A virulence-associated gene microarray: a tool for investigation of the evolution and pathogenic potential of Staphylococcus aureus

Nicholas A. Saunders,1 Anthony Underwood,1 Angela M. Kearns2 and Gillian Hallas1

Genomics Proteomics and Bioinformatics Unit1 and Laboratory of Healthcare Associated Infection2, Health Protection Agency (Colindale), 61 Colindale Avenue, London NW9 5HT, UK

An oligonucleotide probe microarray for investigation of the evolution of epidemic Staphylococcus aureus strains has been constructed. The array comprises 383 probes based on virulence-associated genes present in four key strains. Twelve strains including seven for which the complete chromosomal nucleotide sequence was available were tested on the array. Twenty-six per cent of the probes were able to differentiate between strains to give a minimum of two gene differences between pairs. A gene difference distance tree based on the array data had approximately the same topology as one prepared using concatenated MLST sequences. Differences in the topologies of these trees were found to indicate that large-scale recombination events had occurred during the evolution of the species. One such occurrence appears to have been a key event in the genesis of the EMRSA-15 clone (ST22) that currently represents the most prevalent methicillin-resistant S. aureus (MRSA) in the UK.

INTRODUCTION

Staphylococcus aureus is an important human pathogen responsible for bacterial infections seen both in the community and in hospitals. Over the last decades the emergence of multi-drug resistant strains has greatly increased the economic and health importance of staphylococcal infection (Chambers, 2001; Stefani & Varaldo, 2003; Wenzel, 1982). S. aureus is a highly adapted and extremely successful colonizer of the human nasopharynx, other mucosal surfaces and skin (Kluymans et al., 1997; Williams, 1963). Colonization is generally asymptomatic but occasionally results in disease with a wide variety of signs and symptoms. These range from mild self-limiting infections of the skin to fulminating septicemia (Lowy, 1998; Projan & Novick, 1997). Staphylococcal disease appears to result from an imbalance between the bacterium and the immune system and is associated with tissue invasion or unchecked growth. S. aureus is able to maintain its colonist status by producing a variety of proteins that interact with host-cell components (Baba et al., 2002; Kuroda et al., 2001). These proteins are described as virulence-associated and have functions such as extracellular matrix binding and host cell lysis. Expression of the S. aureus virulence-associated proteins is very carefully controlled by means of a network of regulatory genes including the agr and sar operons (Chien & Cheung, 1998; Sabersheikh & Saunders, 2004). Strain variation in these regulatory mechanisms may be an important factor influencing the switch from colonization to disease. Such variations may also contribute towards the success of the species by helping to minimize the emergence of immunity to essential S. aureus-specific antigens.

Genome sequencing has shown that the species maintains a large number of virulence-associated genes but that not all are carried by each individual strain (Baba et al., 2002; Holden et al., 2004; Kuroda et al., 2001). Pairwise comparisons show as much as 20% variation in the gene inventory of different S. aureus with most of the difference contributed by virulence-associated genes (Fitzgerald et al., 2001). This ensures that strains present a variable antigenic profile to the host and may limit the effective immunity of the population to colonization.

S. aureus strains carry a range of genes that confer resistance to a growing list of antibiotics and vary greatly in their susceptibility profiles. The level of resistance conferred on any individual strain by a particular gene is now known to depend not only on the presence of related genes but also on the degree of gene expression that occurs. In the same way the tendency of strains to make the transition from benign colonizer to pathogen is now thought to be related to the presence of some, but not all, of the virulence-associated genes maintained by the species (Peacock et al., 2002). Other associations between pathogenicity, the genome and gene

Abbreviations: EMRSA, epidemic methicillin-resistant S. aureus; MRSA, methicillin-resistant S. aureus; MSSA, methicillin-sensitive S. aureus.

A supplementary table listing the genes included on the microarray is available in Microbiology Online.
expression are suspected but remain unproven. For example, it is thought that epidemic methicillin-resistant \textit{S. aureus} (EMRSA) strains have a higher cell density threshold for the switch from the colonization to the persistence phase than sporadic strains (Papakyriacou \textit{et al}., 2000). Surface-expressed proteins are dominant during colonization while secreted proteins become more important during persistence (Novick \textit{et al}., 1993; Sabersheik & Saunders, 2004).

Now that many complete sequences are available, gene arrays are being used increasingly for bacterial genome analysis. This approach has proven to be of great value in helping to elucidate the genomic diversity and evolutionary relationships within species (Chan \textit{et al}., 2003; Dorrell \textit{et al}., 2001; Fitzgerald \textit{et al}., 2001; Fukiya \textit{et al}., 2004; Leonard \textit{et al}., 2003; Porwollik \textit{et al}., 2002; Salama \textit{et al}., 2000). Here the development of a \textit{S. aureus} microarray consisting of 383 oligonucleotide probes representing virulence and housekeeping genes is reported. The microarray was tested on 12 strains, of which seven have complete genome sequence information available. Data derived from the array will allow comparative analysis of isolates from geographically and clinically different groups as well as provide useful insights into the evolutionary history of \textit{S. aureus}.

**METHODS**

**Bacterial strains.** The strains included in this study are listed in Table 1.

**DNA extraction.** Strains were subcultured onto nutrient agar plates (PHLS, Colindale) and incubated overnight at 37 °C. Bacterial cells were harvested with a 10 μl plastic loop and lysed in an aqueous solution of lysozyme (2 mg ml⁻¹; Sigma) and lyostaphin (0.1 mg ml⁻¹; Sigma). DNA was extracted as described by Pitcher \textit{et al}., (1989) and the nucleic acid concentration was determined spectrophotometrically.

**DNA labelling.** DNA (1–5–2–0 μg) was labelled with fluorescent Cy3 or Cy5 dUTP (Amersham Biosciences) by random-primer labelling (BioPrime DNA labelling system; Invitrogen) to produce target for hybridization to arrays. Labelling was allowed to proceed for 3 h at 37 °C. Labelled DNA was isolated from the reaction mixture with a QIAquick PCR purification kit (Qiagen), following the manufacturer’s protocol. The bound DNA was eluted from the column with 50 μl distilled water and stored at −20 °C until required for hybridization.

**Array design.** The array was designed to include \textit{S. aureus} genes that have a potential virulence association. A genome browser was developed and used to search the \textit{S. aureus} sequences of strains MW2, Mu50 and N315, for which a complete annotation was available at the NCBI website (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html). The browser (http://193.129.245.227/cgi-bin/genome_browser/genome_browser.cgi) is a free-of-charge web tool that searches annotated genomes for user-defined keywords. The text format output includes the names, sequences (nucleotide and amino acid) and positions of any genes that have the defined keyword included in their annotation. The keywords used for searching were toxin, leukocidin, coagulase, adhesin, protease, nuclease, lipoprotein, lipase, capsular polysaccharide, binding, regulatory and iron. The un-annotated genome of \textit{S. aureus} strain 252 was searched with a series of perl scripts that identify protein motifs associated with virulence. The scripts use glimmer (http://www.tigr.org/software/glimmer/) to accurately predict genes and then translate them into predicted amino acid sequences. These sequences are piped into another script that searches them for a subset of protein motifs from the PRINTS database implicated in bacterial virulence (http://www.jenner.ac.uk/BacBix3/PPprints.htm).

**Table 1. Strains used in the study**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Abbrev.</th>
<th>Year of isolation</th>
<th>Sequence</th>
<th>Methicillin susceptibility</th>
<th>Notes</th>
<th>MLST and/or EMRSA type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCTC 8325</td>
<td>8325</td>
<td>1960</td>
<td>Univ. of Oklahoma</td>
<td>MSSA</td>
<td>Laboratory strain</td>
<td>ST8</td>
</tr>
<tr>
<td>COL*</td>
<td>COL</td>
<td></td>
<td>TIGR</td>
<td>MRSa</td>
<td></td>
<td>ST250</td>
</tr>
<tr>
<td>MW2*</td>
<td>MW2</td>
<td>1998</td>
<td>National Institute of Technology and Evaluation, Tokyo</td>
<td>MRSa</td>
<td>High virulence community-acquired</td>
<td>ST1</td>
</tr>
<tr>
<td>Mu50†</td>
<td>Mu50</td>
<td>1996</td>
<td>Univ. of Juntendo</td>
<td>VISA†</td>
<td>Hospital acquired</td>
<td>ST5</td>
</tr>
<tr>
<td>N315†</td>
<td>N315</td>
<td>1982</td>
<td>National Institute of Technology and Evaluation, Tokyo</td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>ST5</td>
</tr>
<tr>
<td>Sanger 252*</td>
<td>SA252</td>
<td></td>
<td>Sanger Centre</td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>ST36; EMRSA-16</td>
</tr>
<tr>
<td>Sanger 476*</td>
<td>SA476</td>
<td></td>
<td>Sanger Centre</td>
<td>MRSa</td>
<td>High virulence community-acquired</td>
<td>ST1</td>
</tr>
<tr>
<td>NCTC 11939</td>
<td>E1</td>
<td></td>
<td></td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>ST239; EMRSA-1</td>
</tr>
<tr>
<td>NCTC 13130</td>
<td>E3</td>
<td></td>
<td></td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>ST5; EMRSA-3</td>
</tr>
<tr>
<td>97/35481§</td>
<td>E15</td>
<td>1997</td>
<td></td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>EMRSA-15</td>
</tr>
<tr>
<td>95/7924§</td>
<td>E16</td>
<td>1995</td>
<td></td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>EMRSA-16</td>
</tr>
<tr>
<td>00/5307§</td>
<td>E17</td>
<td>2000</td>
<td></td>
<td>MRSa</td>
<td>Hospital acquired</td>
<td>EMRSA-17</td>
</tr>
</tbody>
</table>

*Strains obtained from NARSA (Network on Antimicrobial Resistance in \textit{S. aureus}) at the National Institute of Allergy and Infectious Diseases.
†Strains provided by Dr Mark Enright (Department of Biology and Biochemistry, University of Bath).
‡MRSA displaying intermediate resistance to vancomycin.
§Strains from the collection of \textit{S. aureus} cultures sent to the Laboratory of Healthcare Associated Infection, Health Protection Agency.
using the FingerPRINTscan tool (http://bioinf.man.ac.uk/fingerPRINTscan/). Potential genes encoding toxins and adhesins were flagged using this method. The sequences selected using the two software tools were compared to identify those genes retrieved more than once (i.e. gene homologues from different strains). When sequences with >98% homology were detected, only one example was retained for probe design. In addition to the sequences selected using the software tools, additional virulence-associated genes that were not present in the genomes of the four sequenced strains (e.g. staphylococcal enterotoxin E) and positive control genes were included. The positive controls were the seven house-keeping genes used in Multi Locus Sequence Typing (MLST) (Enright & Spratt, 1999; Enright et al., 2000; van Leeuwen et al., 2003) together with the 16S and 23S rRNA subunits.

Oligonucleotide probes were designed for a total of 383 S. aureus genes using the OligoArray program (Rouillard et al., 2002). Each oligonucleotide was a 50mer with a calculated Tm of approximately 80 °C and minimal internal structure. A local BLAST (Altschul et al., 1997) database comprising the complete genomes of strains MW2, Mu50 and N315 was used to ensure probe specificity. Table A (available as Supplementary Material in Microbiology Online) lists the genes selected for inclusion on the array.

Printing. Probes were synthesized (Illumina) with 5’-end C6 aminolithers. The modified oligonucleotides were printed onto epoxy slides (MWG) at a concentration of 25 pmol μl−1 in ArrayIT micro-spotting solution (Genetix). Six replicates were printed for each oligonucleotide using a BioRobotics Microgrid II arrayer equipped with MicroSpot 2500 pins (Genomic Solutions). Spotting solution alone was printed as a negative control. The complete array was printed to each slide in duplicate.

Slide processing. Printed microarray slides were placed in a chamber at 42 °C with 50–75% relative humidity provided by a saturated solution of NaCl for 16–18 h. Excess spotting solution was washed off with 0.2% SDS by agitation in a slide washer for 2 min at 20–25 °C. The slides were then washed, with shaking, three times in distilled water for 1 min each wash. The microarrays were then incubated for 20 min at 50 °C in distilled water, dried by centrifugation (1000 g for 5 min) and stored in the dark at room temperature.

Hybridization. Hybridization experiments were performed with a mixture of two target preparations applied to each array. Genomic DNA from strain Mu50 (Kuroda et al., 2001) was labelled with Cy5 and was used as an internal control strain in each of the hybridizations. DNA from the strain under test was labelled with Cy3. Half of the labelled DNA from the control strain and an equal aliquot of Cy3-labelled test strain were dried. The two probes were mixed together by resuspension in 18 μl distilled water. The target was denatured by heating at 95 °C for 3 min and allowed to cool before the addition of 18 μl genHYB hybridization buffer (Genetix). The hybridization mixture was warmed at 42 °C for 2 min. The duplicate arrays on the slide were hybridized under a single 22 × 64 mm glass coverslip. Slides were incubated for 16 h at 43 °C in hybridization chambers (Corning) wrapped in foil to exclude the light and sealed inside plastic bags. Following hybridization slides were rinsed in 1× SSC, 0.05% SDS and then washed for a further 5 min in 1× SSC, 0.05% SDS followed by a rinse in 0.06× SSC. After a final wash for 10 min with 0.06× SSC the slides were dried by centrifugation (1000 g for 4 min).

Arrays were scanned in the red (Cy5) and green (Cy3) channels of an Affymetrix 428 scanner. The two images were processed using ImaGene software (BioDiscovery). Briefly, the images were superimposed, the backgrounds were subtracted and the ratios of the two colour signals were calculated. The values were then corrected using the factor required to give a ratio of one for the housekeeping and rRNA genes. This adjusts for the variable specific activities of the target preparations. Signals that were three times the value for the negative control were considered positive when the ratio of fluorescence values (Cy3/Cy5) was in the range >0.25–4. This range included all ratios found in the analysis of an array hybridized to Mu50 DNA labelled with Cy3 and Cy5. Values above this range were taken to show that the test strain was positive and Mu50 was negative. When the ratio was <0.25 the test strain was scored as either negative or intermediate positive if the value was <0.25 and >0.1.

Phylogenetics. Distance matrices were calculated for pairwise comparisons of the probe hybridization patterns of selected gene sets using BioNumerics software (Applied Maths). Positive, negative and intermediate probe reactions were each considered as mutually exclusive states. Distance trees were computed from this matrix using FITCH from the PHYLIP suite of programs (Felsenstein, 1993). Trees were represented using the program TreeView (Page, 1996). Concatenated MLST sequences of the types represented in the study were prepared from data downloaded from the S. aureus MLST website (http://saureus.mlst.net/). A Jukes and Cantor model distance matrix was calculated using the program DNADIST from the PHYLIP suite and trees were then calculated and drawn as for the array data.

RESULTS

Selection of genes for the array

Searching the annotated genomes of the S. aureus strains N315, Mu50 and MW2 (Baba et al., 2002; Kuroda et al., 2001) using keywords listed in Methods gave a total of 340 genes, after careful editing to remove duplicate sequences. Duplicates occurred frequently in the first draft due to the strategy of recovering genes from multiple genomes. A further 22 genes that were not found using the browser keyword search were identified within the unannotated MRSA 252 sequence (http://www.sanger.ac.uk/Projects/S_aureus/) by searching for protein motifs linked with virulence. Eight other virulence genes that had not been selected using one of the genome tools were found in the literature and included. Finally, the seven housekeeping genes used in MLST (Enright et al., 2000), the 16S and 23S rRNA genes and genes representative of accessory gene regulator groups one to four (Jarraud et al., 2000, 2002; Ji et al., 1997) were included (Table A, Supplementary Material). The final array comprised probes for 393 genes.

Hybridization results

The probes’ reactions were scored as positive, intermediate or negative depending upon the intensity and ratio of the fluorescence signals for the reference and experimental strains. Using ratios allowed correction of the variation in signal for different probes. The ratios were normalized using the mean ratios found for the combined MLST and rRNA probes. These probes gave uncorrected hybridization ratios that fell in a narrow range (+/- 50%) on all slides indicating a good match to the target derived from all test strains. A total of 271 probes were found to hybridize to the DNA of all 12 strains tested. Twelve probes were negative for all strains (Table A). A minority of probes showed either negative or positive reactions depending upon the
target DNA. The genes belonged to a range of categories considered to be virulence-associated. The largest group of variable genes comprised 39. These encoded the family of toxins including the exotoxins, the enterotoxins and toxic shock syndrome toxin. Inter-strain variation also occurs among the genes encoding leukotoxins, proteases, adhesins, capsule synthesis proteins, drug resistance-related proteins, iron uptake proteins, regulatory proteins and various transporters (Table A). A few probes gave positive reactions with varying intensities, depending upon the target strain. Probes that exhibited intermediate positive reactions are indicated in Table A (Supplementary Material).

A total of 271 probes were found to hybridize to the DNA of all 12 strains tested, indicating that these genes or close homologues were present. Testing of further strains is likely to reveal that more of these genes are not present in all strains, particularly isolates derived from animal sources. Twelve probes were negative for all strains indicating either that they were not present in any of the strains tested or that the probe was defective in design or synthesis. The former possibility may be the case for the probes for the genes eta, etb, sed, see, sej and mapR, which are associated with extrachromosomal elements and were not present in strains used for design of the array. The intermediate positive hybridization values, e.g. for gene vraF in strains COL, NCTC 8325, 97/35481 (EMRSA-15) and 00/8357 (EMRSA-17), likely result from cross-reactions with gene homologues or alleles that match the probe sequence poorly. This is supported by the finding that most low positive reactions were found for genes that belong to families comprising many homologues. Overall, the results were as predicted from the available genome sequences.

Phylogenetics

Fig. 1 shows phylogenetic trees calculated on the basis of pairwise distance matrices in which the distances were the number of non-matching probe reactions for each pair. Three trees (a–c) are shown for genes associated with mobile elements, genes not associated with mobile elements and all genes. The trees show clustering of the two EMRSA-16 strains (SA252 and 95/7924) and that the closest neighbour of these isolates, EMRSA-15 (97/35481) was relatively distant. Strains Mu50 and N315 that belong to clonotype II-A (Kuroda et al., 2001) clustered with the EMRSA-3 isolate (NCTC 13130). The EMRSA-17 strain (00/5307) formed a loose cluster with strains COL, NCTC 8325, MW2 and SA476. Within this group, SA476 and MW2 were the most closely related strains. UK EMRSA-1 (NCTC 11939) clustered differently in the trees drawn for mobile element and non-mobile element associated genes. Fig. 1(d) shows a distance tree prepared using the concatenated MLST sequences for these strains. With the exception of the branch leading to strain NCTC 11939, this tree has relative branch lengths and topology most similar to those of the array data tree based on genes that are not associated with mobile elements. When strain NCTC 11939 is included, the topology is closest to that of the mobile-element tree.

Strains belonging to the same MLST type also clustered tightly on the array tree, i.e. MW2 and SA476 (ST1), Mu50, N315 and NCTC 13130 (ST5), and SA252 and EMRSA-16 (ST36). Furthermore, strains with different STs that cluster either in BURST analyses (Enright et al., 2002; Feil et al., 2004) or on concatenated MLST sequence trees were also clustered on the trees prepared using array data, i.e. 8325/4 (ST8), COL (ST250) and EMRSA-17 (ST247). Differences in the tree topologies and relative branch lengths were noted particularly for strain EMRSA-1 (ST239) and in the positions of strain EMRSA-15 and the ST5 strains.

Specific points

The probes designed to react with the MLST house-keeping genes and those for the 16S and 23S rRNAs reacted with all strains.

The array showed that, as expected, the mecA gene (penicillin-binding protein) was present in all MRSA strains and absent from the genomes of both methicillin-sensitive Staphylococcus aureus (MSSA) tested. The methicillin resistance regulatory protein gene (mecI) was only found to be present in the two EMRSA-16 strains, Mu50, N315 and the EMRSA-1 strain. MW0040, a hypothetical gene on the staphylococcal cassette chromosome mec (SCCmec), was present in all MRSA strains tested except EMRSA-1. MW0040 is therefore neither present within all SCCmec types nor specific for the type IV SCCmec. Three of the eight genes thought to encode drug transport proteins were only present in a proportion of the strains but there was no clear correlation with methicillin susceptibility status.

The exotoxin genes present on the genomic island vSa2 varied between strains except in the case of strains Mu50 and N315. These strains carried identical complements of the array exotoxin genes. The genes carried by the vSa2 islands in these strains (Fitzgerald et al., 2003; Kuroda et al., 2001) differ only in set9, absent in Mu50. Similar vSa2 islands were present in the following groups of strains as judged by their array hybridization patterns: MW2 and SA476, EMRSA-15 and the two EMRSA-16 strains, EMRSA-17 and EMRSA-1. Distinctive hybridization patterns were also seen for the enterotoxins, toxic shock toxins, leukotoxins and protease genes associated with the other S. aureus genomic islands vSa3, vSa3, vSa3 and vSa3. Similar groupings of strains were observed for these islands as in vSa2. The probe reactions for the variable reaction group are shown in Table 2.

The four agrC probes included on the array were designed to give an indication of the agr specificity group (Jarraud et al., 2000, 2002; Ji et al., 1997). The probes for groups I and II each reacted with five strains each but the groups III and IV were negative for all strains. Sequence comparisons show that strain MW2 would be expected to react with the probe complementary to specificity group III.

Three pairs of strains were found to differ by only a very
small number of reactions. SA252 and 95/7924 (EMRSA-16) differed only by the absence of *tst* and presence of SA 1833 (a hypothetical protein that has structural features of a transcription regulator) in SA252. Mu50 and N315, which are both MLST sequence type 5 strains, differed in the presence of *sea* and absence of the genes E16 1949 (a β-lactamase) and SA1804 (a hypothetical protein that has structural features of a transcription regulator) in Mu50. Finally, the MLST sequence type 1 strains MW2 and SA476 differed at eight loci. The eight genes, all of which were present in strain MW2 and absent from SA476, were *seb*, *sel2*, *sec3*, *lukF*, *lukS*, *mecA*, MW0040 (part of the type-IV staphylococcal cassette chromosome *mec*) and SA0066 (a hypothetical protein, similar to the *kdp* operon transcriptional regulatory protein). The reaction between the MW2 target DNA and the *sec3* probe was relatively weak, indicating the presence of sequence mismatches.

**DISCUSSION**

Many questions concerning staphylococcal disease remain unanswered. This is in part due to the adaptability of the organism, which in turn is a function of its well-organized but highly plastic genome (Baba *et al*., 2002; Fitzgerald *et al*., 2001; Kuroda *et al*., 2001). The microarray described here should provide insights into the pathogenicity, epidemicity and evolution of *S. aureus* clinical strains by providing profiles comprising information on their individual repertoire of virulence-associated genes. The profile also provides information relating to the distribution of genetic islands in each genome allowing detailed categorization of strains.

The gene categories chosen for arraying corresponded to the list given by Kuroda *et al.* (2001) for strains N315 and Mu50. Genes selected included the large family of toxins comprised of the superantigenic enterotoxins, toxic shock syndrome toxins and exotoxins together with leukocidins and haemolysins. Adhesins and binding proteins that may be involved in colonization, and exoenzymes such as proteases and lipases were also important components of the array. A large proportion of the array was made up of probes for regulatory genes to determine the degree to which these genes might differ between clonal lineages. These genes were expected to be relatively conserved in terms of their presence in the genome but may vary greatly in their levels of expression (Sabersheikh & Saunders, 2004).
### Table 2. Genes with differential hybridization

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene</th>
<th>COL</th>
<th>MW2</th>
<th>NH2</th>
<th>MSSA</th>
<th>SA14</th>
<th>E16</th>
<th>SA1</th>
<th>E15</th>
<th>E17</th>
<th>NCTC13310</th>
<th>NCTC11999</th>
<th>Ma50</th>
</tr>
</thead>
<tbody>
<tr>
<td>MW0031</td>
<td>mesC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0040</td>
<td>mezI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0051</td>
<td>mezJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0062</td>
<td>mezK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0066</td>
<td>mezL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0071</td>
<td>mezM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0072</td>
<td>mezN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0091</td>
<td>mezO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0094</td>
<td>mezP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0097</td>
<td>mezQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0099</td>
<td>mezR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0076</td>
<td>mezS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0078</td>
<td>mezT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0090</td>
<td>mezU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0091</td>
<td>mezV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0094</td>
<td>mezW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0097</td>
<td>mezX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0099</td>
<td>mezY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0076</td>
<td>mezA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0078</td>
<td>mezB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0090</td>
<td>mezC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0091</td>
<td>mezD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0094</td>
<td>mezE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0097</td>
<td>mezF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0099</td>
<td>mezG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0076</td>
<td>mezH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0078</td>
<td>mezI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0090</td>
<td>mezJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0091</td>
<td>mezK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0094</td>
<td>mezL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0097</td>
<td>mezM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0099</td>
<td>mezN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0076</td>
<td>mezO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0078</td>
<td>mezP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0090</td>
<td>mezQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0091</td>
<td>mezR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0094</td>
<td>mezS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0097</td>
<td>mezT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0099</td>
<td>mezU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0076</td>
<td>mezV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0078</td>
<td>mezW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0090</td>
<td>mezX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0091</td>
<td>mezY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW0094</td>
<td>mezZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend: Positive probe reactions; Intermediate probe reactions; Negative probe reactions.
The selection of genes from strains belonging to three sequence types ST-1 (MW2), ST-5 (N315 and Mu50) and ST-36 (SA252) (Enright et al., 2002; Feil et al., 2003) ensured that the array was relatively representative of the species. Although ST-1 and ST-5 strains cluster relatively closely in concatenated sequence analyses derived from MLST data, ST-36 is relatively distant (Fig. 1d). The genes were selected using a genome browser and annotation software developed during the course of this study to assist the process of handling the data available in the fully sequenced Staphylococcus aureus genomes. These tools facilitated the process of oligonucleotide array development. A number of genes were omitted in error due to non-optimal choice of keywords for searching the complete genome sequences of MW2, N315 and Mu50. For example the hypothetical lipoproteins on genomic island vSa2 were not returned by the genome browser using the keyword ‘lipoprotein’.

Sergeev et al. (2004) have described a microarray consisting of oligonucleotide probes for an assembly of enterotoxin genes. The array was shown to be able to analyse multiple enterotoxin genes of test strains. An advantage of this type of microarray is that it is practical to derive the target nucleic acid by multiplex PCR. Other workers have used PCR products as probes on S. aureus microarrays (Fitzgerald et al., 2001). Oligonucleotides have considerable advantages over PCR products as probes for array construction. The major advantage of oligonucleotide probes is their greater specificity, which is a function of their length. A relatively large PCR-derived gene probe may hybridize to homologous sequences within a related gene as well as to its target. This problem occurs far less frequently with a well-designed oligonucleotide, since any cross-hybridizing sequences will be too short to give a signal. High specificity was a particularly important factor in the design of this virulence gene array because many of the genes belong to extensive gene families that share some homology at the nucleotide sequence level (Baba et al., 2002; Holden et al., 2004; Kuroda et al., 2001). Further advantages of using oligonucleotide probes are that the design is simpler and that the costs are lower in both materials and technical effort. Potential disadvantages of oligonucleotide probes are their lower sensitivities and that they may not detect all alleles of a gene if several mutations are clustered at the relevant locus. For the array described here 50 base probes, end modified with a C6 aminolinker, were chosen to give the most advantageous combination of specificity, sensitivity and cost-effectiveness. The amino group on the probe reacts with the epoxy residue on the slide surface under mild conditions increasing the probability that it will become covalently attached in the conformation that provides least steric hindrance to hybridization. The choice of longer linker molecules reduces the level of interference between the slide coating and the probe. However, satisfactory signals were produced using the C6 linker under the hybridization and buffer conditions described here.

Mu50 was selected as the control strain because it was used in the design of the array. A possible disadvantage of this approach was that a positive control was not available for all probe reactions. Alternative methods of providing positive controls for each probe were considered, including the use of control target comprising mixtures of labelled DNAs extracted from different strains. However, it was considered that this solution would introduce additional sources of error and would complicate the interpretation of the data. A great advantage of the microarray format is that it is possible to include multiple replicates of each probe on a single slide. This greatly reduces the potential for false-positive and false-negative reactions. In the future it would also be possible to include multiple distinct probes for each open reading frame (ORF) allowing greater certainty over the accuracy of the results.

A large proportion of the genes (26 %) were not found in all strains. As expected, 60 % of these variable genes were associated with mobile genetic elements, even though these represented only approximately 20 % of the array. The array of strain COL ORF-specific probes (Fitzgerald et al., 2001) showed a similar high level of variation between the genomes of different S. aureus strains. The strain clustering deduced from the microarray data was supported by multilocus enzyme electrophoresis (MLEE) and gene sequencing of coa and spa (Koreen et al., 2004).

In order to investigate the relationship between the MLST and array datasets, distance trees were prepared for subsets of the array gene data. The genes were divided into two groups depending upon whether they were carried on mobile genetic elements (Table A, Supplementary Material). The branch lengths were longer on the tree derived from the mobile element genes than on that for the ‘non-mobile’ genes. Branches at the periphery of the non-mobile gene tree were particularly short, e.g. the branch lengths between strains COL and 8325/4 and strains SA252 and EMRSA-16 were zero. This might be anticipated since the genes on the mobile elements are expected to be more variable than genes in more stable chromosomal locations.

The topologies of the mobile and non-mobile element derived trees were not identical. Strain EMRSA-1 (ST239) clusters with the ST8 group in the tree derived from the mobile gene set. This cluster also appears with the MLST data by either BURST analysis or using the concatenated sequences. However, EMRSA-1 is remote from the other ST8 clonal complex strains on the non-mobile gene tree. This can be explained by the evolutionary origin of ST239 (Robinson & Enright, 2004). ST239 appears to be a chimera of strains from the ST8 (one allele difference) and ST30 (six alleles difference) clonal complexes (Robinson & Enright, 2004). Excluding SCCmec, the pattern of hybridization for EMRSA-1 genes surrounding the origin of replication was found to match that of EMRSA-16 (ST36) but not that of the ST8 clonal complex strains. The ST36-like region stretched from MW2612 (cna) through to MW0263 (an incomplete diarrhoeal toxin ORF) consistent with the splice sites described by Robinson & Enright (2004). The
remaining 'non-mobile' genes, i.e. those remote from the origin, clearly matched the ST8 clonal complex strains and mis-matched the ST36 strains. Thus, the array data support the evidence from sequence typing that large-scale recombination events have occurred within the *S. aureus* chromosome (Robinson & Enright, 2004). The simplest explanation for the genesis of ST239 (including EMRSA types 4 and 9) is that an ancestral strain belonging to clonal complex ST30 acquired most of its chromosome by homologous recombination during conjugation. This strain may have been an MSSA but was more likely an MRSA clone carrying SCCmec type III (Enright et al., 2002). ST36 is a single locus variant of the widespread ST30 that carries SCCmec type II. The donor strain that contributed the greater part of the chromosome would have been similar to strains in the ST8 clonal complex.

There was a significant association between the ST1 (MW2 and SA476) strains and the ST8 clonal complex strains (8325/4, EMRSA-17 and COL) in the mobile element gene but not in the non-mobile gene tree. This appears to be largely because the type I *vSaβ* pathogenicity island is carried by ST5 while ST1 and ST8 clonal complex strains carry type II *vSaβ*. The non-mobile gene tree shows no strong clustering between these three sequence types and suggests that the clustering between STs 1 and 5 seen in the concatenated MLST sequence tree is an artefact due to the greater weighting given to the MLST loci with variable sequences. This illustrates the value of the preferred eBURST method (Enright et al., 2002; Feil et al., 2004) of viewing MLST data.

The closest neighbour of the EMRSA-16 strain(s) SA252 and 95/7924 in the array trees was the EMRSA-15 strain (Fig. 1a–c). This relative relationship is not apparent in the MLST data and is potentially very interesting since these two types are currently responsible for a large proportion of healthcare-associated MRSA infection in the UK (Johnson et al., 2001). Common features of these lineages have been discussed previously (Moore & Lindsay, 2002). Data for the mobile genes indicate that the EMRSA-15 (ST22) and 16 (ST36) strains have similar pathogenicity islands, but the association also appears for the 'non-mobile' genes. Analysis of the data indicates that EMRSA-15 may have been the result of another large-scale recombination event involving a strain belonging to the ST30 clonal complex. In this case however, the ST30 parent appears to have contributed a large proportion of the chromosome remote from the origin of replication (from approximately MW0172 to MW1938) including the large pathogenicity islands (Table 2). The array evidence for this recombination event is less clear than that for the chimeric nature of ST239. This indicates either that the genesis of ST22 was an ancient event or that the parent strains were relatively distant relatives of strains included in the study. It seems likely that the ability of EMRSA-15 and -16 strains to colonize patients and cause invasive disease is significantly influenced by the carriage of a similar complement of genes. However, it may be that chimeric strains are more likely to be involved in invasive disease due to impaired functioning of control mechanisms. This may be a result of the poor integration of different sets of genes as suggested by Robinson & Enright (2004). This is partly supported by the evidence (Okuma et al., 2002) that ST239 strains have a lower growth rate than either of the parent clones (ST30 and ST8 clonal complexes).

The data presented here show that the virulence-associated gene array can provide detailed information about gene carriage in *S. aureus*. In addition, it provides a powerful tool for the detection of large-scale genome recombination events that may have been key factors in the evolution of pathogenic strains of the species.

**ACKNOWLEDGEMENTS**

Strains MW2, COL, Sanger 252 and Sanger 476 were obtained through the Network on Antimicrobial Resistance in *Staphylococcus aureus* (NARSA program: supported under NIAID, NIH Contract No. NO1-AI-95359). We thank Professor Saheer Gharbia for critical reading of the manuscript.

**REFERENCES**


