Journal of Hospital Infection 83 (2013) 14–21

Available online at www.sciencedirect.com

Journal of Hospital Infection

journal homepage: www.elsevierhealth.com/journals/jhin

Review

Whole genome sequencing in the prevention and control of *Staphylococcus aureus* infection

J.R. Price [a,*], X. Didelot [b], D.W. Crook [c], M.J. Llewelyn [a], J. Paul [a,d]

[a] *Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton, UK*
[b] *Department of Infectious Disease Epidemiology, Imperial College, London, UK*
[c] *Nuffield Department of Medicine, Experimental Medicine Division, John Radcliffe Hospital, Oxford, UK*
[d] *Health Protection Agency, Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton, UK*

ARTICLE INFO

Article history:
Received 8 March 2012
Accepted 2 October 2012
Available online 16 November 2012

Keywords:
Outbreak investigation
Staphylococcus aureus
Typing
Whole genome sequencing

SUMMARY

Background: *Staphylococcus aureus* remains a leading cause of hospital-acquired infection but weaknesses inherent in currently available typing methods impede effective infection prevention and control. The high resolution offered by whole genome sequencing has the potential to revolutionise our understanding and management of *S. aureus* infection.

Aim: To outline the practicalities of whole genome sequencing and discuss how it might shape future infection control practice.

Methods: We review conventional typing methods and compare these with the potential offered by whole genome sequencing.

Findings: In contrast with conventional methods, whole genome sequencing discriminates down to single nucleotide differences and allows accurate characterisation of transmission events and outbreaks and additionally provides information about the genetic basis of phenotypic characteristics, including antibiotic susceptibility and virulence. However, translating its potential into routine practice will depend on affordability, acceptable turnaround times and on creating a reliable standardised bioinformatic infrastructure.

Conclusion: Whole genome sequencing has the potential to provide a universal test that facilitates outbreak investigation, enables the detection of emerging strains and predicts their clinical importance.

Introduction

In the field of infection control, our understanding of *Staphylococcus aureus* transmission is limited by the methods used to determine the relatedness of micro-organisms in the context of time and space. Conventional typing methods, such as phage typing, multi-locus sequence typing (MLST) and pulsed-field gel electrophoresis (PFGE), have been used successfully to describe the global population structure of *S. aureus*, to provide a framework for the description of the major lineages associated with healthcare-associated infections in different countries and to monitor their emergence, dispersal and decline in different settings. However, conventional typing methods have serious limitations when used to investigate the finer details of infection outbreaks.[1]

* Corresponding author. Address: Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Eastern Road, Brighton BN2 5BE, UK. Tel.: +44 (0) 1273 696955x7516.
E-mail address: jrprice@doctors.org.uk (J.R. Price).

0195-6701/$ – see front matter © 2012 The Healthcare Infection Society. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jhin.2012.10.003

Conventional methods are often insufficiently discriminatory to 'rule out' suspected transmission events in the absence of additional epidemiological information, since representatives of the same type are often found all around the world. For example, when investigating a cluster of MRSA cases in a healthcare setting where a particular MRSA strain has become endemic, conventional methods would most likely fail to distinguish between those unlinked cases that happen to belong to the same lineage and other cases that are truly connected via recent transmission. Conversely, when the typing method depends on phenotypic characterization such as antibiotic susceptibility profiles, isolates that are truly linked via transmission events might not be recognized as such because the characteristics they measure are encoded on mobile genetic elements.[2,3]

Whole genome sequencing (WGS) allows clinical isolates of *S. aureus* to be compared with each other and with reference sequences across time and space, down to a resolution of a single nucleotide difference.[4] This enhances our knowledge of the population structure of *S. aureus*, allowing greater precision in describing and defining the different lineages, provides insights into the evolutionary history of lineages and offers the potential for an outbreak investigator to determine unambiguously the relatedness of isolates. By comparing the relatedness of two isolates with the resolution offered by whole genome sequences, and using estimates of the genome's mutation rate, it is possible to estimate the time elapsed since their descent from a common ancestor with precision.[4] Combined with epidemiological data, such as information on dates of admission to hospital, it is then possible to draw inferences about the probability that a transmission event occurred or not, with sufficient accuracy to direct better targeting of infection control resources.[5] In fact, refinements in genealogical approaches to sequence data analysis offer the prospect of being able to make such inferences even in the absence of supporting epidemiological information (X. Didelot, D. Eyre, M. Cule, *et al.*, unpublished data).

These properties give WGS the potential to revolutionize infection control practice on local, national and international scales. Sequence data interpreted in the context of epidemiological surveillance data will allow the rapid detection of new emerging strains. At a local level awareness of patterns of transmission and prompt outbreak recognition (as early as the detection of the first secondary case) will permit more effective interventions to be instigated. At an individual patient level, the genetic basis for phenotypic characteristics of relevance to clinical case management, such as antibiotic susceptibility and virulence factors, may also be determined using the same method. However, there remain major hurdles to be overcome to translate WGS from a research tool into clinical practice. These include cost, turnaround time and bioinformatic analysis. Current rates of progress suggest that most of these difficulties should be overcome in the fairly near future.

Evolutionary history and population structure of *Staphylococcus aureus*

Staphylococcus aureus is a human commensal that is carried by approximately one-third of the general population.[6] Sites for colonization include the anterior nares, the throat, the axilla, and the perineum.[7] Different patterns of carriage have been described: people may be persistent carriers (20%), intermittent carriers (30%), or non-carriers (50%).[6] Prior asymptomatic carriage is a significant risk factor for the development of invasive disease (relative risk: 16.7; 95% confidence interval: 8.6–32.5), and recent acquisition is associated with an increased risk of poor medical outcome.[8,9] Overall, hospital-acquired *S. aureus* bacteraemia is associated with a mortality rate of 24%.[10] This capacity for superficial carriage and aggressive infection underlies the importance of *S. aureus* as a nosocomial pathogen.

The circular genome of *S. aureus* is composed of about 2.8 million nucleotides and is about one thousand times smaller than the human genome (about three billion nucleotides).[11] Most of the genome (the core genome) is composed of genes present in all strains that encode proteins involved in fundamental functions such as cellular metabolism, growth and replication. About 10% of the genome consists of sets of genes that vary between different lineages and is designated the 'core variable genome'.[12] Between 10% and 20% of the genome consists of 'mobile genetic elements'; regions that are gained and lost by organisms at high frequencies (lateral gene transfer) and which often encode virulence factors and resistance genes.

Staphylococcus aureus has a markedly clonal population structure.[13–16] Most disease-causing isolates belong to a small number of lineages or clonal complexes. Indeed most of the strains that colonize humans belong to one of ten dominant lineages.[12] Within this structure differences in the core genome occur as a result of point mutation and to a lesser extent through recombination events.[13] The necessity to disentangle the evolutionary signals caused by mutation and recombination is a common issue in any sequence-based analysis of bacteria, and statistical methods are being developed to deal with this difficulty by explicitly accounting for the role of recombination.[17–19] The conserved genomic structure of the successful lineages has been explained by the presence of enzymic restriction modification systems that limit acquisition of foreign DNA.[20] In recent decades, two epidemic lineages, designated EMRSA-15 and EMRSA-16 (originally defined by phage typing patterns), became the dominant healthcare-associated strains in the UK.

What methods are currently used to type *S. aureus*?

Phenotypic typing methods that exploit variations in observable strain characteristics such as antibiotic susceptibility pattern, phage typing profile, and serotype are relatively inexpensive but poorly discriminatory.[21,22] Among the most widely used molecular typing methods are multi-locus sequence typing (MLST), staphylococcal protein A (*spa*)-typing, pulsed-field gel electrophoresis (PFGE) and multi-locus variable number tandem repeat analysis (MLVA). MLST is based on sequence variation in housekeeping genes.[23,24] MLST classifies *S. aureus* strains into groups that reflect phylogeny, allowing the study of population structure and evolutionary history.[13,25,26] Different MLST sequence types can be grouped into clonal complexes (CC) on the basis that they share some of the seven (or more) loci.[27] By contrast, *spa*-typing is based on the highly variable X-region of a single gene (*spa*) that encodes protein A.[28] Concordance between MLST and *spa*-typing is high so that *spa*-typing is now used widely for MRSA typing.[29] In PFGE,