Published in final edited form as:

Curr Opin Microbiol. 2018 February ; 41: 43–50. doi:10.1016/j.mib.2017.11.014.

After the deluge: Mining *Staphylococcus aureus* genomic data for clinical associations and host pathogen interactions

Richard Copin^{1,*}, Bo Shopsin^{1,2,*}, and Victor J. Torres^{2,*}

¹Department of Medicine, New York University School of Medicine, 430 East 29th Street, Alexandria West, 10016, New York, NY

²Department of Microbiology, New York University School of Medicine, 430 East 29th Street, Alexandria West, 10016, New York, NY

Abstract

The genome of *Staphylococcus aureus* has rapidly become one of the most frequently sequenced among bacteria, with more than 40,000 genome sequences uploaded to public databases. Computational resources required for analysis and quality assessment have lagged behind accumulation of sequence data. Improved analytic pipelines, in combination with the development of customized *S. aureus* reference databases, can be used to inform *S. aureus* biology and potentially predict clinical outcome. Here, we review the currently available data about *S. aureus* genome in public databases, and discuss their potential utility for understanding *S. aureus* evolution. Also discussed are ways to overcome challenges to the application of whole genome sequencing data for prevention and management of *S. aureus* disease.

INTRODUCTION

Staphylococcus aureus is one of the most important pathogens in industrialized countries, infecting more Americans than HIV and causing more deaths than *Streptococcus pneumoniae*, *Neisseria meningitidis*, *Haemophilus influenzae*, and *Streptococcus pyogenes* combined [1]. Although traditionally linked to the hospital environment, community-associated methicillin-resistant *S. aureus* (MRSA) strains have recently emerged, increasing both the frequency and severity of *S. aureus* infection [2]. The high mortality due to infection by *S. aureus*, despite intensive care and the best-available antibiotic therapy, emphasizes the need for improved methods for prevention and treatment. For that, it is necessary to understand how *S. aureus* responds to changing environments.

Genome plasticity has allowed *S. aureus* to accommodate a high degree of environmental variation. For instance, it is estimated that 44% of *S. aureus* genome is not shared by all *S. aureus* strains [3]. Additionally, mutation, alone or in combination with phage and plasmid

*Corresponding authors: Richard.Copin@nyumc.org; Bo.Shopsin@nyumc.org; and Victor.Torres@nyumc.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

mobilization, modulate virulence and antimicrobial resistance in *S. aureus* (for example [4–7]). Whole-genome sequencing (WGS) is now the gold standard for cataloguing and understanding genetic diversity. Recent advances in the speed and ease of genome sequencing and analysis, have enabled a profusion of in-depth bacterial genomic surveillance during outbreaks of infection [8–10]. Indeed, the past five years have witnessed a deluge of *S. aureus* sequenced genomes, propelling *S. aureus* to the top of the list of the most frequently sequenced bacterial species. Large-scale genomic datasets have revealed epidemiological consequences of adaptive changes in *S. aureus* [11,12], and they facilitate identification of key molecular features defining *S. aureus* traits associated with specific infectious syndromes [13]. However, with thousands of genomes deposited online, it has become increasingly difficult to manage the expanding volume of sequencing information. Moreover, accessing sequence data is not user-friendly; computational skills are required to translate the data into a workable form. Below we describe and summarize the current knowledge available in these databases, discuss ways to broaden access, and highlight the potential of such information to uncover unique characteristics of the *S. aureus* genome that influence host-pathogen interaction.

Proliferation of *S. aureus* genomic sequences in public databases

The development of cost-effective, high throughput sequencing instruments and multiplex sequencing protocols have broadened the appeal of genomic analysis, transforming the field. From August 2010 to August 2017, public databases expanded from 19 to 48,259 *S. aureus* genome sequences derived from 4,155 independent studies (Figure 1). Most sequence data are deposited at the Sequence Read Archive at the US National Center for Biotechnology Information (NCBI), which is part of the International Nucleotide Sequence Database Collaboration (<https://www.ncbi.nlm.nih.gov/sra/>). Although 84% (42,285) of the genome sequences represent short DNA reads or small fragments, the remaining fraction (7,974 – 16%) have been assembled into larger DNA segments and represent partially complete *S. aureus* genomes. Partial assemblies are usually annotated and can be retrieved in the NCBI Genbank assembly database (<https://www.ncbi.nlm.nih.gov/assembly/>, Table S1). *S. aureus* has the largest number of annotated sequences when compared to other bacterial species such as *Mycobacterium tuberculosis* or *Escherichia coli* (Figure 2). However, only 2% (166/7,974) of *S. aureus* assemblies are gapless and therefore represent fully assembled genomes. This small fraction of assembled genomes is biased towards epidemic lineages of MRSA from multilocus sequencing type clonal complexes (CC) 5 and 8 (Figure 3). Additionally, although it is likely that hundreds of *S. aureus* CCs exist, only 32 are believed to be represented in Genbank [14]. In sum, fully assembled genomes represent a small fraction of *S. aureus* genomes available in NCBI (0.4%), and are biased towards a small subset of *S. aureus* lineages. *S. aureus* genetic diversity, immunomodulation, and damage by individual strains can vary dramatically [3,15–19]. Thus, studies that capture the breadth of clones to account for the diversity in host-pathogen interactions at the genomic level will likely yield unique understanding of the biology of MRSA. That understanding should then promote the development of more effective and personalized approaches for preventing infection and improving management of MRSA.

Mobile genetic element (MGE) diversity in public databases

MGE discovery and identification are important goals for genome analysis because almost all *S. aureus* strains harbor one or more MGE with potentially syndrome- and tissue-specific functions. MGEs encompass any intra- or extra-chromosomal DNA segment that can be independently mobilized within or between *S. aureus* cells. Staphylococcal MGEs include plasmids, transposons, integrons, genomic islands, *S. aureus* pathogenicity islands (SaPIs), integrative conjugative elements, staphylococcal chromosome cassettes, and phages [20–22]. MGEs are abundant, representing 15–20% of the *S. aureus* genome, and are highly repetitive. High level of sequence identity facilitates genetic recombination and contributes to the emergence of mosaic MGEs. As such, MGE nomenclature in NCBI databases is constantly evolving and inconsistently annotated among *S. aureus* strains. Improvements in annotation and “long-read” sequencing, discussed below, will enhance our understanding of the evolutionary processes that contribute to the emergence and persistence of MGEs and epidemically successful pathogen clones.

Phages and plasmids are the main source of MGE diversity among *S. aureus* NCBI sequences. Plasmids – auto-replicating, covalently closed, circular double-stranded DNA molecules – usually do not contain genes required for essential cellular functions. The 327 unique *S. aureus* plasmid sequences in the NCBI database (Table S1) can be differentiated into at least 18 plasmid types based on replication mechanism [23]. *S. aureus* plasmid sizes range from ~1kb up to more than 100kb, with a median size of ~20kb (Figure 4A). Small, cryptic plasmids are prevalent and often comprise only replication genes and a few hypothetical coding sequences. Larger plasmids tend to be implicated in detoxification of heavy metals, dissemination of multidrug resistance, and production of virulence-associated proteins [24]. Although most *S. aureus* plasmids are mobilizable, the vast majority (~90%) do not encode all the elements needed for transfer. They rely on “helper” phages or larger conjugative plasmids (5–6% of *S. aureus* plasmid diversity) to transmit between *S. aureus* strains [25,26].

Phages are bacterial viruses that can integrate into the *S. aureus* chromosome. As the primary vehicles for horizontal transfer of MGEs between strains, phages are major drivers of staphylococcal genome evolution [27]. All *S. aureus* genomes sequenced to date contain at least one phage, and many strains carry up to four [28]. The 87 unique *S. aureus* phage sequences uploaded to NCBI average around ~55kb and range from ~15kb to more than 148kb, making phages the largest MGEs in *S. aureus* (Figure 4B, Table S2). In addition to numerous hypothetical proteins, phages carry antibiotic resistance genes and virulence factors, such as the Panton-Valentine leucocidin (PVL) toxin [29,30]. The PVL phage is epidemiologically associated with specific clinical syndromes, including purulent skin and soft tissue infection and necrotizing pneumonia [31,32]. Phages also mobilize other MGEs such as plasmids and staphylococcal pathogenicity islands (SaPIs) [21]. SaPIs, which integrate into specific attachment sites in the chromosome, employ phage machinery for replication and dissemination [20]. Thus, phages impact both the virulence and adaptability of *S. aureus*.

Identification and discovery of MGEs is complicated by a multitude of repeat regions and mosaic elements that cannot be assembled using traditional “short-read” sequencing instruments. *De novo* sequence assembly using a combination of short-read and long-read third-generation sequencing can provide a comprehensive characterization of MGEs and repeat regions [4,33]. Although “hybrid” long and short-read sequencing is currently costlier than short-read sequencing alone, the cost is offset by much greater accuracy in MGE analysis. Moreover, increases in long-read sequencing accuracy, throughput, and reduced multiplex sequencing costs will likely eliminate the need for hybrid analysis in the near future.

Genome mis-annotation in *S. aureus* databases

Public databases serve not only as storage devices, but they also represent a powerful resource for the study of *S. aureus* biology, ecology, and evolution. Rather than accounting for the breath of genomic diversity that is available in public databases, comparative genomic analyses are often guided by a handful of “reference” genomes. Utilization of available sequences from non-reference strains requires well-curated, up-to-date genome annotation across the *S. aureus* lineage. Unfortunately, inconsistency in annotation among *S. aureus* strains is common and complicates functional inter- and intra-species analyses. Particularly problematic are orthologous genes (gene of common ancestry) that are often mis-annotated as encoding distinct products in different strains. For example, *S. aureus* encodes a family of highly related bi-component, pore-forming cytotoxins that injure host leukocytes (Table 1) [34]. In this family, the product of *lukA* (also known as *lukH*), which encodes one of the 2 subunits of leucocidin LukAB [35], is annotated as a hypothetical protein, a putative leukocidin subunit, or a succinyl-diaminopimelate desuccinylase in *S. aureus* Mu50, MSSA476, and USA500 strains, respectively. Overall, 40% of *S. aureus* genes have inconsistent annotation across *S. aureus* lineages, complicating comparative analysis based on product annotation. To support studies aimed at understanding and exploiting genomic diversity in *S. aureus*, improved curation of genome annotation is needed to standardize and improve the consistency of comparative analyses.

Current annotation methods, such as Reciprocal Best Hits BLAST analysis, fail to provide unambiguous assignment of ortholog genes. Consequently, tools such as the Staphopia R package are being developed to provide faster and more accurate annotation of *S. aureus* whole-genome datasets (<https://staphopia.emory.edu/>). Currently the gold standard for comparing genomes remains a combination of manually curated gene annotations and computational analyses [36]. The development of customized *S. aureus* reference databases similar to the Pseudomonas Genome Database (PGD; <http://www.pseudomonas.com>) [36], a community-assisted initiative for assigning orthologs, would be a major advance for comparative analysis and population genomics studies of *S. aureus* genomes. The PGD assigns, maintains, and updates high-quality annotations using genome-scale, high-precision computational predictions and manual curation. Application of this approach to *S. aureus* will benefit not only genomic comparisons but also downstream analysis, such as gene and protein expression quantification. Moreover, integration of detailed isolate metadata (e.g., virulence and resistance phenotype, clinical syndrome, and outcome) and intuitive tools to improve data visualization would facilitate clone and syndrome-specific comparative

analyses. Adaption of a PGD-like program with enforcement of high standards to *S. aureus* could transform the diagnosis and treatment of *S. aureus* infection.

***S. aureus* genomics and pathogen adaptation**

The proliferation of *S. aureus* genomic sequences reflects strong interest in understanding *S. aureus* genome diversity and evolution. Genome-wide surveys of natural *S. aureus* populations have shed light on *S. aureus* pathobiology. For instance, genomic studies have pinpointed the dramatic impact of MGE on the success of MRSA clones such as USA300, the predominant community-acquired MRSA clone in the US. USA300 acquired not only *pvl* but also other potentially syndrome-specific virulence factors such as arginine catabolic mobile element (ACME) [37]. Additional studies have found that even a modest genetic change can have a dramatic impact on virulence factor expression and host-pathogen interaction [38]. Examples include inactivating mutations in global regulators *agr* and *rsp*, which arise during the course of infection [7,39–41]. Within-host evolution has implications not only for patients, but also for establishing thresholds to differentiate relatedness in *S. aureus* for epidemiological purposes in hospitals [42].

WGS data has also set the stage for genome-wide association studies (GWAS) aiming to identify the most predictive determinants of poor infection outcomes. Variant linkage analysis has the potential to bridge genotypes to phenotypes and enhance our understanding of bacterial pathogenesis [43]. Although GWAS has been applied successfully to identify human genetic variants contributing to rare disorders, they have fared poorly at identifying bacterial genes that influence common infections. Although recent studies have begun to shed light upon bacterial polymorphisms influencing *S. aureus* infection [13,44], marked heterogeneity in the clinical presentation of *S. aureus* disease and the complexity of staphylococcal virulence can obscure the relative contributions of host and bacterial factors to patient outcomes. Unlike specialized pathogens such as *Mycobacterium tuberculosis* and HIV, *S. aureus* does not lead to a single prognosis. Instead, it is linked to a wide spectrum of detrimental conditions. Robust datasets from patients with carefully defined epidemiology-host-microbe relationships, including antibiotic use and outcome, are needed to determine the significance and clinical utility of changes in the core and accessory genome.

Concluding remarks

Genome sequencing has caused research on *S. aureus* to take a significant leap forward [3,5,9,11,38]. New technologies, such as long-read sequencing, will further improve our ability to integrate *S. aureus* MGE mobilization and mutation. As whole-genome sequencing tools improve and the number of sequenced genomes multiply, the need for curated databases will be crucial for rapid sequence analysis across strains. Elimination of the time-constraining process of creating tools for analyzing strain-specific orthology promises to produce a significant resource for researchers seeking a greater understanding of *S. aureus* and other pathogens. We emphasize that the influence of host diversity and bacterial ecology must also be considered. Such complexity requires multi-disciplinary programs to characterize genotypic differences among strains. The needed areas of expertise include genomics, molecular genetics, evolutionary biology, epidemiology, and host-pathogen

interactions. We expect that such multidisciplinary programs will lead to novel approaches for managing *S. aureus* and provide a foundation for examining the adaptations of other opportunistic pathogens.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Karl Drlica for critical comments on the manuscript and Karen Clark and Ben Busby for helping navigating the NCBI databases. R.C., V.J.T. and B.S. were supported by National Institutes of Health grants R01 AI103268 and NIAID HHSN272201400019C. V.J.T. was also supported by National Institutes of Health grants R01 AI099394 and R01 AI105129. V.J.T. is a Burroughs Wellcome Fund Investigator in the Pathogenesis of Infectious Diseases.

References

1. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, et al. Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. *JAMA*. 2007; 298:1763–1771. [PubMed: 17940231]
2. David MZ, Daum RS. Community-associated methicillin-resistant *Staphylococcus aureus*: epidemiology and clinical consequences of an emerging epidemic. *Clin Microbiol Rev*. 2010; 23:616–687. [PubMed: 20610826]
- 3(••). Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BO. Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc Natl Acad Sci U S A*. 2016; 113:E3801–3809. Characteristics of the *Staphylococcus aureus* core genome and pangenome and implication on *Staphylococcus aureus* adaption. [PubMed: 27286824]
4. Benson MA, Ohneck EA, Ryan C, Alonzo F 3rd, Smith H, Narechania A, Kolokotronis SO, Satola SW, Uhlemann AC, Sebra R, et al. Evolution of hypervirulence by a MRSA clone through acquisition of a transposable element. *Mol Microbiol*. 2014; 93:664–681. [PubMed: 24962815]
5. Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, Lin F, Lin J, Carleton HA, Mongodin EF, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired methicillin-resistant *Staphylococcus aureus*. *Lancet*. 2006; 367:731–739. [PubMed: 16517273]
6. Kennedy AD, Otto M, Braughton KR, Whitney AR, Chen L, Mathema B, Mediavilla JR, Byrne KA, Parkins LD, Tenover FC, et al. Epidemic community-associated methicillin-resistant *Staphylococcus aureus*: recent clonal expansion and diversification. *Proc Natl Acad Sci U S A*. 2008; 105:1327–1332. [PubMed: 18216255]
7. Young BC, Golubchik T, Batty EM, Fung R, Lerner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci U S A*. 2012; 109:4550–4555. [PubMed: 22393007]
8. Planet PJ, Diaz L, Kolokotronis SO, Narechania A, Reyes J, Xing G, Rincon S, Smith H, Panesso D, Ryan C, et al. Parallel Epidemics of Community-Associated Methicillin-Resistant *Staphylococcus aureus* USA300 Infection in North and South America. *J Infect Dis*. 2015; 212:1874–1882. [PubMed: 26048971]
9. Tong SY, Holden MT, Nickerson EK, Cooper BS, Koser CU, Cori A, Jombart T, Cauchemez S, Fraser C, Wuthiekanun V, et al. Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Res*. 2015; 25:111–118. [PubMed: 25491771]
- 10(•). Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, et al. Whole-genome sequencing for analysis of an outbreak of

- meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis*. 2013; 13:130–136. An example of the utilization of whole genome sequencing to characterize and understand *Staphylococcus aureus* spread during an outbreak. [PubMed: 23158674]
- 11(••). Uhlemann AC, Dordel J, Knox JR, Raven KE, Parkhill J, Holden MT, Peacock SJ, Lowy FD. 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; 111:6738–6743. A prospective study using whole genome sequencing to characterize and understand *Staphylococcus aureus* spread in the community. [PubMed: 24753569]
 12. Li M, Diep BA, Villaruz AE, Braughton KR, Jiang X, DeLeo FR, Chambers HF, Lu Y, Otto M. Evolution of virulence in epidemic community-associated methicillin-resistant *Staphylococcus aureus*. *Proc Natl Acad Sci U S A*. 2009; 106:5883–5888. [PubMed: 19293374]
 13. Recker M, Laabei M, Toleman MS, Reuter S, Saunderson RB, Blane B, Torok ME, Ouadi K, Stevens E, Yokoyama M, et al. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nat Microbiol*. 2017
 14. Planet PJ, Narechania A, Chen L, Mathema B, Boundy S, Archer G, Kreiswirth B. Architecture of a Species: Phylogenomics of *Staphylococcus aureus*. *Trends Microbiol*. 2017; 25:153–166. [PubMed: 27751626]
 15. Peres AG, Stegen C, Li J, Xu AQ, Levast B, Surette MG, Cousineau B, Desrosiers M, Madrenas J. Uncoupling of pro- and anti-inflammatory properties of *Staphylococcus aureus*. *Infect Immun*. 2015; 83:1587–1597. [PubMed: 25644014]
 16. Parcina M, Miranda-Garcia MA, Durlanik S, Ziegler S, Over B, Georg P, Foermer S, Ammann S, Hilmi D, Weber KJ, et al. Pathogen-triggered activation of plasmacytoid dendritic cells induces IL-10-producing B cells in response to *Staphylococcus aureus*. *J Immunol*. 2013; 190:1591–1602. [PubMed: 23325892]
 17. Hilmi D, Parcina M, Stollewerk D, Ostrop J, Josten M, Meilaender A, Zaehring U, Wichelhaus TA, Bierbaum G, Heeg K, et al. Heterogeneity of host TLR2 stimulation by *Staphylococcus aureus* isolates. *PLoS One*. 2014; 9:e96416. [PubMed: 24810614]
 18. Cheung GY, Wang R, Khan BA, Sturdevant DE, Otto M. Role of the accessory gene regulator *agr* in community-associated methicillin-resistant *Staphylococcus aureus* pathogenesis. *Infect Immun*. 2011; 79:1927–1935. [PubMed: 21402769]
 19. Chaves-Moreno D, Wos-Oxley ML, Jauregui R, Medina E, Oxley AP, Pieper DH. Exploring the transcriptome of *Staphylococcus aureus* in its natural niche. *Sci Rep*. 2016; 6:33174. [PubMed: 27641137]
 - 20(•). Penades JR, Christie GE. The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites. *Annu Rev Virol*. 2015; 2:181–201. A thorough review describing the diversity of a family of highly mobile genetic elements that contribute substantively to *Staphylococcus aureus* adaptation. [PubMed: 26958912]
 21. Novick RP, Christie GE, Penades JR. The phage-related chromosomal islands of Gram-positive bacteria. *Nat Rev Microbiol*. 2010; 8:541–551. [PubMed: 20634809]
 22. Malachowa N, DeLeo FR. Mobile genetic elements of *Staphylococcus aureus*. *Cell Mol Life Sci*. 2010; 67:3057–3071. [PubMed: 20668911]
 23. Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol*. 2015; 6:242. [PubMed: 25873913]
 24. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*. 2005; 3:722–732. [PubMed: 16138100]
 25. Ramsay JP, Kwong SM, Murphy RJ, Yui Eto K, Price KJ, Nguyen QT, O'Brien FG, Grubb WB, Coombs GW, Firth N. An updated view of plasmid conjugation and mobilization in *Staphylococcus*. *Mob Genet Elements*. 2016; 6:e1208317. [PubMed: 27583185]
 - 26(•). Ramsay JP, Firth N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr Opin Microbiol*. 2017; 38:1–9. An interesting review discussing the diverse mobilization strategies employed by mobile genetic elements in *Staphylococcus aureus*. [PubMed: 28391142]

27. Xia G, Wolz C. Phages of *Staphylococcus aureus* and their impact on host evolution. *Infect Genet Evol.* 2014; 21:593–601. [PubMed: 23660485]
28. Lindsay JA. Genomic variation and evolution of *Staphylococcus aureus*. *Int J Med Microbiol.* 2010; 300:98–103. [PubMed: 19811948]
29. Haaber J, Leisner JJ, Cohn MT, Catalan-Moreno A, Nielsen JB, Westh H, Penades JR, Ingmer H. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat Commun.* 2016; 7:13333. [PubMed: 27819286]
30. Nguyen LT, Vogel HJ. Staphylokinase has distinct modes of interaction with antimicrobial peptides, modulating its plasminogen-activation properties. *Sci Rep.* 2016; 6:31817. [PubMed: 27554435]
31. Gillet Y, Issartel B, Vanhems P, Fournet JC, Lina G, Bes M, Vandenesch F, Piemont Y, Brousse N, Floret D, et al. Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *Lancet.* 2002; 359:753–759. [PubMed: 11888586]
32. Lina G, Piemont Y, Godail-Gamot F, Bes M, Peter MO, Gauduchon V, Vandenesch F, Etienne J. Involvement of Panton-Valentine leukocidin-producing *Staphylococcus aureus* in primary skin infections and pneumonia. *Clin Infect Dis.* 1999; 29:1128–1132. [PubMed: 10524952]
33. Altman DR, Sebra R, Hand J, Attie O, Deikus G, Carpini KW, Patel G, Rana M, Arvelakis A, Grewal P, et al. Transmission of methicillin-resistant *Staphylococcus aureus* via deceased donor liver transplantation confirmed by whole genome sequencing. *Am J Transplant.* 2014; 14:2640–2644. [PubMed: 25250641]
34. Spaan AN, van Strijp JAG, Torres VJ. Leukocidins: staphylococcal bi-component pore-forming toxins find their receptors. *Nat Rev Microbiol.* 2017; 15:435–447. [PubMed: 28420883]
35. DuMont AL, Yoong P, Liu X, Day CJ, Chumbler NM, James DB, Alonzo F 3rd, Bode NJ, Lacy DB, Jennings MP, et al. Identification of a crucial residue required for *Staphylococcus aureus* LukAB cytotoxicity and receptor recognition. *Infect Immun.* 2014; 82:1268–1276. [PubMed: 24379286]
- 36(••). Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FS. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic Acids Res.* 2016; 44:D646–653. A pioneer paper describing a community-assisted initiative for assigning curated annotation to *pseudomonas* genomes. [PubMed: 26578582]
37. Goering RV, McDougal LK, Fosheim GE, Bonnstedter 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; 45:1981–1984. [PubMed: 17409207]
38. Viana D, Comos M, McAdam PR, Ward MJ, Selva L, Guinane CM, Gonzalez-Munoz BM, Tristan A, Foster SJ, Fitzgerald JR, et al. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat Genet.* 2015; 47:361–366. [PubMed: 25685890]
39. Tachibana S, Nishiura S, Ishida S, Kakehi K, Honda S. Quality check of heparin injections by ¹H-nuclear magnetic resonance spectroscopy. *Chem Pharm Bull (Tokyo).* 1990; 38:2503–2506. [PubMed: 2285982]
40. Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, et al. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc Natl Acad Sci U S A.* 2007; 104:9451–9456. [PubMed: 17517606]
41. Traber KE, Lee E, Benson S, Corrigan R, Cantera M, Shopsis B, Novick RP. agr function in clinical *Staphylococcus aureus* isolates. *Microbiology.* 2008; 154:2265–2274. [PubMed: 18667559]
42. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. *Nat Rev Microbiol.* 2016; 14:150–162. [PubMed: 26806595]
- 43(•). Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol.* 2012; 8:e1002822. An insightful review about Genome-wide association studies. [PubMed: 23300413]

44. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, et al. Predicting the virulence of MRSA from its genome sequence. *Genome Res.* 2014; 24:839–849. [PubMed: 24717264]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights

1. 40,000 *S. aureus* sequences are available in NCBI but fully assembled genomes represent only 0.4% of sequences.
2. *S. aureus* genomic heterogeneity derives from a combination of mutation and mobile genetic element diversity; more than 300 unique plasmids and 80 phages have been identified.
3. A combination of short-read and long-read third-generation sequencing offers a new way to capture the full extent of *S. aureus* mobile genetic element diversity.
4. Advances in whole genome sequencing techniques have enabled real-time genomic surveillance of *S. aureus* clones, but inconsistent gene product annotation confounds genomic analysis.
5. Systems and methods developed for annotating other pathogens, such as community-assisted curation of genomic data, can be applied to *S. aureus*.

Figure 1. Few *S. aureus* genomes are assembled or annotated

Pie chart showing the proportion of *S. aureus* genomes available in NCBI that are fully (0.4%), partially (16%) or not assembled (83.6%). SRA, Sequence Read Archive; WGS, whole genome sequencing.

Figure 2. *S. aureus* has the highest number of annotated genomes

Number of annotated bacterial genomes deposited in the NCBI assembly database from 2012 to 2017 (<https://www.ncbi.nlm.nih.gov/assembly/>).

Figure 3. Clonal complexes 5 and 8 represent the majority of genomes sequenced
Phylogeny reconstruction of fully assembled *S. aureus* genomes (n=166) show a bias towards sequencing of MRSA strains from the clonal complexes 5 (blue) and 8 (yellow).

Figure 4. Number and size of *S. aureus* plasmids and phages available in NCBI

List of plasmid and phages sequences used in these graphs are available in supplementary table 1 and 2. Dotted line indicates the median size.

Product annotation of leukocidin-encoding genes in 6 *S. aureus* strains (USA300_FPR3757, Newman, Mu50, MSSA476, MRSA252 and USA500).

Table 1

	Staphylococcus aureus genomes					
Gene name	USA300 (NC_007793.1)	Newman (NC_009641.1)	Mu50 (NC_002758_2)	MSSA476 (NC_002953.3)	MRSA252 (NC_002952.2)	USA500 (NZ_CP007499)
lukA	Aerolysin/Leukocidin family protein	leukocidin/hemolysin toxin family S subunit	hypothetical protein	putative leukocidin S subunit	putative leukocidin S subunit	succinyl-diaminopimelate desuccinylase
lukB	Leukocidin/Hemolysin toxin family protein	leukocidin/hemolysin toxin family F subunit	hypothetical protein	putative leukocidin F subunit	putative leukocidin F subunit	gamma-hemolysin subunit B
lukD	leukotoxin LukD	leukocidin LukD precursor	leukotoxin F-subunit	gamma-hemolysin component B precursor	absent	leucotoxin LukD
lukE	leukotoxin LukE	leukocidin LukE precursor	leukotoxin S-subunit	gamma-hemolysin component A precursor	absent	gamma-hemolysin subunit A
lukF-PV	Panton-Valentine leukocidin LukF-PV	absent	absent	absent	absent	absent
lukS-PV	Panton-Valentine leukocidin LukS-PV	absent	absent	absent	absent	absent
hlgC	gamma-hemolysin component C	gamma-hemolysin component C	gamma-hemolysin component C	gamma-hemolysin component C precursor	gamma-hemolysin component C precursor	gamma-hemolysin subunit A
hlgB	gamma-hemolysin component B	gamma hemolysin%2C component B	gamma-hemolysin component B	gamma-hemolysin component C precursor	gamma-hemolysin component C precursor	gamma-hemolysin subunit B
hlgA	gamma-hemolysin component A	gamma-hemolysin component A	gamma-hemolysin chain II precursor	gamma-hemolysin component A precursor	gamma-hemolysin component A precursor	gamma-hemolysin subunit A
hla	alpha-hemolysin precursor	alpha-hemolysin precursor	alpha-hemolysin precursor	alpha-hemolysin precursor	alpha-hemolysin precursor (pseudogene)	alpha-hemolysin