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 © 2012 The Healthcare Infection Society. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
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, enzymes are used to cleave DNA into fragments of different sizes which form strain-specific patterns when separated by gel electrophoresis.[30] Compared with MLST and *spa*-typing, PFGE is relatively good at resolving differences between strains and many authors have promoted it as a tool for local outbreak investigation.[31,32] MLVA uses multiplex polymerase chain reaction (PCR) to amplify known genetic loci containing varying length random repeats, which again are separated into strain-specific patterns by electrophoresis.[33] MLVA has been shown to have powers of discrimination similar to those of PFGE, and MLVA results are highly concordant with those generated by MLST and *spa*-typing.[33–37]

The determinant of meticillin resistance in MRSA, the *mec*A gene, is part of a mobile genetic element, the staphylococcal cassette chromosome element (SCC*mec*).[38] Hence, MRSA strains may be classified and typed according to the composition of SCC*mec*.[39,40] Matrix-assisted laser desorption/ionization—time of flight (MALDI-TOF) allows rapid identification of organisms through molecular mass profiling of protein biomarkers.[41] MALDI-TOF technology has recently been evaluated as a typing tool and has been shown to allow categorization of MRSA into lineages based on diversity of protein profiles.[42]

Why sequence the whole genome to type *S. aureus*?

The current diversity of methods for characterizing *S. aureus* results in inconsistencies in nomenclature and hinders communication. Table I shows the differences in cost, practicalities and achievable resolutions of conventional typing techniques and WGS. A unified typing scheme linked to a freely accessible universal database would be ideal. WGS resolves differences between isolates down to the level of individual nucleotides. Because of poor resolution, conventional methods generate inconclusive results. For example, when two isolates are indistinguishable it does not mean that they are identical. WGS makes it possible to determine when sequences really are identical, and, if not, to state exactly by how much they differ. This allows inferences to be made about the relationships between strains not just in terms of global population structure but also in terms of local patterns of transmission.

Conventional typing methods fail to reveal the fine details of the genetic differences that accumulate between strains as *S. aureus* diversifies, as DNA polymerases make copying mistakes, point mutations occur, recombination events take place and mobile elements are gained and lost. Even when used in combination, conventional methods are limited in their potential to distinguish between isolates within a major lineage. WGS offers a portable, reproducible method to detect the smallest of genetic differences.

Whole genome sequencing of *Staphylococcus aureus*: practicalities

Platforms

Whole genome sequencing refers to the construction of the complete nucleotide sequence of a genome. In 1977, Sanger *et al.* developed a method that revealed the first complete genome sequence of a virus.[43] This 'first generation sequencing' used capillary—electrophoresis methods to sequence DNA fragments, a process that was expensive and slow.[44] The terms 'second generation' or 'next generation' sequencing (NGS) refer to methods that parallelize the sequencing process, thus dramatically lowering costs and increasing capacity. NGS first became available in 2004 and a number of different sequencing platforms are currently commercially available.[45] These include the Illumina Platform (Illumina, San Diego, CA, USA), 454 genome sequencer (Roche Applied Science, Rotkreuz, Switzerland), SOLiD platform (Life Technologies, Applied Biosystems, Carlsbad, CA, USA) as well as the first bench-top machines such as the MiSeq (Illumina) and IonTorrent platforms (Life Technologies) which are anticipated to be among the first platforms to be routinely implemented for clinical applications. Late in 2012 it is anticipated that the GridION platform (Oxford Nanopore, Oxford, UK) will enter the marketplace.[46]

Genome assembly

Sequencing machines generate thousands of small DNA sequences, called reads, which can be between 40 and 1000 base pairs in length depending on the technique used. Each read represents the sequence of a small fraction of the

Table I
Comparison of the *Staphylococcus aureus* typing techniques

Technique	Set-up cost	Cost per isolate	Current availability	Time to results	Data analysis	Data transferability	Common nomenclature	Level of resolution
PFGE	Low	£4–7	Local and reference laboratories	2–3 days	Minimal	Limited	USA type (USA-)	Lineage
MLVA	Low	£3–5	Research and reference laboratories	24 h	Minimal	Limited	MLVA complex (Sa-)	Lineage
MLST	High	£20	Research and reference laboratories	Days	Moderate	Yes – widely	Sequence type (ST-)	Lineage
spa	High	£3–5	Local and reference laboratories	24 h	Moderate	Yes – widely	Spa-type (t-)	Lineage
WGS	High	~£100	Research laboratories	Real time[a]	High	Being addressed	To be determined	Base pair

PFGE, pulsed-field gel electrophoresis; MLVA, multi-locus variable-number tandem-repeat analysis; MLST, multi-locus sequence typing; spa, *spa*-typing; WGS, whole genome sequencing.

[a] Third generation sequencing platforms.

genome. The reads overlap so that each position of the genome is included in several reads. This number is called 'the coverage'. To be useful, the reads need to be assembled into the whole genome sequence. This can be done in one of two ways; either by comparison with a previously sequenced 'reference' strain (mapping-based assembly), or by piecing the reads together on the basis of their overlaps (*de novo* assembly).[47,48] Mapping-based assembly produces results that are easier to interpret, but are dependent upon the choice of reference genome, so that sequences not present in the reference genome remain unmapped. In principle, *de novo* assembly can recover the whole genome, but in practice it returns several assembled regions (called contigs) whose further assembly tends to remain unresolved. In either case, longer read lengths and higher coverage make assembly easier and more accurate.[49] Storing the reads for a single *S. aureus* genome requires about 1 GB of storage space; roughly equivalent to 10 music albums in MP3 format. The term high-throughput sequencing is used to refer to newer sequencing technologies that can generate sequence data faster and more economically than previous platforms. Consequently, sequencing platforms are rated according to their throughput speeds, set-up and processing costs and length of reads produced.[45] A medium-sized sequencing facility can process thousands of bacterial isolates in a year. New platforms (Oxford Nanopore and Ion Proton) offer the promise of projected cost reductions to the order of US $1,000 for a whole human genome, sequenced in 15 min to 2 h. Such performance would translate into sequencing costs for bacteria of about US $1 per whole genome.

Limitations

Current approaches to whole genome sequencing have inherent limitations. At present, using high throughput sequencing machines takes several weeks to generate a DNA sequence due to complex sample preparation, large numbers of scanning/washing cycles and reliance on PCR amplification of DNA templates. Furthermore the reads produced are relatively short which can make genome assembly challenging. At the time of writing it is important to realize that the limitations of mapping-based and *de novo* assembly techniques for short read sequences mean that the term 'whole genome sequence' actually refers to the 80–90% of the entire genome that is actually rendered visible by those techniques. The introduction of new platforms that generate much longer reads will make visible the remaining 'concealed' sequences to routine analysis.

Third generation sequencing

A major recent milestone has been the development of technology to allow sequencing of genomes to be performed in real-time. This is termed 'third generation sequencing' (TGS).[44] Instead of fragmenting and reconstructing read sequences, TGS platforms allow direct observation of the DNA polymerase enzyme while it constructs a strand of DNA. By not having to pause between individual base identification, and by taking advantage of the enzyme's speed in adding nucleotide bases to the growing chain, the throughput is greatly increased and the reads are considerably longer.[50] This lowers costs and aids read assembly. Currently available TGS technologies have limitations: error rates are high at 5% and the output data are formatted differently from those yielded by Sanger *et al.* and NGS.[44] These issues are being addressed but, for the time being, second generation machines are the platforms of choice. Table II shows a comparison of different sequencing technologies.

Analysis

The first complete *S. aureus* genome was published in 2001.[11] As recently as 2008 only 12 complete genomic sequences of *S. aureus* were available in the public domain.[51] At the time of writing (July 2012) there are 178 fully annotated *S. aureus* genomes publically available in the National Centre for Biotechnology Information (NCBI) Reference Sequences (RefSeq) database.[52] The first few genomes of *S. aureus* to be sequenced were fully annotated with the position of each gene and probable function stated when known. The wealth of information contained in these annotations makes these genomes desirable choices for use as references. With higher sequencing throughput now achievable, much larger numbers of genomes can be sequenced and the emphasis has shifted from annotation to data analysis. The depth of analysis required depends on the question being asked. For example, full annotations as described above are not necessary for determining the relatedness of isolates in transmission and outbreak investigations. In such situations comparisons are required between strains in terms of single nucleotide variants (SNVs) and insertion or deletion events.

Molecular clock calibration

As a bacterial population evolves and diversifies from a common ancestor its constituent members accumulate differences in their genome sequences. By comparing sequence data it is possible to infer phylogenetic relationships. A key requirement when interpreting such data is to be able to calibrate the molecular clock, i.e. estimate the rate at which mutations accumulate during the evolution of a genome. With knowledge of the molecular clock, it is possible directly to convert a number of differences between two genomes into the length of evolutionary time that has elapsed since they last

Table II
Comparison of sequencing technologies

	Sequencing technology platform		
	First generation	Second generation	Third generation
Resolution	Average of multiple DNA copies	Average of multiple DNA copies	Single molecule
Read length generated	800–1000 bp	<400 bp	1000–10,000 bp
Financial cost per base	High	Low	Moderate
Financial cost per run	Low	High	Low
Sample preparation	Moderate	Complex	Variable
Time to result	Hours	Days	Minutes to hours

shared a common ancestor. This is an area where there is still some uncertainty. The molecular clock may not tick at a constant speed in different settings (environment, commensal and in blood for example), in different hosts or even across lineages. Harris *et al.* used a linear regression between isolation dates and root-to-tip distances in a phylogenetic reconstruction to estimate an average rate of 3.3×10^{-6} mutations per site per year in *S. aureus*.[53] Over the whole genome, this represents an average of 9.2 mutations per genome per year. A similar rate was estimated by Young *et al.* for carriage isolates sampled serially from the same patients using the Bayesian phylogenetic method BEAST.[4,54] This rate is high enough to envision the application of WGS to the investigation of patient-to-patient transmission.

Hurdles and expectations

Besides the technical challenge of specimen preparation and sequencing, even greater hurdles to putting WGS into routine practice lie in how to perform bioinformatic analysis and optimize information technology infrastructure for data storage and exchange. The implementation of WGS to near real-time infection control practice would require expertise in technical issues of sequencing plus collaboration with specialist teams of bioinformaticians and analysts. However, it seems reasonable to predict that bench-top sequencers will become increasingly accessible and user-friendly. A secure central database could be developed for exchange of data that would allow interpretation of sequence data within a given epidemiological context. Being able to cope with huge volumes of data will be a challenge. Furthermore, with the rapid evolution of technologies, flexibility for forward and backward integration will be required; to be compatible with previous typing techniques and to support output data from newly evolving platforms.

Translating WGS of *S. aureus* from research into practice

Harris *et al.* used WGS to assess the diversity of an MRSA lineage across time, across continents and within the individual healthcare setting.[53] They provided estimates of mutation rate, demonstrated geographical clustering of isolates and showed how the difference in number of mutations can be used to distinguish transmission from endemic infection. Using this method, they found evidence suggestive of hospital-based transmission of MRSA between patients staying in the same hospital in Thailand. Another research group used WGS to demonstrate the emergence in a localized hospital setting of a new MRSA clone within ST36 that was indistinguishable from other members of the same sequence type by conventional methods (R. Miller, J. Price, E. Batty *et al.*, unpublished data). WGS appears to be uniquely well suited to detect new pathogen variants as they emerge, track their spread and direct effort to formulate infection control strategies. WGS is increasing our understanding of the micro-evolution of *S. aureus* during long-term carriage and its relationship with disease states. Young *et al.* used WGS to investigate serial isolates of *S. aureus* during long-term carriage and later during an episode of invasive disease.[4] This showed a mutation rate (roughly one mutation per genome every seven weeks) during long-term carriage similar to that estimated by Harris *et al.* for the mutation rates within a single lineage, ST239. However, blood culture isolates differed from the ancestral colonizing strain by eight mutations. As a consequence, the invasive strain's genome incorporated stop codons that encoded a truncated regulatory protein, thereby providing potential clues about the nature of the changes associated with pathogenesis. McAdam *et al.* applied WGS to 87 isolates to show how EMRSA-16 emerged 35 years ago, and spread in the UK by transmission from hospitals in large cities to regional locations.[55]

Rapid desktop sequencers are becoming affordable for use in routine diagnostic laboratories. Eyre *et al.* showed that using desktop machines it was possible to 'rule in' isolates from a suspected outbreak as being truly related at the core genome level despite conflicting evidence of differing antibiotic resistance profiles within a five-day turnaround period.[56] This demonstrated the unreliability of typing tools based on phenotypic characteristics encoded by genes present in mobile genetic elements. Conversely, this same study demonstrated that WGS could be used to 'rule out' cases during an outbreak investigation even when all of the isolates examined were from the same geographical area and belonged to the same unusual *spa* type. Köser *et al.* used desktop sequencing to characterize a suspected outbreak on a neonatal intensive care unit of an MRSA strain with an unusual antibiotic susceptibility pattern.[57] Phylogenetic reconstructions of seven MRSA isolates highlighted tight genetic clustering of suspected strains and confirmed the outbreak.

Experience with other pathogens

Whole genome sequencing has also been applied to other major pathogens in order to improve our understanding of their epidemiology. He *et al.* used WGS to demonstrate that the strains of *Clostridium difficile* associated with disease emerged from multiple lineages.[58] This contradicted the received wisdom that bacterial pathogens arise from a single lineage as a result of gaining genetic properties through gene transfer or mutation. Didelot *et al.* used WGS to investigate cases of *C. difficile* infection that appeared to be linked according to conventional typing results (X. Didelot, D. Eyre, M. Cule *et al.*, unpublished data). The resolution provided by WGS made it possible to 'rule out' transmission for the majority of cases, thereby challenging the assumption that *C. difficile* infection is usually acquired as a result of direct transmission from symptomatic patients. WGS has already been applied during the investigation of some high-profile outbreaks. The Asian origins of the Haitian outbreak of cholera that killed several thousand people in 2010–2011 were determined through WGS of the outbreak strain of *Vibrio cholera*.[59] In 2003 WGS was used to characterize the severe acute respiratory syndrome virus, showing it to be a previously unrecognized coronavirus.[60] In 2011 sequencing technology was used to characterize a Shiga-toxin producing *Escherichia coli* O104:H4 strain that affected nearly 4000 people and caused 47 deaths in Europe and North America. Sequencing rapidly confirmed clonal expansion from a common ancestor and allowed identification of virulence factors.[61] Gardy *et al.* used WGS to evaluate a Canadian outbreak of tuberculosis.[62] Outbreak isolates were indistinguishable by conventional genetic fingerprinting methods (MIRU-VNTR), and epidemiological information suggested a probable point source. WGS of isolates revealed that two distinct lineages were circulating simultaneously. Evolutionary analysis using historical isolates from the region showed that both lineages were present long before the dual outbreak was recognized. WGS was able to 'rule in' four historical tuberculosis cases that had not previously been recognized as part of the outbreak. Hence, WGS provides a tool for optimizing contact tracing resources.

Potential for WGS and possible consequences

Whole genome sequencing has the potential to improve our understanding of phylogeny, transmission and pathogenesis and to provide information that will enhance surveillance, guide outbreak investigation and improve disease management. When put into practice WGS could revolutionize the principles of reference microbiology and taxonomy.

Transmission and outbreaks

In the UK in recent decades, strategies to prevent nosocomial acquisition of *S. aureus* have been focused on the prevention and control of MRSA infection. This focus is partly explained by the ease in being able to differentiate MRSA from sensitive strains and the recognition of the burden of disease caused by MRSA. However, the lack of resolution offered by conventional typing methods has hampered our ability to form a sound evidence base from which we might optimize our practice. WGS will allow us to 'rule in' and 'rule out' links between otherwise indistinguishable isolates during the course of an investigation and give us the tool that we need to evaluate infection control practice.

Epidemiological surveillance

WGS offers the prospect of a typing system that could be applied globally to *S. aureus* surveillance. Genome sequencing of isolates in local laboratories would permit real-time generation of locally and nationally relevant epidemiological data, allowing strains to be tracked, their relative importance evaluated and control efforts to be meaningfully targeted. Furthermore, the higher level of discrimination will permit the detection and monitoring of newly emerging strains. Linked with clinical surveillance data, this could provide an early warning system.

Phylogeny

As the WGS knowledge base grows, a much better understanding of staphylococcal population structures will be developed. Genomic comparisons over time and space of isolates will provide clearer insights into the evolutionary history of *S. aureus*. WGS will help us to understand the biological significance of genetic variation in bacteria. It will enhance our understanding of the effects of selective pressures (e.g. antibiotic exposure) on bacterial populations and the biology of pathogenesis.

Phenotypic predictions

Since a whole genome sequence contains the sum of the genetic information that determines an organism's phenotype, the current plethora of phenotypic tests could in principle be substituted by WGS. The relationship between genotype and differences in phenotype in clinically relevant strains can be investigated. For example, desktop sequencing has been shown to allow rapid identification of mutations associated with unusual antibiotic susceptibility patterns, variations in strain growth rates, and the development of small colony variants.[57] As long as the association between genotype and phenotype are well understood, *in silico* interrogation of a WGS could be used to predict antibiotic resistance and virulence and effectively guide management earlier. For example using a 'basic local alignment search tool' (BLAST).[63] An automated framework exists (BIGSdb) for simultaneously querying several genes from several genomes.[64] This approach also provides a bridge between WGS and previous methods based on sequencing such as *spa*-typing or MLST since the loci targeted by these methods can be extracted from the whole genomes. PFGE and MLVA types should also be predicted directly from WGS data. This will ensure that the wealth of information that has been accumulated using these previous typing methods does not go to waste. Genome-wide association studies to uncover the genetic basis of severe disease phenotypes could be undertaken.[65] In addition to known associations there is the potential to identify previously unknown virulence genes associated with clinical phenotypes. This is highlighted by the discovery of more than 70 novel candidate virulence factors that were identified when the first two *S. aureus* genomes were sequenced.[11] A recent comparison of three *S. aureus* whole genomes isolated sequentially over a period of 26 months from the same cystic fibrosis patient revealed several genes under diversifying selection which could indicate that they play an essential role for chronic infection.[66] As sequencing of the human genome becomes more routine it will be possible to investigate in great detail the genetic basis for the interplay between host and commensal or pathogen. It might for example become possible to determine the reasons why some people become persistent carriers while others do not. It may become possible to predict susceptibility to invasive staphylococcal infection and to target preventive measures accordingly.

With the continuing fall in sequencing costs it is conceivable that sequence-based methodologies might replace traditional methods as frontline diagnostic tests. There would then be a need to develop systems to check the reliability of WGS outputs as predictors of phenotypic characteristics such as antibiotic susceptibility. Such needs could define one of the core functions of future reference laboratories.

Consequences for identification and nomenclature

Bacteria are identified through a process of comparison of the features exhibited by the test isolate with those of the type culture that has been designated and archived by the author of the species. For convenience, when identifying a bacterium, most microbiologists make do with reference to a published account of the distinguishing features of a species rather than to make a direct comparison with the actual type culture. Widespread access to WGS makes it possible to extend the concept of a designated 'type culture' for every species to that of a universally accessible 'type sequence'. Identification will become more reliable. It will be possible to define the limits of genetic variation within a species, to define lineages more precisely and to define new species with greater accuracy.

Conclusion

Whole genome sequencing has the potential to replace a multitude of diagnostic and reference tests. For the purposes of outbreak investigation and surveillance, WGS of *S. aureus* yields results of greater utility than can be achieved using conventional techniques. Furthermore, WGS data may be used to predict phenotypic characteristics. WGS is also applicable to studies that seek to address questions about the biology of *S. aureus* and the relationship between pathogen and host.

At the time of writing, WGS of *S. aureus* is technically difficult and is the preserve of the research community. Compared with many conventional approaches, WGS is still relatively expensive. The bioinformatic processing and interpretation of sequence data are particularly challenging. Also the issues around storing and transmitting cannot be underestimated.

Despite these challenges, the steady decline in sequencing costs and the increasing ease of use of sequencing machines and their data outputs suggest that WGS will eventually replace a whole suite of current test strategies. It seems clear that WGS will revolutionize our understanding of *S. aureus* and our ability to manage it as a pathogen.

Conflict of interest statement

None declared.

Funding sources

J.R.P. and X.D. were supported by the UKCRC Modernising Medical Microbiology Consortium, which is funded under the UKCRC Translational Infection Research Initiative supported by the Wellcome Trust (Grant 087646/Z/08/Z) and by the Medical Research Council, Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the Department of Health (Grant G0800778).

References

1. Harris SR, Clarke IN, Seth-Smith HMB, *et al*. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nat Genet* 2012;**44**:413–419.
2. Zhu W, Clark NC, Mcdougal LK, Hageman J, Mcdonald LC, Patel JB. Vancomycin-resistant *Staphylococcus aureus* isolates associated with Inc18-like vanA plasmids in Michigan. *Antimicrob Agents Chemother* 2008;**54**:452–457.
3. Zhu W, Murray PR, Huskins WC, *et al*. Dissemination of an Enterococcus Inc18-like vanA plasmid associated with vancomycin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* 2010;**52**:4314–4320.
4. Young BC, Golubchik T, Batty EM, *et al*. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci USA* 2012;**109**:4550–4555.
5. Walker AS, Eyre DW, Wyllie DH, *et al*. Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. *PLoS Med* 2011;**9**:e1001172.
6. Wertheim HF, Melles DC, Vos MC, *et al*. The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis* 2005;**5**:751–762.
7. Williams RE. Healthy carriage of *Staphylococcus aureus*: its prevalence and importance. *Bacteriol Rev* 1963;**27**:56–71.
8. Robicsek A, Suseno M, Beaumont JL, Thomson Jr RB, Peterson LR. Prediction of methicillin-resistant *Staphylococcus aureus* involvement in disease sites by concomitant nasal sampling. *J Clin Microbiol* 2008;**46**:588–592.
9. Wertheim HF, Vos MC, Ott A, *et al*. Risk and outcome of nosocomial *Staphylococcus aureus* bacteraemia in nasal carriers versus non-carriers. *Lancet* 2004;**364**:703–705.
10. Thwaites GE, United Kingdom Clinical Infection Research Group (UKCIRG). The management of *Staphylococcus aureus* bacteremia in the United Kingdom and Vietnam: a multi-centre evaluation. *PLoS One* 2010;**5**:e14170.
11. Kuroda M, Ohta T, Uchiyama I, *et al*. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* 2001;**357**: 1225–1240.
12. Lindsay JA, Moore CE, Day NP, *et al*. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *J Bacteriol* 2006;**188**:669–676.
13. Feil EJ, Cooper JE, Grundmann H, *et al*. How clonal is *Staphylococcus aureus*? *J Bacteriol* 2003;**185**:3307–3316.
14. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 2009;**3**:199–208.
15. Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol* 2006;**6**:97–112.
16. Hanage WP, Fraser C, Spratt BG. The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol* 2006;**239**:210–219.
17. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 2007;**175**:1251–1266.
18. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. *Trends Microbiol* 2010;**18**:315–322.
19. Didelot X, Lawson D, Darling A, Falush D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* 2010;**186**:1435–1449.
20. Waldron DE, Lindsay JA. Sau1: a novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into *Staphylococcus aureus* and between *S. aureus* isolates of different lineages. *J Bacteriol* 2006;**188**:5578–5585.
21. Rossney AS, Coleman DC, Keane CT. Antibiogram-resistogram typing scheme for methicillin-resistant *Staphylococcus aureus*. *J Med Microbiol* 1994;**41**:430–440.
22. Weller TM. Methicillin-resistant *Staphylococcus aureus* typing methods: which should be the international standard? *J Hosp Infect* 2000;**44**:160–172.
23. Maiden MC, Bygraves JA, Feil E, *et al*. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 1998;**95**:3140–3145.
24. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J Clin Microbiol* 2000;**38**:1008–1015.
25. Robinson DA, Enright MC. Multilocus sequence typing and the evolution of methicillin-resistant *Staphylococcus aureus*. *Clin Microbiol Infect* 2004;**10**:92–97.
26. Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, Spratt BG. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc Natl Acad Sci USA* 2002;**99**:7687–7692.
27. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of

related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* 2004;**186**:1518–1530.

28. Shopsin B, Gomez M, Montgomery SO, *et al*. Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. *J Clin Microbiol* 1999;**37**:3556–3563.
29. Strommenger B, Braulke C, Heuck D, *et al*. spa typing of *Staphylococcus aureus* as a frontline tool in epidemiological typing. *J Clin Microbiol* 2008;**46**:574–581.
30. Ichiyama S, Ohta M, Shimokata K, Kato N, Takeuchi J. Genomic DNA fingerprinting by pulsed-field gel electrophoresis as an epidemiological marker for study of nosocomial infections caused by methicillin-resistant *Staphylococcus aureus*. *J Clin Microbiol* 1991;**29**:2690–2695.
31. McDougal LK, Steward CD, Killgore GE, Chaitram JM, McAllister SK, Tenover FC. Pulsed-field gel electrophoresis typing of oxacillin-resistant *Staphylococcus aureus* isolates from the United States: establishing a national database. *J Clin Microbiol* 2003;**41**: 5113–5120.
32. Faria NA, Carrico JA, Oliveira DC, Ramirez M, de Lencastre H. Analysis of typing methods for epidemiological surveillance of both methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* strains. *J Clin Microbiol* 2008;**46**:136–144.
33. Sabat A, Krzyszton-Russjan J, Strzalka W, *et al*. New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. *J Clin Microbiol* 2003;**41**:1801–1804.
34. Malachowa N, Sabat A, Gniadkowski M, *et al*. Comparison of multiple-locus variable-number tandem-repeat analysis with pulsed-field gel electrophoresis, spa typing, and multilocus sequence typing for clonal characterization of *Staphylococcus aureus* isolates. *J Clin Microbiol* 2005;**43**:3095–3100.
35. Pourcel C, Hormigos K, Onteniente L, Sakwinska O, Deurenberg RH, Vergnaud G. Improved multiple-locus variable-number tandem-repeat assay for *Staphylococcus aureus* genotyping, providing a highly informative technique together with strong phylogenetic value. *J Clin Microbiol* 2009;**47**:3121–3128.
36. Tenover FC, Vaughn RR, McDougal LK, Fosheim GE, McGowan JE Jr. Multiple-locus variable-number tandem-repeat assay analysis of methicillin-resistant *Staphylococcus aureus* strains. *J Clin Microbiol* 2007;**45**:2215–2219.
37. Schouls LM, Spalburg EC, van Luit M, *et al*. Multiple-locus variable number tandem repeat analysis of *Staphylococcus aureus*: comparison with pulsed-field gel electrophoresis and spa-typing. *PLoS One* 2009;**4**:e5082.
38. Katayama Y, Ito T, Hiramatsu K. A new class of genetic element, Staphylococcus cassette chromosome mec, encodes methicillin resistance in *Staphylococcus aureus*. *Antimicrob Agents Chemother* 2000;**44**:1549–1555.
39. Chongtrakool P, Ito T, Ma XX, *et al*. Staphylococcal cassette chromosome mec (SCCmec) typing of methicillin-resistant *Staphylococcus aureus* strains isolated in 11 Asian countries: a proposal for a new nomenclature for SCCmec elements. *Antimicrob Agents Chemother* 2006;**50**:1001–1012.
40. International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG – SCC). Classification of staphylococcal cassette chromosome mec (SCCmec): guidelines for reporting novel SCCmec elements. *Antimicrob Agents Chemother* 2009;**53**:4961–4967.
41. Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem* 1991;**63**:1193A–1203A.
42. Wolters M, Rohde H, Maier T, *et al*. MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant *Staphylococcus aureus* lineages. *Int J Med Microbiol* 2011;**301**:64–68.
43. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;**74**:5463–5467.
44. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;**19**:R227–R240.
45. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet* 2010;**11**:31–46.
46. Branton D, Deamer DW, Marziali A, *et al*. The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008;**26**: 1146–1153.
47. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**:443–451.
48. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–829.
49. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res* 2010;**20**: 1165–1173.
50. Eid J, Fehr A, Gray J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**:133–138.
51. Holden MTG, Lindsay JA. Whole genomes: sequence, microarray and systems biology. In: Lindsay J, editor. *Staphylococcus: molecular genetics*. 1st ed. Caister, UK: Caister Academic Press; 2008. p. 1–28.
52. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012;**40**:D130–D135.
53. Harris SR, Feil EJ, Holden MT, *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;**327**:469–474.
54. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007;**7**:214.
55. McAdam PR, Templeton KE, Edwards GF, *et al*. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. *Proc Natl Acad Sci USA* 2012;**109**:9107–9112.
56. Eyre DW, Golubchik T, Gordon NC, *et al*. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open* 2012;**2**:e001124.
57. Köser CU, Holden MT, Ellington MJ, *et al*. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012;**366**:2267–2275.
58. He M, Sebaihia M, Lawley TD, *et al*. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci USA* 2010;**107**:7527–7532.
59. Chin C-S, Sorenson J, Harris JB, *et al*. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 2011;**364**:33–42.
60. Marra MA, Jones SJ, Astell CR, *et al*. The genome sequence of the SARS-associated coronavirus. *Science* 2003;**300**:1399–1404.
61. Cheung MK, Li L, Nong W, Kwan HS. 2011 German *Escherichia coli* O104:H4 outbreak: whole-genome phylogeny without alignment. *BMC Res Notes* 2011;**4**:533.
62. Gardy JL, Johnston JC, Ho Sui SJ, *et al*. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011;**364**:730–739.
63. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–3402.
64. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;**11**:595.
65. Falush D, Bowden R. Genome-wide association mapping in bacteria? *Trends Microbiol* 2006;**14**:353–355.
66. McAdam PR, Holmes A, Templeton KE, Fitzgerald JR. Adaptive evolution of *Staphylococcus aureus* during chronic endobronchial infection of a cystic fibrosis patient. *PLoS One* 2011;**6**:e24301.