low support show higher amounts of reticulation. Note that the two epidemic lineages NAE and SAE can be robustly distinguished from each other but exhibit high levels of reticulation internally. Whole genomes of BR-VRSA, COL, FPR3757, Newman, RN4220, TCH1516, USA500 and VC40 are shown for comparative purposes. Chronogram **(B)** was constructed using core genome SNPs from the 48 sequenced strains in a Bayesian phylogenetic analysis that co-estimated the phylogenetic relationships among isolates, the times since the divergence of the major clades and evolutionary rate of change. We used a relaxed clock model and a constant-size coalescent tree prior probability. Calendar dates of isolation were used to calibrate the clock. Branches are colored based on parsimony reconstruction of geographical location in either North America (blue) and South America (red). Numbers on or near branches are Bayesian posterior probability support values. Ancestral branches with uncertain geographical reconstructions are black. Pie charts show proportional likelihood reconstructed for each of the key branches based on the mkt1 model (see methods). Clades representing North and South American epidemics are designated as NAE and SAE respectively. Early branching (EB), strains are also indicated. Key evolutionary acquisition events (eg. ACME and COMER acquisition events) are indicated with arrows. The binary matrix shows presence (green;1) and absence (white;0) of selected genes associated with each lineage. *lukSF PVL*, Panton Valentine Leukocidin; *mecA*, PBP2a; *seq*, enterotoxin Q; *sek*, enterotoxin K; *speG*, spermidine N(1)-acetyltransferase; *arcA*, arginine deiminase; *copB*, putative ATPase copper exporter; lipo,

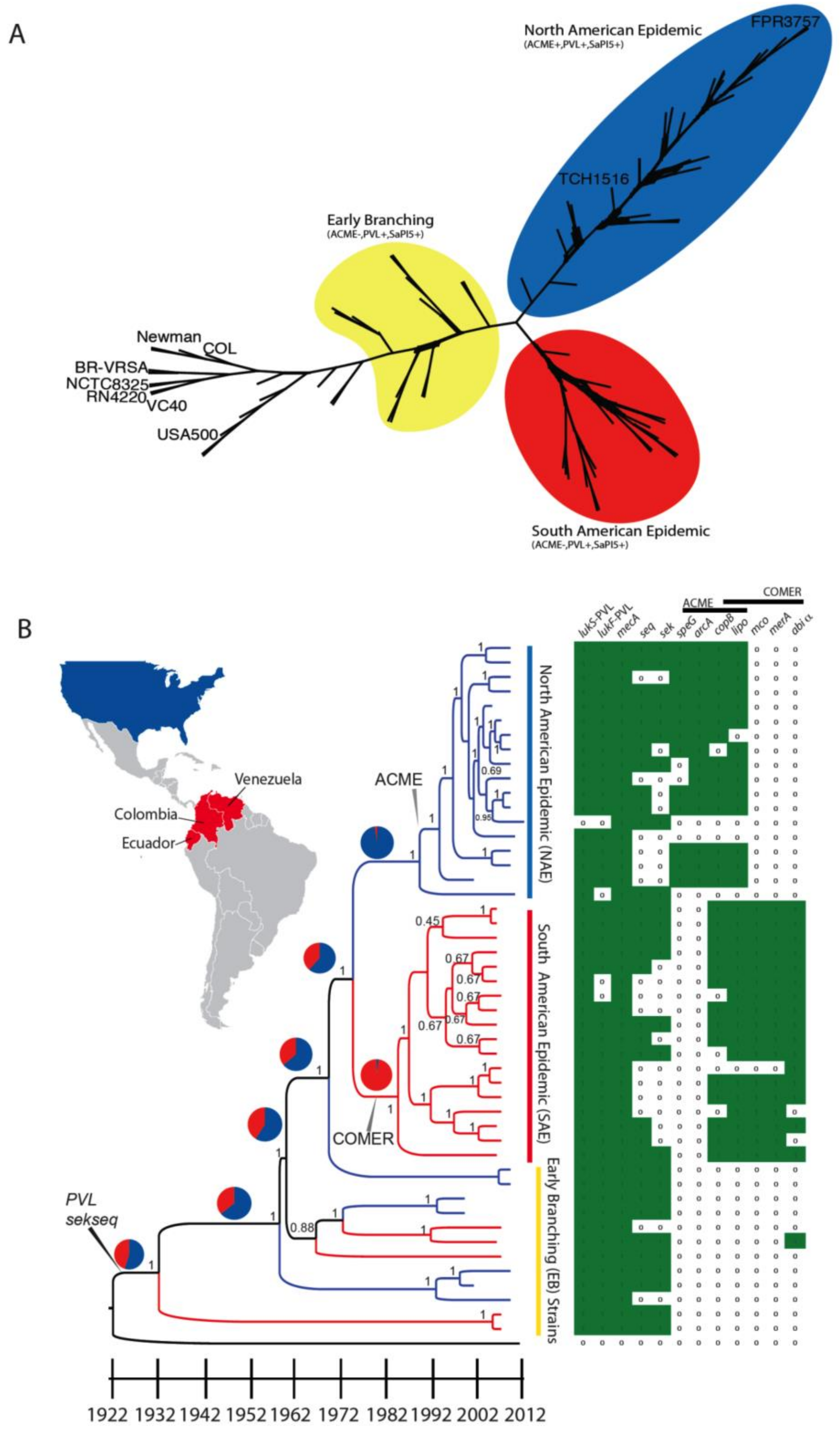
putative lipoprotein; *mco*, multicopper oxidase; *merA*, mercuric reductase; *abi* a, abortive phage infection.

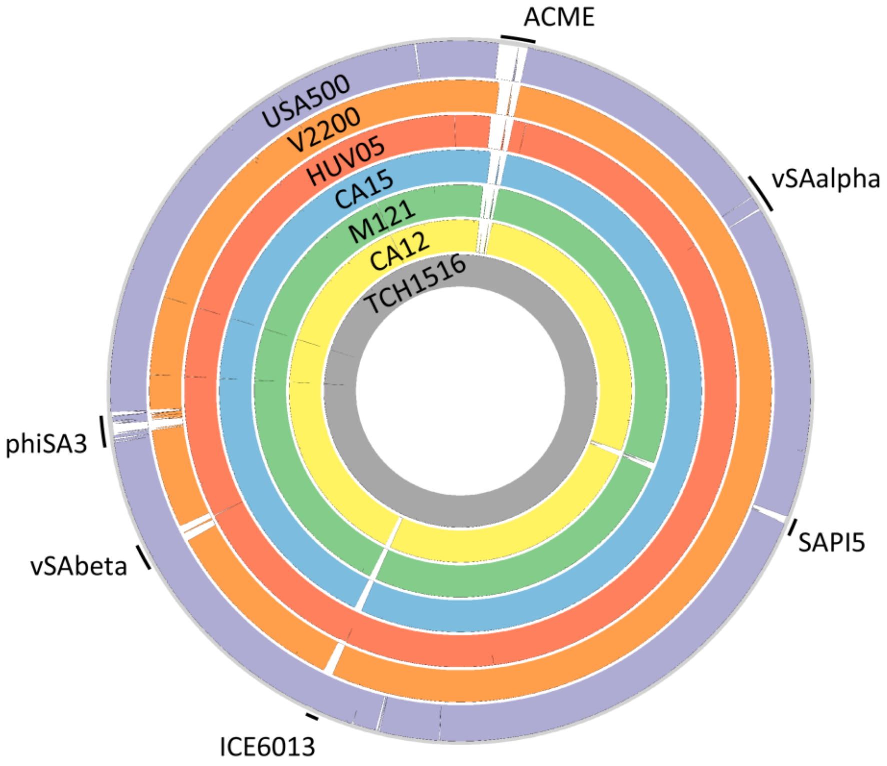
Figure 2 Circular whole-genome alignment of closed circular USA300 lineage genomes.

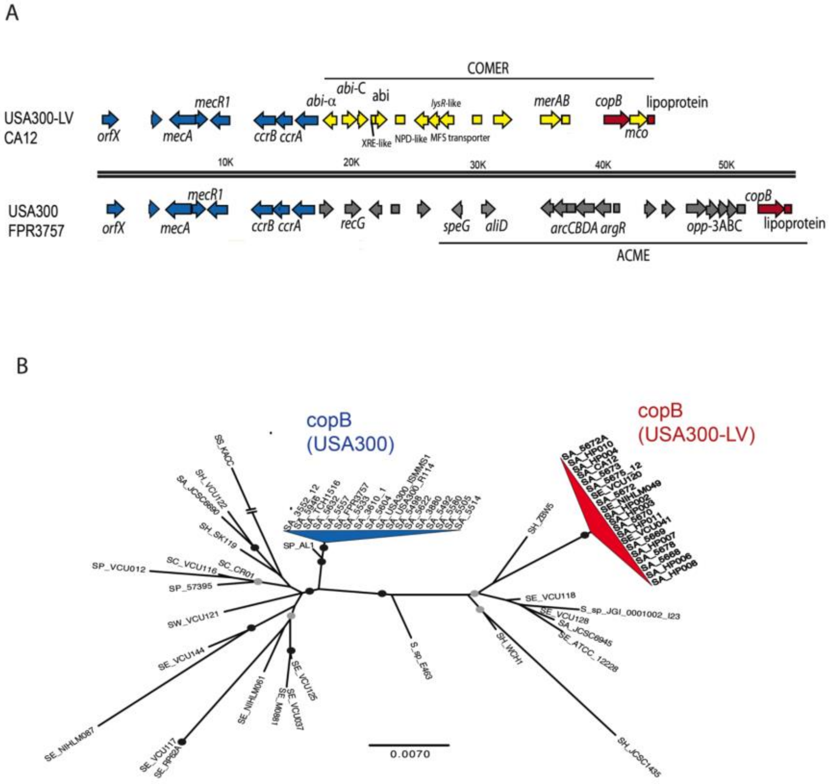
Width of each circular track is relative to nucleotide identity as calculated by whole genome alignment based on mummer (http://mummer.sourceforge.net). From inside out the tracks are TCH1516 (grey), CA12 (yellow), M121 (green), CA15 (blue), HUV05 (red), V2200 (orange). Areas of difference corresponding to mobile elements are noted.

Figure 3 Comparison of locus adjacent to SCC*mec* element in NAE and SAE strains.

(A) The arginine catabolic mobile element (ACME) is characteristic of NAE strains, and the **co**pper and **mer**cury resistance (COMER) element is characteristic of SAE strains. Only the *copB* gene (corresponding to ABD21730.1 in FPR3757) and an adjacent putative lipoprotein (corresponding to ABD21221.1) are found in both elements. **(B)**, unrooted gene maximum likelihood gene genealogy of *copB* genes from USA300 genomes and publically available staphylococcal genome sequences. Clades for *copB* genes from North American USA300 (blue) and USA300-LV (red) are distinct and appear to have been acquired in independent horizontal gene transfer events from other staphylococcal species. Black circles on branches represent RAxML non-parametric bootstrap values 80% to 100%; grey circles indicate bootstrap values of 50% to 79%. Length of branches is proportional to substitutions per site. Taxa on branches are labeled SA (*S. aureus*), SE (*S. epidermidis*), SW (*S. warneri*), SP (*Stahpylococcus pettenkoferii*), SS (*S. saprophyticus*), SC (*S. capitis*), SH (*S. haemolyticus*), SX (*S. xylosus*) or S_sp (*Staphylococcus spp.).*


A
North American Epidemic
(ACME+,PVL+,SaPI5+)
FPR3757
TCH1516
Early Branching
(ACME-,PVL+,SaPI5+)
Newman
COL
BR-VRSA
NCTC8325
RN4220
VC40
USA500
South American Epidemic
(ACME-,PVL+,SaPI5+)
B
Venezuela
Colombia
Ecuador
ACME
COMER
PVL
sekseq
North American Epidemic (NAE)
South American Epidemic (SAE)
Early Branching (EB) Strains
lukS-PVL
lukF-PVL
mecA
seq
sek
speG
arcA
copB
lipo
mco
merA
abi α
1
0.69
0.95
0.45
0.67
0.88
1922
1932
1942
1952
1962
1972
1982
1992
2002
2012


ACME
vSAalpha
SAPI5
ICE6013
vSAbeta
phiSA3
USA500
V2200
HUV05
CA15
M121
CA12
TCH1516


A
COMER
USA300-LV
CA12
orfX
mecA
mecR1
ccrB ccrA
abi-α
abi-C
abi
XRE-like
NPD-like
lysR-like
MFS transporter
merAB
copB
mco
lipoprotein
10K
20K
30K
40K
50K
USA300
FPR3757
orfX
mecA
mecR1
ccrB ccrA
recG
speG
aliD
arcCBDA argR
opp-3ABC
copB
lipoprotein
ACME
B
copB
(USA300)
copB
(USA300-LV)
SS_KACC
SH_VCU122
SA_JCSC6690
SH_SK119
SP_VCU012
SC_VCU116
SC_CR01
SP_57395
SW_VCU121
SE_VCU144
SE_NIHLM087
SE_VCU117
SE_RP62A
SE_NIHLM061
SE_M0881
SE_VCU037
SE_VCU125
SP_AL1
SA_3552_12
SA_5946
SA_TCH1516
SA_5632
SA_5557
SA_FPR3757
SA_5533
SA_3610_1
SA_5604
SA_USA300_ISMMS1
SA_USA300_R114
SA_5498
SA_5622
SA_3880
SA_5492
SA_5580
SA_5505
SA_5514
S_sp_E463
SH_ZBW5
SA_5672A
SA_HP010
SA_HP004
SA_CA12
SA_5673
SA_5675_12
SE_VCU120
SA_5672
SE_NIHLM049
SA_HP002
SA_HP003
SA_5670
SA_HP011
SE_VCU041
SA_5669
SA_HP007
SA_5678
SA_5668
SA_HP006
SA_HP008
SE_VCU118
S_sp_JGI_0001002_I23
SE_VCU128
SA_JCSC6945
SE_ATCC_12228
SH_WCH1
SH_JCSC1435
0.0070